École polytechnique fédérale de Lausanne

Master Thesis

Master in Applied Mathematics

A Study of Extreme Rainfall in Costa Rica with Multivariate Extreme Value Analysis

Author:
Antoine Bourret

Supervisor:
Prof. Anthony Davison

A thesis submitted in the fulfillment of the requirements for the Master Degree of Applied Mathematics

at the

École polytechnique fédérale de Lausanne

January 20, 2022

Abstract

The study of extreme rainfall events through the construction of intensity-duration-frequency (IDF) curves is of great importance for water resource management and drainage basin systems design. This report focuses on creating IDF curves for extreme precipitation in Costa Rica, through the use of extreme value theory and deep learning. Bivariate and multivariate analyzes with non-stationary max-stable processes are used to quantify the joint probability of extreme rainfall at multiple sites, along with a regionalization approach based on hierarchical clustering to decompose the territory into regions with similar behaviours with respect to extreme events. This study highlights regions with strong seasonality and a clear demarcation between the Pacific and Caribbean coasts, especially for extreme precipitation events that extend over long periods. The use of machine learning tools allowed to model the local effects and the different weather systems present in Costa Rica.

Acknowledgements

First of all, I would like to express my sincere thanks to my thesis professor, Anthony Davison, for the encouragement, suggestions and inspirations he gave me during the development of this thesis. His patience and enthusiasm have greatly contributed to my taste for mathematics and statistics over the years.

I also thanks the expert involved in this research project, Juan José Leitón-Montero, from the Instituto Costarricense de Electricidad, for his great help and participation.

Finally, I would also like to express my profound gratitude to my relatives as well as to my classmates, for their continuous support during all these years of study.

Contents

Chapter 1

Introduction

Extreme rainfall can have a significant negative impact on local populations, with increased risks of flooding, crop failure and landslide. Motivated by the opportunities to potentially reduce the direct and indirect impacts of such events on the environment and the populations, the study of extreme events in domain such as finance, oceanography, meteorology or health care has also been used to develop and illustrate numerous statistical tools from both theoretical and practical levels, especially through the lens of extreme value theory. The analysis of rare events from an univariate point of view is well explored in Coles [2001], Beirlant et al. [2004] and de Haan and Ferreira [2006], and Chapter 3 provides a review of these methods.

Because precipitation is generally monitored at specific locations, one needs to consider a number of locations to be able to derive relevant statistical and spatial behaviour of extremes across a region of interest, which then raises concerns about dependence of extremes across multiple locations. The study of such quantities often lies at the intersection of extreme value theory and geostatistics, for which an extended review of the most common tools and practices can be found in Cressie [1993], Banerjee et al. [2004], Schabenberger and Gotway [2005] and Cressie and Wikle [2011]. Because extreme events are unusual by nature, the amount of data that can be used for statistical inference is limited, regardless of the amount of available data. Thus, it is crucial to use flexible statistical models with strong mathematical background.

This report focuses on the application and extension of such tools for the study of extreme rainfall in Costa Rica. Located in Central America Isthmus, Costa Rica has an intercontinental and interoceanic position, between North and South America, and between the Pacific and Atlantic oceans. This characteristic makes it a region with a diversified climate and a rich biodiversity, dominated by complex climate systems and diversified geographical attributes, notably the Cordillera Central, large volcanic mountains stretching the length of the country. Widely known worldwide for its conservation efforts, Costa Rica is a country in the environmental service payment system, with more than three quarters of its land under protection[1]. Among the main sectors are tourism, agriculture and the exploitation of its natural resources [GFDRR, 2011]. However, its major vulnerabilities lie in extreme climate events and natural hazards, mainly due to the presence of populations in areas sensitive to landslides and volcanic eruptions.

The objective of this report is to provide methods to assess rainfall risks over different periods of interest, taking into account seasonal and spatial variability, as well as spatial dependence of extremes. We start with an exploratory analysis in Chapter 2, in which we present

[1] Costa Rica's Second National Plan (2009)

the source and availability of the rainfall data. We then review and apply in Chapter 3 the extreme value theory in an univariate setting. This section also explores how one can estimate return levels for extreme precipitation at any location in Costa Rica, through the use of a conditional density network to approximate the parameters of the distribution of maximal rainfall as a function of spatial and temporal covariates.

Chapter 4 examines the relationship between extreme precipitation of any two rain gauges in Costa Rica, within a bivariate framework. In this chapter, we review the theory that revolves around bivariate models in an extreme value analysis framework, apply it for a sample of stations across Costa Rica, and then propose models that can provide measure of dependency between any two locations. In chapter 5, we review the theory of max-stable processes and use it to estimate extremal dependence of extreme rainfall between different precipitation gauges. Due to the presence of local climatic phenomena, a regionalization approach with hierarchical clustering based on Saunders et al. [2019] is used to decompose the territory into regions with similar behaviours with regards to extreme precipitation. Another approach from Huser and Genton [2016] involving non-stationary max-stable processes is then reviewed and extended with an artificial neural network to model the different weather systems of Costa Rica.

Chapter 2

Data and exploratory analysis

2.1 Generalities

Costa Rican rainfall undergoes different regimes created by the interaction of trade winds with mountains and plains, and can be broadly associated with the Pacific and Caribbean coasts of the country, separated by the Cordillera Central [Hastenrath, 1967]. Alfaro [2002] and Alfaro et al. [2017] explore the distribution of rainfall in relation with the aforementioned oceans. The Pacific coast is characterized by a dry season from November to April and a rainy season from May to October. The maxima are reached at two different periods in the year (May–June and September–October), in what is usually known as the Little summer [Karnauskas et al., 2013]. The Caribbean coast has a much less pronounced seasonality, with slightly decreasing precipitation levels along the coast from north to south [Amador, 1998]. A review of Costa Rican climate and a projection of precipitation and temperature at the end of the century is given in Castillo and Amador [2020]. Aside from trade winds, elevation is also an important predictor for average rainfall, as shown for example in Song et al. [2019], Gonfiantini et al. [2001] and Wei et al. [2008]. Temperature is the key element driving this relationship, as cooler temperatures (usually linked with high altitudes) result in a decrease in maximum precipitable moisture, which then causes the appearance of an altitude of maximum precipitation [Vuille, 2011]. The elevation-precipitation relationship can vary depending on continentality, rain shadow and barrier effect. Change of elevation also affects rainfall, with high precipitation on the windward side of the mountains, and dry and quickly (adiabatic) warming winds on the other side. This is known as the Foehn effect, observable for example in the north of the Alps and in the Andes. Because of the presence of very mountainous regions and trade winds, extreme rainfall is likely to be affected in the same way as average rainfall. Hence, this study uses precipitation and topographical data to estimate the distribution of extreme rainfall.

Topographical data of Costa Rica were obtained from the General Bathymetric Chart of the Oceans (GEBCO, see `https://download.gebco.net`) and the Japan Aerospace Exploration Agency (JAXA, see `https://global.jaxa.jp/`). Figures 2.1 and A.1 (Appendix A) show the topography of Costa Rica, along with the 161 gauge stations considered in this study, grouped by drainage basin. The stations tend to be located at strategic places, such as mountains, rivers and lakes, and so their topographical characteristics do not constitute a representative sample of Costa Rica's topography, which also has vast areas with low-elevation deltas. This is especially visible for the northern part of the country (which overlaps the Heredia, Lemón and Alajuela provinces) and the province of Guanacaste, where few or no stations are present. The southern part of the province of Lemón has only one station that can provide observations, as this region is heavily wooded with tropical forests.

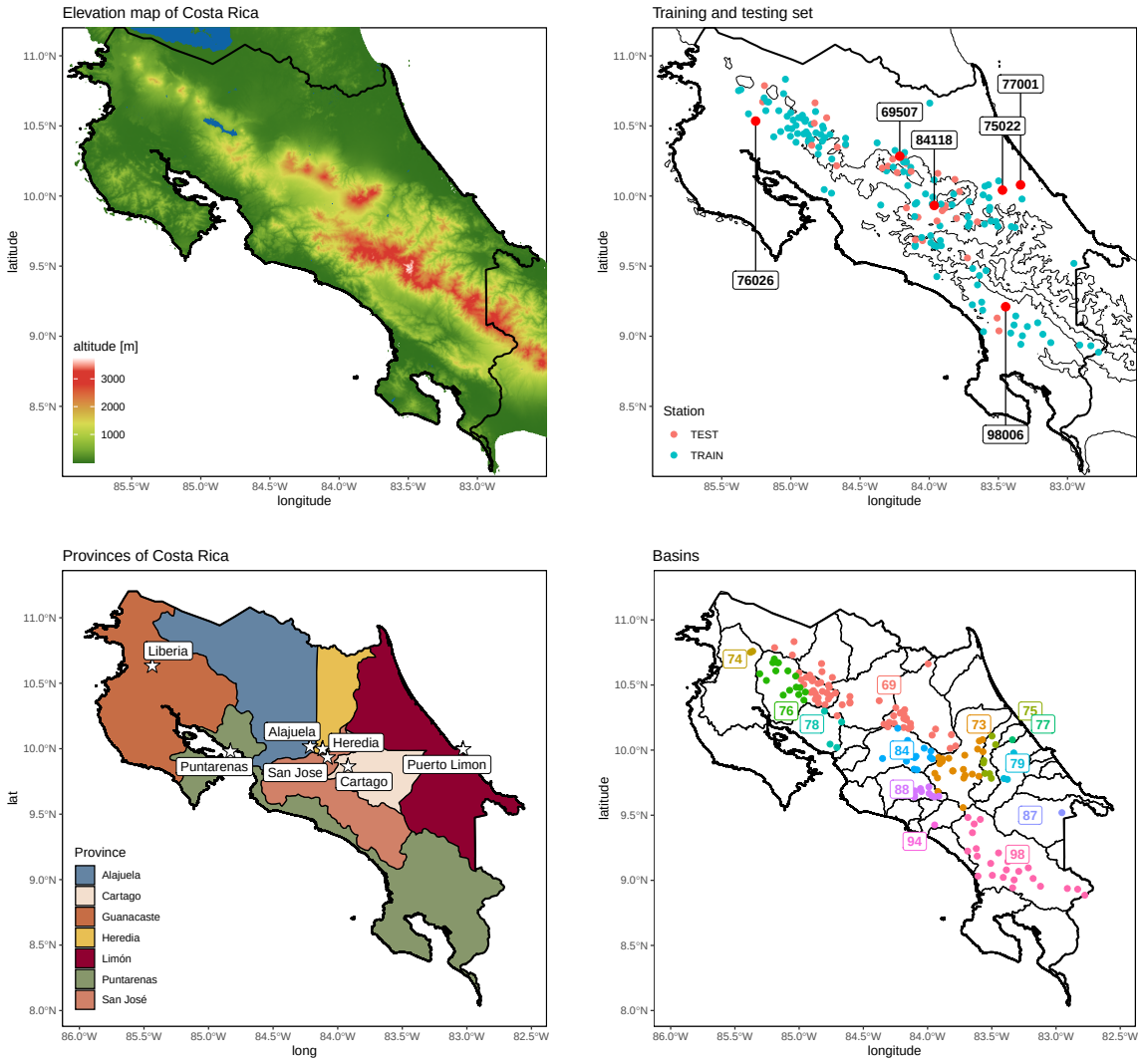


Figure 2.1: Elevation maps of Costa Rica (top left), set of the locations of training and testing stations (top right), the provinces of Costa Rica (bottom left) and of drainage basins (bottom right).

2.2 Rainfall data

The characterization of rainfall precipitation is often done with intensity-duration-frequency (IDF) curves, as introduced in Koutsoyiannis et al. [1998]. These curves are of paramount importance for water resource management, drainage basin systems design and extreme rainfall events. In the following, we first provide general mathematical formulations of IDF curves, and then describe some of the methods to obtain them.

Let $\tau(t)$ denote the instantaneous rainfall process intensity, at time t. In practice, $\tau(t)$ is not observed, but only the average intensity $\tau_\delta(t)$ over some resolution δ

$$\tau_\delta(t) = \frac{1}{\delta}\int_{t-\delta}^{t} \tau(s)\,\mathrm{d}s. \tag{2.2.1}$$

The resolution typically ranges from 5 to 60 minutes, depending on precipitation levels and technical details. One can then construct the moving average rainfall process as

$$\tau_d(t) = \frac{\delta}{d}\sum_{i=0}^{N-1} \tau_\delta(t - i\delta), \quad d = N\delta, \tag{2.2.2}$$

where d, usually referred to as the duration, is a multiple of the resolution. The series of maximum average intensities can then be constructed as

$$i_l(d) = \max_{l^- < t < l^+} \{\tau_d(t)\}, \tag{2.2.3}$$

where l^-, l^+ correspond respectively to the beginning and the end of the period of interest l, typically daily, monthly or yearly intervals. One can then consider the series of maximal intensities $\{i_l(d), l = 1, \ldots, n\}$ as multiple realizations of a random variable $I(d)$ with probability distribution function F_d. The return period T associated with level x is defined as $T = 1/\{1 - F(x)\}$, and can be interpreted as the average time between two successive rare events of magnitude greater than x. Stedinger and Foufoula-Georgiou [1993] and Koutsoyiannis et al. [1998] provide an extensive review of possible distributions for F, including the generalized extreme value (GEV) distribution, discussed in more detail in Chapter 3.

In this study, we consider daily average rainfall intensities for 161 stations across Costa Rica, provided by the Instituto Costarricense de Electricidad. The observations range from January, 1990 to June, 2021, and are given in mm/hr. Our dataset has rainfall intensities with durations 5, 10, 15, 30, 60, 120, 180, 360, 720 and 1440 minutes. Several recording instruments were used to extract these intensities, of which the three main ones are the digitizing table (also called Mesa, labelled M), the datalogger (D) and telemetry (T). The digitizing table is a method to extract information from pluviograph strip charts [Sansom, 1987]. Although recent methods were created to automatically digitize them [Jaklič et al., 2015], this process is generally conducted manually, and often results in non-standard time increments between observations. Pluviographs can be replaced by a logger, which automatically records information, but its extraction is still performed manually every one or two months. Extraction can take more time when the precipitation gauge is hard to access. The last data source, telemetry, corresponds to a datalogger with automatic transmission of information from precipitation gauges. Owing to the distribution of recording stations across Costa Rica and technical constraints, time steps between measurements can vary from one data source to another. As some stations can provide data from two different sources, we used the following prioritization to keep only one value per day and duration: M, D and finally T.

Figure 2.2 shows the number of observations per day for all 161 stations and 10 durations. There is a significant increase of records after 1999, as new stations started to provide daily observations, and a lower number of daily measurements for the first months of the year. This is highlighted in Figure A.2 (Appendix A), where darker regions are linked with fewer observations. This is especially visible for February, March and April (which are also linked with lower daily intensities). The stations are grouped by drainage basin, and exhibit similar patterns of data availability. Basin 76, corresponding to the stations that are the closest to the Pacific coast, has a very low number of observations during the first months of the year, usually an average of 5 to 10 observations per month. The last station of basin 69 (ID 69702) only provides measurements since 2019, and with large discontinuities, so extra attention will be paid to this station, which is at a strategic location in Costa Rica (in the north of the Caribbean coast, where few stations are sited). Finally, the first rainfall intensity measurements (before 1999) are relatively well spread across the country, and not clustered at some specific location, so the distribution of their locations is still representative of the distribution over the entire period of time (1990 to 2021). Since 2019, one can also notice fewer measurements (with fewer stations reporting daily intensities), which motivates the use of multivariate analysis tools to create consistent IDF curves over the whole of Costa Rica.

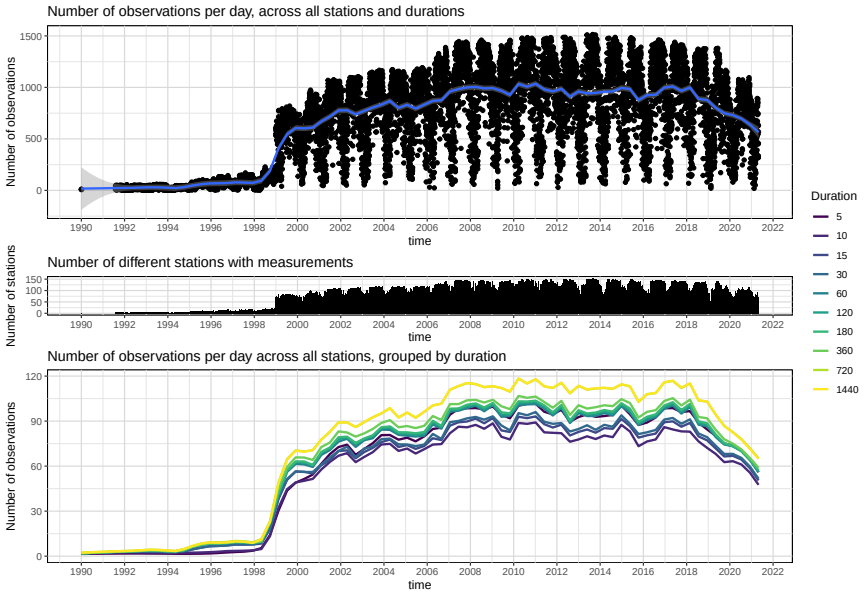
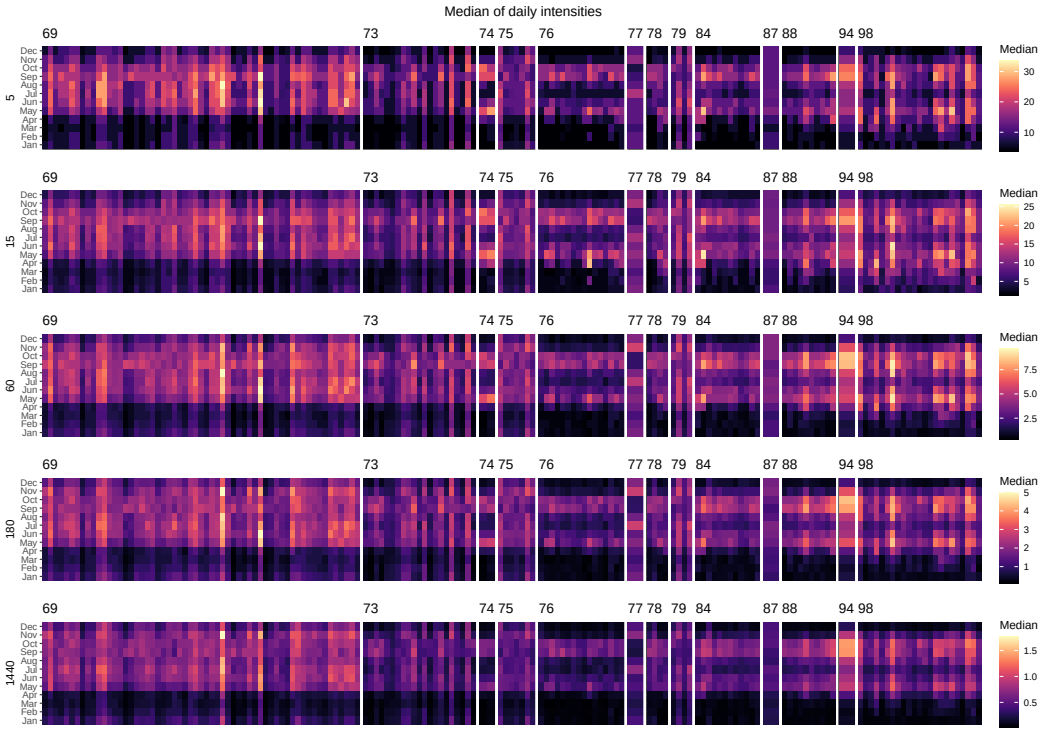


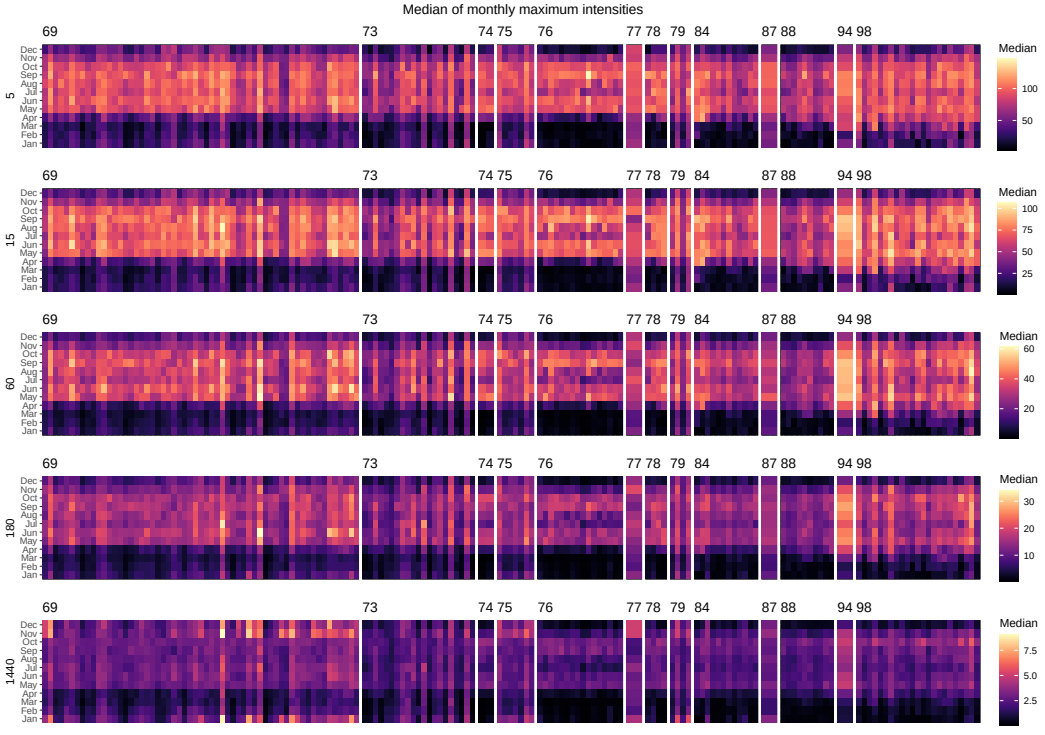
Figure 2.2: Number of observations across time, for all 161 stations and 10 durations. (top) All the durations are aggregated, and a LOESS regression function is shown in blue. Middle panel shows the number of stations that produced measurements, for each day. One can see that the number of stations that started to record since 1999 increased significantly compared to the period 1990–1999. (bottom) A smooth regression LOESS function applied to the number of observations for each day and all stations, colored by duration.

Figure 2.3 shows the median of daily and monthly maximum rainfall intensities grouped by stations, months and durations. We used the median for more robustness, as a couple of stations had very large values. This includes stations 76044, 69682, and 69520, but only for specific durations. As these values were not particularly related to typical extreme events, were very localized in space and few observations were available before the date on which high values were observed, we decided to delete these observations from the dataset, as they can have a large impact on parameter estimates for tail distributions. Median daily values for high durations range from 0 to 5 mm/hr, and for lower durations from 0 to 30 mm/hr. Maximum median rainfall intensities are reached between May and October for all stations, and lowest intensities during December and January. There are lower rainfall levels during July for a couple of stations, in what is called the Little summer. The stations from basins 73, 76, 84 and 88 (located in the middle of the country and on the Pacific coast) have lower median intensities over the year than the other stations, which reveals different behaviour of climatological events between the Pacific and Caribbean coasts. Regarding median monthly maxima, similar conclusions can be drawn. One can observe significantly higher median values between May and October compared to the other months, for most of the durations. This excludes stations on the Caribbean coast for long durations (720 to 1440 minutes), for which the peak is reached between December and February.

In the following chapters, we present models to create IDF curves for various locations in Costa Rica, and identify potential similarities and joint behaviour of extreme events through the lens of extreme value theory. Through this report, we use 20% of the stations as a validation set, as shown in Figure 2.1.



(a)



(b)

Figure 2.3: Median of daily and monthly maximum rainfall intensities across stations, months and durations. The color scale depends on duration, for better visibility. Overall higher intensities are observed during May-June and August-September. Note the Little summer in July, with lower intensities.

Chapter 3

Univariate analysis of extreme rainfall intensities

As shown in Chapter 2, the construction of IDF curves relies on a frequency analysis in order to derive exceedance probabilities. Stedinger and Foufoula-Georgiou [1993] proposed several distributions for modeling maximum intensities, such as the normal, log-normal, and generalized extreme value distribution, which is particularly relevant for modeling the maximum of several random processes, such as mean precipitation.

This chapter is divided into three parts. In the first, we propose a review of extreme value theory (EVT) and its two approaches: the block maximum and the peaks over threshold (POT). The first seeks to model the maximum of a sequence of random variables, while the second models the exceedance probability of these random variables over some threshold. In this report, we mainly focus on the block maximum approach. Possible extensions to accomodate for the POT approach will be discussed at the end of this chapter. More details about EVT can be found in Beirlant et al. [2004] and de Haan and Ferreira [2006], and in the following, we refer to the notation of Coles [2001] and Davison [2019].

In the second part, we apply the theory of extreme values to construct and analyze IDF curves for six stations across Costa Rica. Exploratory analysis revealed different behaviors of daily and monthly mean maxima, depending on duration, month and location. As monthly maxima and multiple durations are considered, the modeling will have to take into account seasonality and its interaction with durations.

In the last part of this chapter, we seek to extend these IDF curves to any location in Costa Rica. This is of particular interest for locations that are hard to access and for which direct measurements of precipitation levels cannot be produced, and therefore used to construct IDF curves. In this part, we will propose two approaches to incorporate the location of precipitation gauges into our modeling, and compare them.

3.1 Extreme value theory

3.1.1 Block maximum approach

Extreme value theory seeks to characterize statistical properties of extreme events, for which few observations are available. The typical setting is to consider a sequence $X_1, \dots, X_n$ of independent random variables with distribution function F. The central limit theorem gives convergence results about the sum $\sum_{i=1}^{n} X_i$, but we will focus here on the maximum $M_n = \max(X_1, \dots, X_n)$. While it is clear that $\mathbb{P}(M_n \leq x) = F^n(x)$ by independence of the

random variables, the distribution function F is in practice unknown and so the distribution of the maxima M_n as $n \to \infty$ can be approximated by

$$F(x)^n \to \begin{cases} 0, & F(x) < 1, \\ 1, & F(x) = 1, \end{cases} \tag{3.1.1}$$

so that $M_n \to x_F$ almost surely as $n \to \infty$, where $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$. Similar to the central limit theorem, one can consider to study the convergence of the centered and scaled quantities $(M_n - b_n)/a_n$, with $b_n \in \mathbb{R}$ and $a_n \in \mathbb{R}_+$. The following theorem provides very useful asymptotic results.

Theorem 3.1.1 (Extremal types)**.** *If there exist sequences of constants $b_n \in \mathbb{R}$ and $a_n \in \mathbb{R}_+$ such that*

$$\mathbb{P}\left\{(M_n - b_n)/a_n \leq x\right\} \to G(x), \to \infty,$$

where G is a non-degenerate distribution function, then G must be the generalized extreme value distribution (GEV),

$$G(x) = \begin{cases} \exp\left[-\left\{1 + \xi(x-\mu)/\sigma\right\}_+^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left[-\exp\left\{-(x-\mu)/\sigma\right\}\right], & \xi = 0, \end{cases} \tag{3.1.2}$$

for $x \in \{x : 1 + \xi(x-\mu)/\sigma > 0\}$, $\xi, \mu \in \mathbb{R}$ and $\sigma > 0$.

The distribution G has three parameters, controlling the location (μ), the scale (σ) and the shape (ξ) of the distribution. The latter is crucial to determine the rate of tail decay, since when $\xi > 0$, one has a heavy-tailed Fréchet distribution with support bounded below at $\mu - \sigma/\xi$, and when $\xi > 0$ we obtain the short-tailed Weibull distribution, with support bounded above at $\mu - \sigma/\xi$. The case $\xi = 0$ corresponds to the light-tailed Gumbel distribution, with support on all real values. A proof of the extremal types theorem can be found in Leadbetter et al. [1983].

Suppose that one has access to a collection of independent observations $X_1, X_2, \ldots$. In order to fit the GEV distribution, one can partition the observations into blocks of lengths n and create the set of maxima $M_{n,1}, \ldots, M_{n,K}$, where

$$M_{n,i} = \max_{j \in B_i} X_j, \tag{3.1.3}$$

with B_i the set of indices corresponding to the ith block. In the following, we will use the notation $Z_i := M_{n,i}$ for clarity. The choice of block length is of primary concern, as it often results in a variance/bias trade-off. A small block length typically results in maxima computed on smaller samples, resulting in a higher bias. Selecting a block length that is too large will create few block maxima, so the estimation variance will be larger. The block size is often based on a familiar time period, leading to the use of weekly, monthly or yearly maxima. For the GEV distribution, extreme quantiles are defined as

$$z_p = \begin{cases} \mu - \sigma/\xi\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0, \\ \mu - \sigma\log\{-\log(1-p)\}, & \xi = 0, \end{cases} \tag{3.1.4}$$

with p such that $G(z_p) = 1 - p$, and z_p corresponding to the p-block return level with return period $1/p$, where $0 < p < 1$.

If the maxima $Z_1, \ldots, Z_K$ are independent GEV variables, one can estimate the parameters μ, σ and ξ by maximizing the log-likelihood function, which (when $\xi \neq 0$) is given by

$$l(\mu, \sigma, \xi) = -K\log\sigma - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{K}\log\left[1 + \xi\frac{z_i - \mu}{\sigma}\right] - \sum_{i=1}^{K}\left[1 + \xi\frac{z_i - \mu}{\sigma}\right]^{-1/\xi}, \tag{3.1.5}$$

for all z_i in the support of the distribution function. When $\xi = 0$, the log-likelihood is

$$l(\mu, \sigma) = -K \log \sigma - \sum_{i=1}^{K} \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^{K} \exp \left(-\frac{z_i - \mu}{\sigma} \right). \tag{3.1.6}$$

As the support of the distribution depends on the parameters, standard asymptotic properties of the MLE might fail. However, as noted in Smith [1985], if $\xi > -0.5$, the MLE is regular and the usual properties are valid. The case $\xi \leq -0.5$ corresponds to a distribution with very short bounded upper tail, which is rarely encountered in practice.

3.1.2 Peaks Over Thresholds

As mentioned in the previous section, one major drawback of the block maximum approach is that it heavily relies on the selection of a correct partition of the data into blocks, resulting in a trade-off between variance and bias. Usually, the selection of the size of the blocks is based of a familiar time period, but this can lead to a waste of information when several high values occur in the same block. The peaks over threshold (POT) approach addresses this issue, by selecting a threshold u and estimating the distribution of the observations above this threshold. More specifically, suppose that one has access to a sequence of independent random variables $X_1, X_2, \ldots$ with marginal distribution function F. One can then be interested in the conditional probability

$$\mathbb{P}(X > u + y \mid X > u) = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0. \tag{3.1.7}$$

As with the block maximum approach, the distribution F might not be known. The exceedance theorem informs us about the convergence of the conditional distribution described above.

Theorem 3.1.2 (Exceedance)**.** *Suppose that X is a random variable with distribution function F, and that $c(u)$ can be chosen so that the limiting distribution of $(X - u)/c(u)$, conditional on $X > u$, is non-degenerate as u approaches the upper support value $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ of X. If such a limiting distribution exists, it must be of generalized Pareto form, i.e.,*

$$H(x) = \begin{cases} 1 - (1 + \xi x/\sigma_u)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-x/\sigma_u), & \xi = 0, \end{cases} \tag{3.1.8}$$

with $x > 0$, $\xi \in \mathbb{R}$ and $\sigma_u > 0$. H is then called the generalized Pareto distribution (GDP).

A proof of this theorem is given in Leadbetter et al. [1983]. Note that

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x \right) = F^n(b_n + a_n x) = \left(1 - \frac{n\{1 - F(b_n + a_n x)\}}{n} \right)^n = \left(1 - \frac{\Lambda_n(x)}{n} \right)^n \tag{3.1.9}$$

using the notation $\Lambda_n(x) = n\{1 - F(b_n + a_n x)\}$, with $x \in \mathbb{R}$. Taking b_n such that $1 - F(b_n) = 1/n$, we have that, for $x \in \mathbb{R}$,

$$\Lambda_n(x) = n\{1 - F(b_n + a_n x)\} = \frac{1 - F(b_n + a_n x)}{1 - F(b_n)} = \mathbb{P}(X > b_n + a_n x \mid X > b_n). \tag{3.1.10}$$

Since $(1 + a_n/n)^n \to e^a$ for a_n if and only if $a_n \to a$, convergence of the conditional probability $\mathbb{P}(X > b_n + a_n x \mid X > b_n)$ to a non-degenerate limiting function $\Lambda(x)$ is equivalent to convergence of the centered and scaled maximum $(M_n - b_n)/a_n$ to a non-degenerate limiting random variable. This shows the relationship between GEV and GPD distributions. The parameter ξ is equal in both, and $\sigma_u = \sigma + \xi(u - \mu)$, where u denotes the high threshold for exceedances.

The first step in the estimation of the parameters of $H(x)$ is to define a convenient threshold u. Following Coles [2001], the selection of such threshold can be done using mean residual life plots, which plot the points

$$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)} - u)\right) : u < u_{\max}\right\}. \tag{3.1.11}$$

If the threshold u_0 is such that the generalized Pareto distribution is valid, then it should also be valid for $u > u_0$, and $\mathbb{E}(X - u \mid X > u)$ should be a linear function of u. One can also estimate the parameters of the generalized Pareto distribution for multiple thresholds, and then select one for which the parameter estimates are stable for thresholds above u, where the approximation holds.

Assuming that a threshold u is already chosen, the parameters of the GDP can be estimated by maximum likelihood. Denoting by $y_1, \ldots, y_m$ the m excesses over u from a sample $x_1, \ldots, x_K$, the log-likelihood for $\xi \neq 0$ is expressed as

$$l(\sigma_u, \xi) = -m \log \sigma_u - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{m} \log\left(1 + \xi y_i/\sigma_u\right), \tag{3.1.12}$$

for $y_i \in \{y : 1 + \xi y/\sigma_u > 0\}$, $\xi \in \mathbb{R}$ and $\sigma_u > 0$. For $\xi = 0$, the log-likelihood becomes

$$l(\sigma_u) = -m \log \sigma_u - \sum_{i=1}^{m} y_i/\sigma_u. \tag{3.1.13}$$

Similar to the block maxima approach, the level x_l that is exceeded on average once every l observations is given by

$$x_l = u + \frac{\sigma_u}{\xi}[(l\zeta_u)^{\xi} - 1], \tag{3.1.14}$$

where $\zeta_u = \mathbb{P}(X > u)$.

3.2 Detailed study of several stations

In this section, we apply the theory of univariate extreme values to some stations across Costa Rica, those with IDs 69507, 98006, 77001, 76026, 75022 and 84118. These stations are located at different and representative places in Costa Rica, from the Caribbean to the Pacific coast, with some at higher altitudes (ranging from 15m for station 77001 to 1720m for station 84118, see Figure A.1 Appendix A). As stations 75022 and 98006 are located at similar altitudes (438m and 427m respectively), this section will reveal if altitude has a major impact on the characterization of IDF curves, or if more complex phenomenon are at play. All these stations began to provide measurements from January 1999, for about 15 to 20 years of daily data. Figure B.1 (Appendix B) shows the rainfall intensities for durations 60 and 720 minutes, as well as the monthly maxima, and Figure 3.1 shows the locations of these stations. For both long and short durations, monthly maxima exhibit seasonality, which differs depending on the locations of the precipitation gauges. The station that is the closest to the Pacific coast and the province of Guanacaste has very low rainfall intensities at the beginning of the year, and thus low monthly maxima for this period. Conversely, stations 75022, 77001 and 69507 have weaker seasonality, especially for long durations.

This section focuses on the use of the block maximum approach to create IDF curves. The objective is to model monthly intensity maxima M_t $(t = 1, \ldots, T)$, for each duration and location, with the GEV distribution introduced in Section 3.1.1. For this, assume that

$$M_{t,d,s} \sim \text{GEV}(\mu_{t,d,s}, \sigma_{t,d,s}, \xi_{t,d,s}), \tag{3.2.1}$$

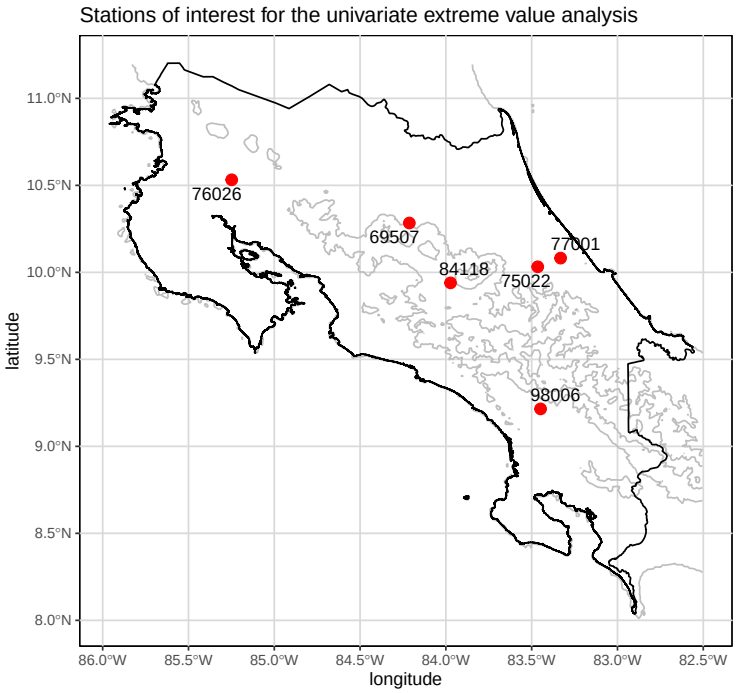


Figure 3.1: Location of the six stations for the example univariate extreme value analysis. Grey lines correspond to the 1000m elevation contours.

where t is the time, d the duration and s the station identification. In this section, each station s is treated separately, and we will present a model that combines intensities for multiple stations in Sections 3.3 and 3.3.2. Non-stationarity for GEV modelling can be introduced by fitting response surfaces to the GEV parameters [Coles, 2001, p.105]. For example, one can assume that maxima $Z_i \sim \text{GEV}(\mu_i, \sigma_i, \xi)$ and

$$\mu_i = x_{\mu,i}^T \beta_\mu + \epsilon_i^1, \quad \sigma_i = x_{\sigma,i}^T \beta_\sigma + \epsilon_i^2, \tag{3.2.2}$$

with $x_{\mu,i}, \beta_\mu \in \mathbb{R}^{p_\mu}$ and $x_{\sigma,i}, \beta_\sigma \in \mathbb{R}^{p_\sigma}$. The parameters β_μ and β_σ can be estimated with maximum likelihood. More complex structures can be used with the addition of link functions between the covariates and GEV parameters, or the use of semiparametric regression models, for which more details can be found in Ruppert et al. [2003]. For example, one can assume that a GEV parameter ϕ is of the form

$$\phi_i = f(x_i) + \epsilon_i, \tag{3.2.3}$$

with covariates x_i and smooth spline function $f(\cdot)$. If $x_i \in \mathbb{R}$, one can use a spline function of the form

$$f(x) = \sum_{i=1}^{l} \gamma_i b_i(x), \tag{3.2.4}$$

with $\{b_i(\cdot)\}_{i=1}^l$ a family of basis functions and $\{\gamma_i\}_{i=1}^l$ the set of coefficients in $\mathbb{R}$. Functions taking multiple inputs can be approximated with tensor product spline functions, such as (in the 2-dimensional case)

$$f(x, y) = \sum_{i=1}^{l_1} \sum_{j=1}^{l_2} \gamma_{i,j} b_i^1(x) b_j^2(y), \tag{3.2.5}$$

where $\{b_i^1(\cdot)\}_{i=1}^{l_1}, \{b_j^2(\cdot)\}_{j=1}^{l_2}$ are two families of basis functions. The choice of basis determines the type of spline function, and can be chosen among cubic and cyclical cubic spline basis,

polynomial basis, B-spline, Legendre polynomial basis, for example [Ruppert et al., 2003]. The models were implemented in `R` with the package `evgam` [Youngman, 2020].

A first approach is to assume the following model for the parameters, called *Model 1*:

$$\begin{aligned} \mu_{t,d,s} &= f^{cc}_{1,s,\mu}\{\text{month}(t)\} + f^{cs}_{2,s,\mu}(d) + \mu_s, \\ \log \sigma_{t,d,s} &= f^{cc}_{1,s,\sigma}\{\text{month}(t)\} + f^{cs}_{2,s,\sigma}(d) + \sigma_s, \\ \xi_{t,d,s} &= \xi_s, \end{aligned} \tag{3.2.6}$$

where $f^{cc}_{\dots}(\cdot)$ and $f^{cs}_{\dots}(\cdot)$ are respectively cyclical cubic and cubic spline functions and $\text{month}(t)$ corresponds to the month of the observation $M_{t,d,s}$. In words, month and duration effects are additive on the location and scale parameters, and the shape parameter only depends on the station identification, and is constant through the months and durations. One direct extension would allow the shape parameter to vary with month and the duration, as for the location and the scale. More specifically, we have the following models

$$\begin{aligned} \textit{Model 2} &: \xi_{t,d,s} = f^{cs}_{2,s,\xi}(d) + \xi_s \\ \textit{Model 3} &: \xi_{t,d,s} = f^{cc}_{1,s,\xi}\{\text{month}(t)\} + \xi_s \\ \textit{Model 4} &: \xi_{t,d,s} = f^{cc}_{1,s,\xi}\{\text{month}(t)\} + f^{cs}_{2,s,\xi}(d) + \xi_s, \end{aligned} \tag{3.2.7}$$

with the spline functions defined as before. One remark about these models is that for each duration, the maximum return level through over the year occurs during the same month. While this seems reasonable, some stations get higher rainfall maxima during December and January for long durations (typically 720 and 1440 minutes), and higher intensities during August and September for very small durations (5 to 60 minutes). As such, one might extend the previous models to allow for the effects of the duration to change with the month. This can be attained using spline tensor products, instead of regular spline functions. We propose the following model (called *Model 5*), with fixed shape parameter,

$$\begin{aligned} \mu_{t,d,s} &= f^{cc,cs}_{s,\mu}\{\text{month}(t), d\} + \mu_s, \\ \log \sigma_{t,d,s} &= f^{cc,cs}_{s,\sigma}\{\text{month}(t), d\} + \sigma_s, \\ \xi_{t,d,s} &= \xi_s, \end{aligned} \tag{3.2.8}$$

which then has the following extensions:

$$\begin{aligned} \textit{Model 6} &: \xi_{t,d,s} = f^{cs}_{2,s,\xi}(d) + \xi_s, \\ \textit{Model 7} &: \xi_{t,d,s} = f^{cc}_{1,s,\xi}\{\text{month}(t)\} + \xi_s, \\ \textit{Model 8} &: \xi_{t,d,s} = f^{cc}_{1,s,\xi}\{\text{month}(t)\} + f^{cs}_{2,s,\xi}(d) + \xi_s. \end{aligned} \tag{3.2.9}$$

In order to choose which model fits best, we used the BIC and AIC criteria and diagnostic plots. As noted in Coles [2001], for the GEV distribution, if one assumes that $Z_i \sim \text{GEV}(\hat{\mu}_i, \hat{\sigma}_i, \hat{\xi}_i)$, the transformed variables

$$\tilde{Z}_i = \frac{1}{\hat{\xi}_i} \log\left(1 + \hat{\xi}_i \frac{Z_i - \hat{\mu}_i}{\hat{\sigma}_i}\right) \tag{3.2.10}$$

have a standard Gumbel distribution, so that $\mathbb{P}(\tilde{Z}_i \leq z) = \exp\{-\exp(-z)\}$, for $z \in \mathbb{R}$. One can then construct the probability plot with the pairs

$$\{i/(m+1), \exp[-\exp\{\tilde{z}_{(i)}\}]; i = 1, \dots, m\}, \tag{3.2.11}$$

and ordered observations $\tilde{z}_{(1)}, \dots, \tilde{z}_{(m)}$. Similarly, the quantile plot is constructed with the pairs

$$\{\tilde{z}_{(i)}, -\log[-\log\{i/(m+1)\}]; i = 1, \dots, m\}. \tag{3.2.12}$$

We used the Kolmogorov–Smirnov [Birnbaum and Tingey, 1951] and Cramer–von Mises [Braun, 1980] tests to assess if the underlying sample distribution corresponds to the theoretical one, for each month, duration and station. Because of the high number of tests and their possible dependence, we used the Benjamini–Yekutieli procedure [Benjamini and Yekutieli, 2001] to control the false discovery rate, which can be applied to a wider range of problems compared to the Benjamini–Hochberg procedure [Benjamini and Hochberg, 1995]. We used a 5% level for our tests. *Model 1*, *Model 2* and *Model 3* are all nested within *Model 4*, and *Model 5*, *Model 6* and *Model 7* are all nested within *Model 8*, so one can also use the deviance statistic

$$D = 2\{l_1(M_1) - l_0(M_0)\} \tag{3.2.13}$$

for two models $M_0 \subset M_1$, where $l_1(M_1)$ and $l_0(M_0)$ are the maximized log-likelihood functions under models M_1 and M_0 respectively. Under the assumption that M_0 is adequate, the deviance statistic D follows a χ^2_k distribution, where k is the difference in the dimensionality of M_1 and M_0 [Coles, 2001].

For most stations, *Model 7* was preferred overall, although it did not get systematically the best BIC and AIC scores. This implies that duration does not have a significant impact on the shape parameter of the GEV distribution, unlike the month of the year. However, the duration and month have a significant influence on the location and scale parameters. The best fit for station 77001 is obtained with *Model 6* (implying that only the duration influences the shape), and for station 76026 it is *Model 5* (with constant shape). Examples of IDF curves for all six stations are shown in Figures B.2–B.7 (Appendix B), as well as the QQPlot for the corresponding model. Following the Benjamini–Yekutieli procedure with a 5% level, all the selected models for each station seems adequate.

The 20-year return levels for each month and durations can be found in Figure 3.2. One can note different behaviour of extrema depending on the duration and the month. For low to middle durations (5 to 60 minutes), return levels are higher between April to October, ranging from 66.93 mm/hr to roughly 140 mm/hr. The smallest return levels are observed from December to February. Station 76026, which is located in the north side of the Pacific coast, has the highest amplitude, reaching the lowest and the highest return levels among all the stations studied in this section. One can still observe this little summer phenomena for return levels, in particular during July. For large durations, there are major differences between stations: stations 77001, 75022, 69507 have their highest return levels during December and January, and their lowest in September. Notice that these stations are located on basins facing the Caribbean coast. On the other hand, stations 76026, 84118 and 98006 have a similar seasonal pattern as for small durations.

These results motivate a larger study of rainfall intensities for all stations of Costa Rica. It can be of great interest to directly incorporate the spatial characteristics of the stations directly into the modelling of the GEV parameters. This will be explored in the next Section.

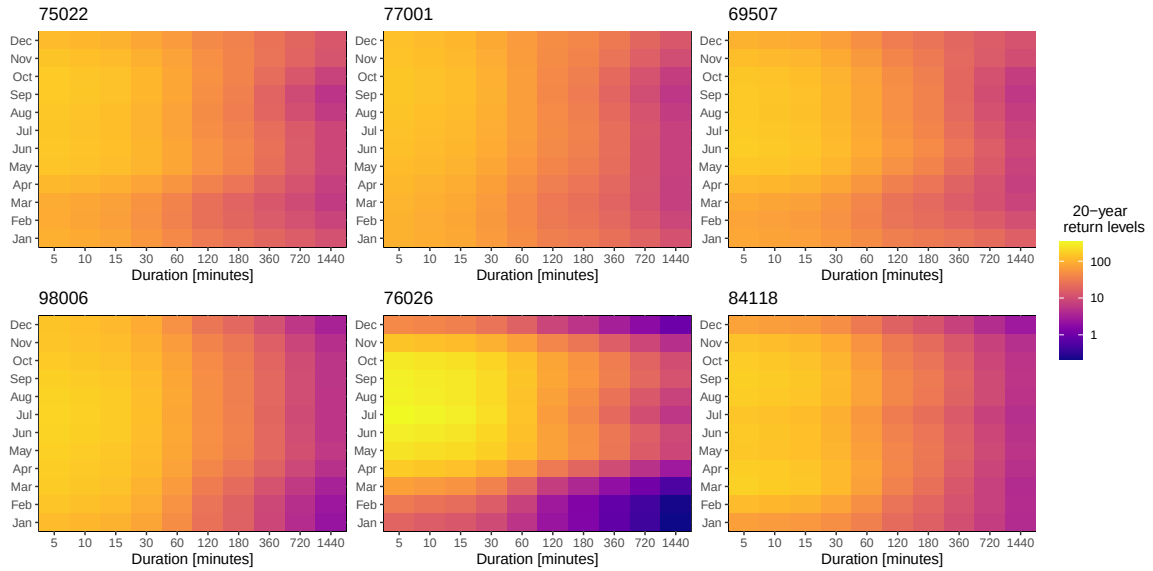


Figure 3.2: 0.95 quantiles for rainfall intensities (in [mm/hr]) for each month and duration, for the six stations.

3.3 Extension to all stations

The study of six stations across Costa Rica revealed important and different behaviour as regards to extreme rainfall intensities, which are greatly affected by a strong seasonality. This seasonality seems to differ depending on the duration and the location of the measurements. Indeed, the three stations facing the Caribbean coast are marked with a seasonality for which the peak occurs during different months depending on the duration, while the other stations have a seasonality pattern that is more consistent.

3.3.1 Extension of the GEV parameters with k-nearest neighbours

We first directly extend the work that was carried in Section 3.2 for all stations. Hence, we obtain one model for each station, selected using the same criteria as before. Note that, for efficiency, we fitted only *Model 5*, *Model 6*, *Model 7* and *Model 8*, and then selected the model that achieved the best BIC score among those that were not rejected during the Benjamini–Yekutieli procedure.

The 20-year return values are shown in Figure 3.3, grouped by basin, duration and month. The stations in basin 76 are marked by extremely high seasonality, reaching the highest and lowest return levels for small and middle durations. Their seasonality pattern is kept across all the durations. As they are also located on the north side of the Pacific coast, basins 74 and 78 share the same behaviour. The stations in basins 84 and 88 also have a seasonality, although with less amplitude. The stations in basin 69 are of particular interest, as they can be divided into two groups. The first consists of the stations that are the closest to basin 76, and thus to the Pacific coast. Those stations have rainfall intensities that peak during August and September, for all durations. The other group consists of the stations that are closest to the Caribbean coast, and for which monthly maximum intensities peak between December and February. This is what we typically observed for station 69507, which falls within this group.

One natural question that needs to be addressed is about extending the estimation of point-wise return levels related to some particular station, to any location in Costa Rica. This can

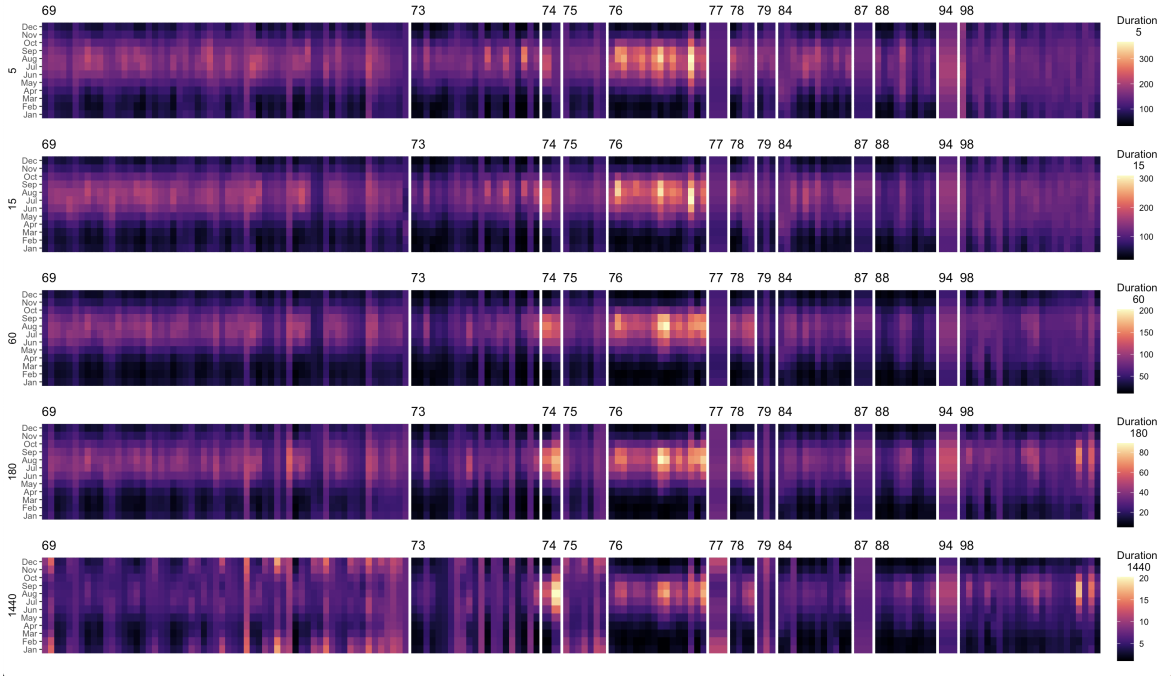


Figure 3.3: Fitted 20-year return levels for each station, duration and month.

be of major interest if one wants to obtain estimates for locations that are hard to access. One approach is to model the GEV and GPD parameters with spatial covariates. This will be explained in the next section. Another simple approach is to use simple regression tools to extend the quantiles of known location to any other point in the country, using spatial covariates. In particular, we used the non-parametric weighted k-nearest neighbours procedure of Dudani [1976], with $k = 2$. The hyper-parameter k was selected using cross validation from a range of potential values, from 1 to 50. The longitude, latitude and altitude of each stations were standardized and used to compute the Euclidean distance.

One example for all months and duration 1440 minutes is shown in Figure B.9. The north of the Pacific coast (and the province of Guanacaste) relies on very few stations for the inference, as the precipitation gauges are located more in the center of the country. The return levels are very low compared to the other regions at the beginning and the end of the year, and higher during August and September. The Caribbean coast has a seasonality that is much less marked, with rainfall return levels that are almost constant (for this particular duration of 1440 minutes) over the year, with small peaks in December and January. Another region of interest is the province of San José, which is characterized by lower return levels than any other region, for all months.

3.3.2 Estimating the GEV parameters with artificial neural networks

In this section, we propose to extend the univariate analysis of Section 3.2 by including the longitude, latitude and altitude as covariates for the GEV parameters. As highlighted by Ulrich et al. [2020], the use of spatial covariates for the GEV parameters can reduce the uncertainties of parameter estimation, in addition to having a unique model which encompasses all the stations. Ulrich et al. [2020] adapted vector generalized linear models [Van de Vyver, 2012; Peter et al., 2017] to produce IDF curves for multiple stations and durations. They proposed to model the location, scale and shape parameters for annual maxima with

orthogonal polynomials for the longitude and latitude as

$$l^{\phi}(\phi) = \phi_0 + \sum_{j=1}^{J} \alpha_{\phi,j} P_j(\text{long}) + \sum_{k=1}^{K} \beta_{\phi,k} P_k(\text{lat}) + \sum_{j=1}^{J} \sum_{k=1}^{K} \gamma_{\phi,j,k} P_j(\text{long}) P_k(\text{lat}), \tag{3.3.1}$$

with $l^{\phi}(\cdot)$ the link function associated with parameter $\phi \in \{\mu, \sigma, \xi\}$, $\alpha_{\phi,j}, \beta_{\phi,k}$ and $\gamma_{\phi,j,k}$ the regression parameters to be estimated and $P_j(\cdot)$ the orthogonal polynomials. They applied the ideas of Koutsoyiannis et al. [1998] to account for the duration, by setting

$$\begin{aligned} \sigma(d) &= \sigma_0 (d + \theta)^{-\eta}, \\ \mu(d) &= \mu_0 (d + \theta)^{-\eta}, \\ \xi(d) &= \xi_0, \end{aligned} \tag{3.3.2}$$

with $\theta > 0$ and $0 < \eta \leq 1$ parameters to estimate. As extreme rainfall might depend upon altitude and month, one needs to add them into equation (3.3.1), which can quickly lead to an explosion of the number of terms for the definition of the GEV parameters, especially when high order polynomials are needed. Also, one limitation of this approach is that the form of the interaction between covariates must be specified a priori. Another solution is to use Bayesian Hierarchical Models, as in Lehmann et al. [2013] and Dyrrdal et al. [2015], for which parameter uncertainties can be estimated directly. However, one still needs to specify the form of the interaction between covariates, as in the previous approach.

Cannon [2010] used a conditional density network (CDN) as a flexible nonlinear approach to model extreme precipitation data. CDNs extend the class of feedforward artificial neural networks (ANN) by modelling the conditional probability $p(y \mid x)$ of the conditioning and dependent variables x and y [Rothfuss et al., 2019]. In the following, we apply a similar approach by modelling the GEV parameters with a feedforward ANN with the longitude, latitude, altitude, month and duration as inputs. More details about ANN can be found in Bishop [2006], Schmidhuber [2014] and Goodfellow et al. [2016].

Let x denote the neural network input, which consists of the longitude, latitude, altitude, duration and the month. Months were transformed into a Fourier basis of order 2 and given as inputs, so that $x \in \mathbb{R}^8$. Increasing the order of the Fourier basis did not improve the results. The architecture of the neural network is shown in Figure F.1 (Appendix F). Increasing the number of layers and neurons within each layer can improve the model predictability. As shown in Telgarsky [2015], the number of neurons in a single hidden layer network g must increase exponentially as the number of layers D of a deep neural network f increases, hinting the benefits of increasing the depth of the neural network instead of its width. This can also improve test performance [Belkin et al., 2019], but comes at the cost of a model that is more complex to train, due to vanishing gradients and heterogeneous gradient amplitude (Bengio et al. [1994], Glorot and Bengio [2010]). In our case, the neural network consists of four hidden layers of dimensions respectively 100, 20, 20 and 10. The jth component of the kth layer is given by

$$a_j^k = l\left(b_j^k + \sum_{i=1}^{N_{k-1}} a_i^{k-1} w_{ij}^k\right), \quad k = 1, \ldots, K, \tag{3.3.3}$$

where N_k is the number of neurons on the kth layer and $w_{ij}^k, b_j^k \in \mathbb{R}$ the weights and bias. In matrix notation, this corresponds to

$$a^k = l\left(W^k a^{k-1} + b^k\right) \in \mathbb{R}^{N_k}, \quad k = 1, \ldots, K, \tag{3.3.4}$$

with $l(\cdot)$ the activation function applied component-wise. We used the sigmoid activation function for all layers except the last one, $a^K \in \mathbb{R}^3$. The parameters W^k, b^k, $k = 1, \ldots, K$ are

optimized by minimizing the negative log-likelihood of the GEV distribution given in 3.1.5, using the Adam stochastic optimizer [Kingma and Ba, 2015]. The output of the feedforward ANN is

$$\begin{aligned} \mu &= a_1^K \in \mathbb{R}, \\ \sigma &= \exp(a_2^K) \in \mathbb{R}_+^*, \\ \xi &= 2h(a_3^K)/3 - 0.5 \in [-0.5, 1.0], \end{aligned} \tag{3.3.5}$$

with $h(x) = \{1 + \exp(-x)\}^{-1}$ the sigmoid function. The constraint on ξ follows from Smith [1985] and avoids parameter explosion during extrapolation (which, for ξ, can greatly impact estimates of return levels). Another approach to constrain ξ would is described in Coles and Dixon [1999], with the addition of a penalty term directly in the likelihood.

Prediction intervals for neural network forecasts can be constructed with different techniques, including the delta, Bayesian, mean-variance and bootstrap methods. For a review of these techniques, see Khosravi et al. [2011]. In the following, we apply a bootstrapping method to create confidence bands, and more specifically residual bootstrapping (Khaliq et al. [2006], Kharin and Zwiers [2005]). The non-stationary GEV model is fitted to the original data y_i, and then the residuals are computed as

$$\epsilon_i = \left(1 - \xi_i \frac{y_i - \mu_i}{\sigma_i}\right)^{1/\xi_i}, \quad i = 1, \dots, N. \tag{3.3.6}$$

Then, we create a bootstrapped version $\tilde{y}_i$ of y_i as

$$\tilde{y}_i = \mu_i - \sigma_i \frac{{\epsilon_k}^{\xi_i} - 1}{\xi_i}, \quad i = 1, \dots, N, \tag{3.3.7}$$

with ϵ_k sampled with repetition from the set of residuals $\{\epsilon_i\}_{i=1}^N$, and the GEV model is fitted with the new set of observations $\{\tilde{y}_i\}_{i=1}^N$. This process is repeated until the desired number of bootstrapped samples is reached. In order to save computation time, the confidence intervals are created as follow: (i) the GEV model is first fitted with the entire set of observations, excluding the stations from the testing set. This model is then used as a starting point for the models created during the bootstrapping procedure. (ii) For $b = 1, \dots, B$, we create a new dataset by excluding some stations from the training set, and perform D residual bootstrapping iterations. (iii) This leads to $B \times D$ models, from which mean estimates and confidence bands can be produced. This process ensures that the models are not overfitting (by excluding stations from the original training set that consists of 80% of the total number of stations, see Chapter 2), and still allows for residual bootstrapping. One major drawback of this approach is the amount of models needed and stored.

3.3.3 Comparison of the two approaches

Examples of 20-year IDF curves for station 69507 and 84118 for January, May and September are presented in Figure 3.4. The IDF curves from the reference model is shown in yellow, and corresponds to the model selected in Section 3.2 for those two specific stations. The two approaches (namely the ones using k-NN and CDN) provide sensible return levels compared to the reference model, and QQplots based on the validation dataset indicate a reasonably good fit (Figures B.11 and B.12, Appendix B). Estimates of the return levels for high durations might however differ, for January. The confidence bands for the CDN approach are globally smaller than those of the two other models.

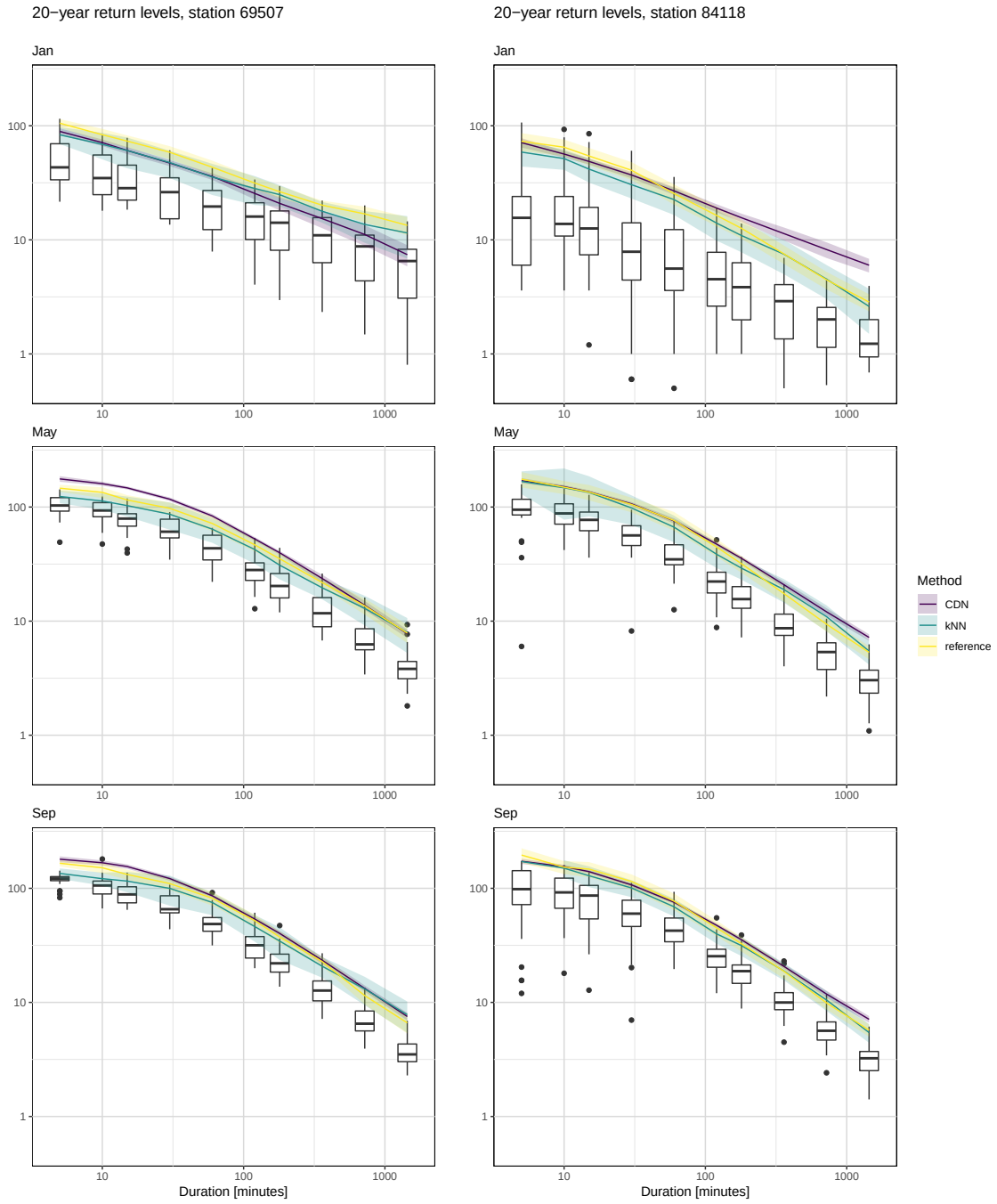


Figure 3.4: Example of IDF curves for stations 69507 and 84118, for January, May and September, computed with the reference model, the model with k-NN and the CDN approach. Positive values of QSI (in blue) indicate improvements compared to the second model, while negative ones (red) an improvement of the second to the first.

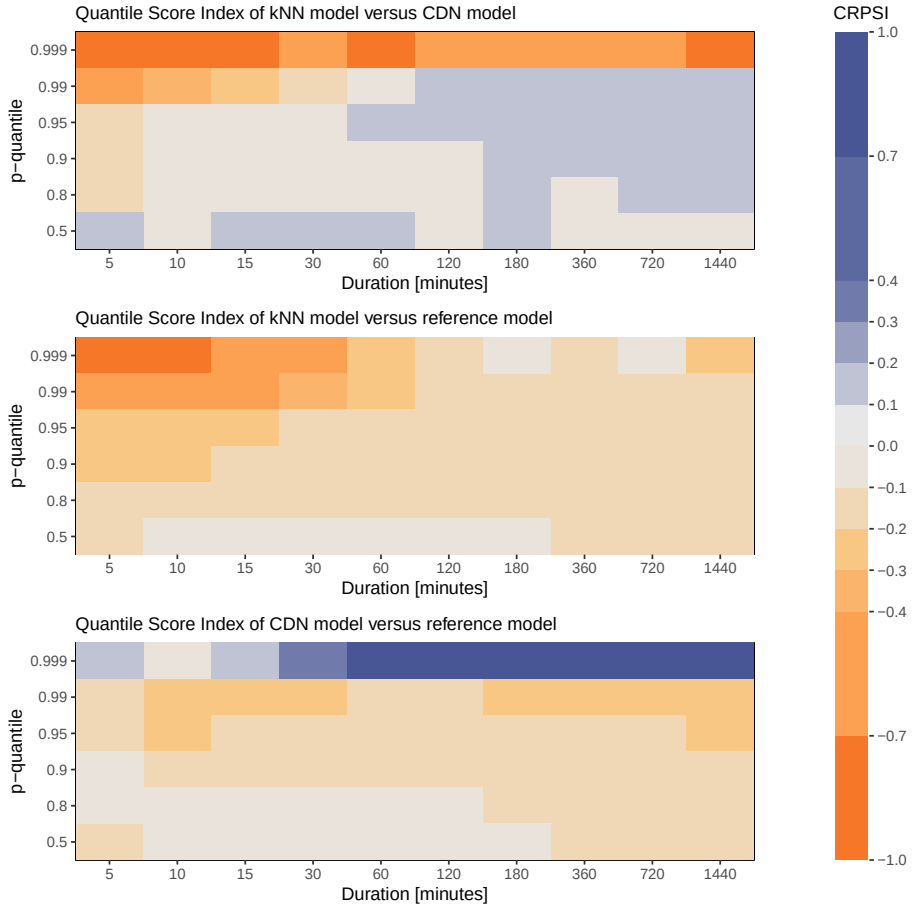


Figure 3.5: Expected QSI for several durations and quantiles, for the reference, k-NN and CDN models. The reference model corresponds to the seasonal GEV model from Section 3.2.

In order to have a better understanding of the strength of these methods, one can use the Quantile Score (QS) [Bentzien and Friederichs, 2014] and the Quantile Skill Index (QSI), for which more details and an application to the construction of IDF curves is presented in Ulrich et al. [2020]. The QS is defined as

$$QS(p) = \frac{1}{N}\sum_{i=1}^{N} \rho_p\{x_i - r(p)\}, \quad p \in (0,1) \tag{3.3.8}$$

with $\{x_i\}_{i=1}^N$ the set of observations, $r(p)$ the p-return level and $\rho_p(\cdot)$ the check function

$$\rho_p(x) = \begin{cases} px, & x \geq 0, \\ (p-1)x, & x < 0. \end{cases} \tag{3.3.9}$$

$QS(p)$ is non-negative for any $p \in (0,1)$, with optimal value 0. In practice, the return level may vary depending on the month m, the duration d and the location s, so that we will refer to $QS_{m,d,s}(p)$ with return levels $r_{m,d,s}(p)$. The Quantile Skill Score (QSS) compares the score of a model $\mathcal{M}$ with a reference model $\mathcal{R}$, and is defined as

$$QSS^{\mathcal{M}}(p) = 1 - \frac{QS^{\mathcal{M}}(p)}{QS^{\mathcal{R}}(p)}, \tag{3.3.10}$$

with $QSS^{\mathcal{M}}(p) > 0$ whenever model $\mathcal{M}$ performs better than the reference model $\mathcal{R}$. In order to maintain the symmetry and allow for a more interpretable score, especially when

the reference model performs better than model $\mathcal{M}$, one can define the Quantile Score Index (QSI)

$$QSI(p) = \begin{cases} QSS^{\mathcal{M}}(p), & QS^{\mathcal{M}}(p) \leq QS^{\mathcal{R}}(p), \\ -QSS^{\mathcal{R}}(p), & QS^{\mathcal{M}}(p) > QS^{\mathcal{R}}(p). \end{cases} \tag{3.3.11}$$

Hence, $QSI(p) \in [-1, 1]$, with positive values whenever model $\mathcal{M}$ performs better than the reference model $\mathcal{R}$. For our application, we decided to adjust the QSI to allow for the uncertainty of the return levels. Hence, let

$$EQS_{m,d,s}(p) = \mathbb{E}_{R(p)}\left[QS(p)\right], \tag{3.3.12}$$

by the expected quantile score, with return level $R_{m,d,s}(p)$ and distribution $F_{m,d,s,p}$. A straightforward estimator for $EQS_{m,d,s}(p)$ would come from the law of large number, with

$$\widehat{EQS}_{m,d,s}(p) = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} \rho_p\{x_i - r^b_{m,d,s}(p)\} \tag{3.3.13}$$

and return level samples $\{r^b_{m,d,s}(p)\}_{b=1}^{B}$. The expected QSS and QSI are defined likewise.

Figure 3.5 shows estimates of EQSI across the set of validation precipitation gauges, depending on the duration and the quantile. A similar panel is shown in Figure 3.5 (Appendix B) for the months and the quantiles. Overall, both approaches do not lose much information compared to the reference model, taking into account the fact that those stations were not used during training. As noted by Ulrich et al. [2020], the addition of spatial and seasonal covariates generally improves the modelling as a larger quantity of data is available for the estimation. This is particularly visible for high quantiles (typically 0.999), where the CDN approach provides a major improvement from the reference model, for all durations. The k-NN method however struggles for small durations (5 to 30 minutes) and large quantiles.

3.4 Discussion

The previous sections reviewed the extreme value theory and provided an application to the creation of IDF curves for a range of precipitation gauges across Costa Rica. The seasonality and durations effects had to be taken into account through different models that we compared in order the obtain the most accurate return levels. In order to expand the station-wise analysis to a global one, we compared the k-NN and CDN approaches, that can estimate the GEV parameters for any duration, month and location across Costa Rica. These analyzes helped us to identify several patterns with regards to extreme costarican precipitation. The Pacific coast is marked by a strong seasonality for monthly maximum rainfall, with a peak reached between May and October. The 20-year return levels for the dry season (November–March) are particularly low compared to the Caribbean coast, for all durations. A particular region of interest is the province of Guanacaste, which has higher estimated return levels between May and October than any other region of Costa Rica. However, heavy extrapolation is needed as only a few precipitation gauges are located in this region. The Caribbean coast, on the other hand, sees weaker seasonality, with a pattern that differs depending on the duration. Indeed, high duration are characterized by return levels that peak during the dry season, between November and February.

As shown in the exploratory analysis, the number of observations per month can vary significantly, with an overall decrease of available data between December and March. This coincides with the dry season in Costa Rica, and stations that face the Caribbean coast usually tend to have more data during this period. The lack of data is largely due to the fact

that no rainfall was observed for these particular days, and could be interpreted as 0 values. One might want to take into account the block length directly into the GEV modelling. Let α be the ratio of the current block length (that is, the number of observation used to compute the maximum) over the maximum block length (for monthly maxima, 30 or 31). Then, one has

$$\exp\left[-\alpha\left(1+\xi\frac{y-\mu}{\sigma}\right)_+^{-1/\xi}\right] = \exp\left[-\left\{1+(\alpha^{-\xi}-1)+\xi\frac{y-\mu}{\sigma\alpha^{\xi}}\right\}_+^{-1/\xi}\right] \tag{3.4.1}$$

$$= \exp\left[-\alpha\left(1+\xi\frac{y-\mu+\sigma\frac{1-\alpha^{\xi}}{\xi}}{\sigma\alpha^{\xi}}\right)_+^{-1/\xi}\right] \tag{3.4.2}$$

$$= \exp\left[-\alpha\left(1+\xi\frac{y-\tilde{\mu}}{\tilde{\sigma}}\right)_+^{-1/\xi}\right], \tag{3.4.3}$$

with $\tilde{\mu} = \mu - \sigma(1-\alpha^{\xi})/\xi$ and $\tilde{\sigma} = \sigma\alpha^{\xi}$. This proves that the GEV distribution $F(y;\mu,\sigma,\xi)$ satisfies the max-stability relation $F(y;\mu,\sigma,\xi)^t = F(y;\mu_t,\sigma_t,\xi_t)$ for $t>0$, with the shape parameter ξ that remains unchanged. Hence, one can estimate the parameters μ, σ and ξ by maximizing the log-likelihood

$$l(\mu,\sigma,\xi) = -\sum_{i=1}^{K}\log\tilde{\sigma}_i - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{K}\log\left[1+\xi\frac{z_i-\tilde{\mu}_i}{\tilde{\sigma}_i}\right] - \sum_{i=1}^{K}\left[1+\xi\frac{z_i-\tilde{\mu}_i}{\tilde{\sigma}_i}\right]^{-1/\xi}, \tag{3.4.4}$$

with $\tilde{\mu}_i = \mu - \sigma(1-\alpha_i^{\xi})/\xi$ and $\tilde{\sigma}_i = \sigma\alpha_i^{\xi}$, where α_i is the ratio of the number of observations to compute z_i over the maximum block length.

Another possibility to deal with the lack of data for these months is to consider the POT approach, described in Section 3.1.2. Due to the seasonality, a time-varying threshold might need to be estimated, based on a pre-defined high quantile [Coles, 2001]. This can be achieved by fitting an asymmetric Laplace distribution to estimate the threshold, and then propagating the estimates and the standard errors for the estimation of the GPD parameters. The POT approach shows similar qualitative and quantitative results as the block maxima approach, and was not considered for the extension process to any location. This approach benefits from a larger amount of available data for the fitting, but it introduces possible dependence between observations (as we consider daily intensities). Parametric models to account for this exist (Coles [2001], Davison [2019]), and the CDN-based approach of Section 3.3.2 might need to be adapted.

In this chapter, rainfall intensities were considered independent across the stations, even for the extended models. In practice, two very close stations are likely to observe extreme events for the same periods, but the extent to which this occurs may vary depending on the duration and the location of the stations. In the next chapter, we provide a review of bivariate models in the context of extreme value theory, and apply it to estimate and study the extremal dependence between any precipitation gauge.

Chapter 4

Bivariate analysis of extreme rainfall

This chapter and the following concentrate on the study of the relationship between extreme precipitation of two or more rain gauges in Costa Rica. In this chapter, we review the theory that revolves around models for multivariate extremes, in a bivariate framework. A major advantage of using a multivariate model is that rainfall data from one station can be used for inference made at another station. This is especially useful for regions with very few precipitation gauges.

The second part of this chapter explores the application of bivariate modeling for the six stations studied in the previous sections. The goal here is to identify asymptotic independence or dependence between each pair of stations, and extend it to any two locations in Costa Rica.

4.1 Theory

In the following, we refer to Coles [2001] and Davison [2019] for a review of bivariate models and multivariate extreme value theory.

Assume that we are given a sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ of independent and identically distributed vectors with distribution function $F(x, y)$. If we denote by $M_{X,n} = \max_{j=1,\ldots,n} X_j$ and $M_{Y,n} = \max_{j=1,\ldots,n} Y_j$, we can define the vector of component-wise maxima

$$M_n = (M_{X,n}, M_{Y,n}). \tag{4.1.1}$$

Our focus is to study the behavior of M_n as $n \to \infty$. Note that M_n may not correspond to an actual observation, as maxima are taken component-wise (for which, separately, the univariate theory from Section 3.1 applies). We then have the following theorem:

Theorem 4.1.1. *Let (Z_1, Z_2) be the linearly rescaled component-wise maxima of n independent vectors (X_i, Y_j), transformed to have limiting unit Fréchet marginal distributions. If*

$$\lim_{n\to\infty} \mathbb{P}(Z_1 \leq z_1, Z_2 \leq z_2) = H(z_1, z_2), \quad z_1, z_2 > 0, \tag{4.1.2}$$

where H is a non-degenerate distribution function, then

$$H(z_1, z_2) = \exp\{-V(z_1, z_2)\}, \quad z_1, z_2 > 0, \tag{4.1.3}$$

where the function V is called the exponent measure, and we can write

$$V(z_1, z_2) = 2\int_0^1 \max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right) Q(\mathrm{d}w) = 2\mathbb{E}\left\{\max\left(\frac{W}{z_1}, \frac{1-W}{z_2}\right)\right\}, \tag{4.1.4}$$

where $W \sim Q$, an angular distribution function on $[0,1]$ such that

$$\mathbb{E}(W) = \int_0^1 wQ(\mathrm{d}w) = 1/2. \tag{4.1.5}$$

If Q is differentiable with angular density function q, then

$$V(z_1, z_2) = 2\int_0^1 \max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right) q(w)\,\mathrm{d}w. \tag{4.1.6}$$

The theorem requires the marginals Z_1 and Z_2 to have limiting Fréchet distribution, $F(z) = \exp(-1/z)$, for $z > 0$. The exponent measure is homogeneous of order -1, that is, it satisfies the property

$$tV(tz_1, tz_2) = V(z_2, z_2), \quad t > 0, \tag{4.1.7}$$

which implies that

$$H^t(tz_1, tz_2) = H(z_1, z_2). \tag{4.1.8}$$

This relates to the notion of max-stability that we will see in Section 5. One can derive two limiting cases from Theorem 4.1.1:

- independent unit Fréchet variables Z_1 and Z_2. In this case, for $z_1, z_2 > 0$, $H(z_1, z_2) = \exp\{-(1/z_1 + 1/z_2)\}$, which corresponds to a measure Q with masses $1/2$ on $w = 0$ and $w = 1$.
- perfectly dependent, $Z_1 = Z_2$ unit Fréchet, so that the measure Q places unit mass on $w = 1/2$, and $H(z_1, z_2) = \exp\{-\max(1/z_1, 1/z_2)\}$, for $z_1, z_2 > 0$.

Because generating parametric families with parameter-free mean and tractable integral for the exponent measure is usually a difficult task (see Coles [2001]), one often relies on standard models that constitute sub families of the set of non-degenerate distribution functions H that satisfy Theorem 4.1.1. One particular case is the logistic dependence function [Gumbel, 1960],

$$H(z_1, z_2) = \exp\left\{-(z_1^{-1/r} + z_2^{-1/r})^r\right\}, \quad z_1, z_2 > 0,\ 0 < r \leq 1. \tag{4.1.9}$$

As $r \to 1$, we have the independent case, and as $r \to 0$, the perfectly dependent one. Other examples of models include the Hüsler–Reiss model [Hüsler and Reiss, 1989],

$$H(z_1, z_2) = \exp\left\{-\frac{1}{z_1}\Phi\left(r^{-1} + \frac{r}{2}\log\frac{z_2}{z_1}\right) - \frac{1}{z_2}\Phi\left(r^{-1} + \frac{r}{2}\log\frac{z_1}{z_2}\right)\right\}, \quad r > 0, \tag{4.1.10}$$

and the negative logistic [Galambos, 1975] model, with

$$H(z_1, z_2) = \exp\left\{-\frac{1}{z_1} - \frac{1}{z_2} + (z_1^r + z_2^r)^{-1/r}\right\}, \quad r > 0. \tag{4.1.11}$$

As $r \to 0$, we have the independent case, and as $r \to \infty$, the perfectly dependent one.

4.1.1 Model fitting

In practice, one has access to a series $(x_1, y_1), \ldots, (x_n, y_n)$ of independent vectors, that are used to create the component-wise block maxima sequence $(z_{1,1}, z_{2,1}), \ldots, (z_{1,m}, z_{2,m})$. Because the margins are often not distributed according to a Fréchet distribution, one needs to transform the variable in order to use Theorem 4.1.1. Suppose that

$$Z_{k,j} \sim \mathrm{GEV}(\mu_{(k,j)}, \sigma_{(k,j)}, \xi_{(k,j)}), \quad j = 1, \ldots, m, \quad k = 1, 2. \tag{4.1.12}$$

The transformed variables

$$\tilde{Z}_{k,j} = \left[1 + \xi_{(k,j)} \left(\frac{Z_{k,j} - \mu_{(k,j)}}{\sigma_{(k,j)}} \right) \right]^{1/\xi_{(k,j)}} \tag{4.1.13}$$

are then distributed according to the standard Fréchet distribution. In practice, $\mu_{(k,j)}$, $\sigma_{(k,j)}$ and $\xi_{(k,j)}$ are replaced by their estimates. The bivariate model is then fitted using maximum likelihood, which is obtained by differentiating $H(z_1, z_2) = \exp(-V(z_1, z_2))$, giving

$$f_H(z_1, z_2) = \left\{ \frac{\partial V(z_1, z_2)}{\partial z_1} \frac{\partial V(z_1, z_2)}{\partial y} - \frac{\partial^2 V(z_1, z_2)}{\partial z_1 \partial z_2} \right\} \exp\{-V(z_1, z_2)\}. \tag{4.1.14}$$

Stephenson and Tawn [2005] emphasise that only one of the leading term in (4.1.14) contributes to the likelihood: if both maxima occur at the same event, we keep the second term and the first one otherwise.

4.1.2 Asymptotic dependence and independence

One might be interested in the asymptotic behaviour of $\mathbb{P}(Z_1 \leq z, Z_2 \leq z)$, with a common $z > 0$. Note that, because of the homogeneity of the exponent measure V,

$$\mathbb{P}(Z_1 \leq z, Z_2 \leq z) = \exp\{-V(z, z)\} = \exp\{-V(1, 1)/z\} = \exp(-1/z)^{V(1,1)}, \quad z > 0. \tag{4.1.15}$$

Hence, one can define a measure of dependence with what is called the extremal coefficient, defined as $\theta = V(1, 1)$. Note also that

$$\lim_{n \to \infty} \mathbb{P}(Z_1 > z \mid Z_2 > z) = 2 - \theta, \tag{4.1.16}$$

so that the extremal coefficient can be interpreted as the conditional probability that one marginal exceeds some z given that the other has exceeded z.

One critical point of this definition is that it relies on some model for the choice of the exponent measure V, and therefore, one can introduce a more general setting in which we suppose that (X, Y) is a bivariate random vector with marginal distribution functions F_X and F_Y. We can then write

$$F(X, Y) = \mathbb{P}(X \leq x, Y \leq y) = C\{F_X(x), F_Y(y)\}, \tag{4.1.17}$$

where C is the copula corresponding to F. If we are given F, F_X and F_Y, then $C(u, v) = F\{F_X^{-1}(u), F_Y^{-1}(v)\}$, for $0 < u, v < 1$. We can then focus on the probability that X and Y are equally extreme, and define

$$\chi = \lim_{u \to 1} \mathbb{P}\{F_Y(Y) > u \mid F_X(X) > u\}, \quad 0 < u < 1. \tag{4.1.18}$$

Notice that $\mathbb{P}\{F_Y(Y) > u, F_X(X) > u\} = 1 - 2u + C(u, u)$, so that

$$\mathbb{P}\{F_Y(Y) > u \mid F_X(X) > u\} \to \frac{1 - 2u + C(u, u)}{1 - u}, \quad u \to 1. \tag{4.1.19}$$

Hence,

$$\chi(u) = 2 - \frac{\log C(u, u)}{\log u}, \quad 0 < u < 1, \tag{4.1.20}$$

where we used the fact that as $u \to 1$, $1 - u \approx -\log u$ and $1 - C(u, u) \approx -\log C(u, u)$, so that $\chi = \lim_{u \to 1} \chi(u)$.

If $C(u,u)$ corresponds to the extreme value copula, so that $F_X^{-1}(u) = -1/\log(u)$, we have

$$C(u,u) = \exp\{-V(-1/\log u, -1/\log u)\} = \exp\{V(1,1)\log u\} = u^{V(1,1)} = u^{\theta}, \quad (4.1.21)$$

with $\theta = V(1,1)$, as in the previous setting. Hence, $\chi(u) = 2 - \theta$.

Here are some properties of the asymptotic coefficient χ:

- $0 \leq \chi \leq 1$;
- if $C(u,u)$ corresponds to an extreme value copula, $\chi = 2 - \theta$;
- if $\chi = 0$, we have the asymptotic independence of X and Y;
- within the class of asymptotically dependent variables, the value of χ increases with strength of dependence at extreme levels [Coles, 2001].

Although χ can provide a useful interpretation of asymptotic dependence among two variables, it does not provide any information about the type of asymptotic independence two variables can have, as we can only have $\chi = 0$ in this case. Alternatively, one can define

$$\bar{C}(u,u) = \mathbb{P}\{F_X(X) > u, F_Y(Y) > u\}, \quad 0 < u < 1, \quad (4.1.22)$$

and

$$\bar{\chi}(u) = \frac{2\log(1-u)}{\log \bar{C}(u,u)} - 1. \quad (4.1.23)$$

We then have that:

- if X and Y are independent, then $\bar{C}(u,u) = (1-u)^2$ and $\bar{\chi}(u) = 0$;
- if X and Y are perfectly dependent, $\bar{C}(u,u) = (1-u)$ and $\bar{\chi}(u) = 1$;
- $-1 \leq \bar{\chi}(u) \leq 1$, and $\bar{\chi}(u)$ grows with increasing dependence;
- if X and Y are asymptotically dependent, $\bar{\chi}(u) \to 1$ as $u \to 1$.

4.2 Application: bivariate analysis of extreme rainfall

4.2.1 Case study of six stations

We applied the theory of bivariate modelling described above to the study of dependencies between the six stations analyzed in Section 3.2. The rainfall intensities from the 15 possible pairs of stations were fitted with different dependence functions. In particular, we first transformed the monthly maxima to have a Fréchet distribution, using the selected models from Section 3.2. We obtain transformed variables that depend on the month, the station and the duration. We fitted different dependence models, including the logistic, negative logistic, bilogistic, Hüsler–Reiss, asymmetric logistic and Coles–Tawn dependence functions. For each pair of stations, the model that was kept was the one achieving the best BIC score.

Figure 4.1 (Appendix C) shows the variation of the dependence coefficient $2 - V_{x_1,x_2}(1,1)$ across the durations, for each pair of stations (x_1, x_2). The probability that the transformed rainfall of one station exceeds some value u given that the other exceeds this value ranges from 0 to 0.5. One can observe a tendency for close stations to have higher dependence measure, although the topography of the stations seems to greatly impact the results. Indeed, stations 75022, 77001 and 69507 have a dependency that increases greatly as the duration grows, while

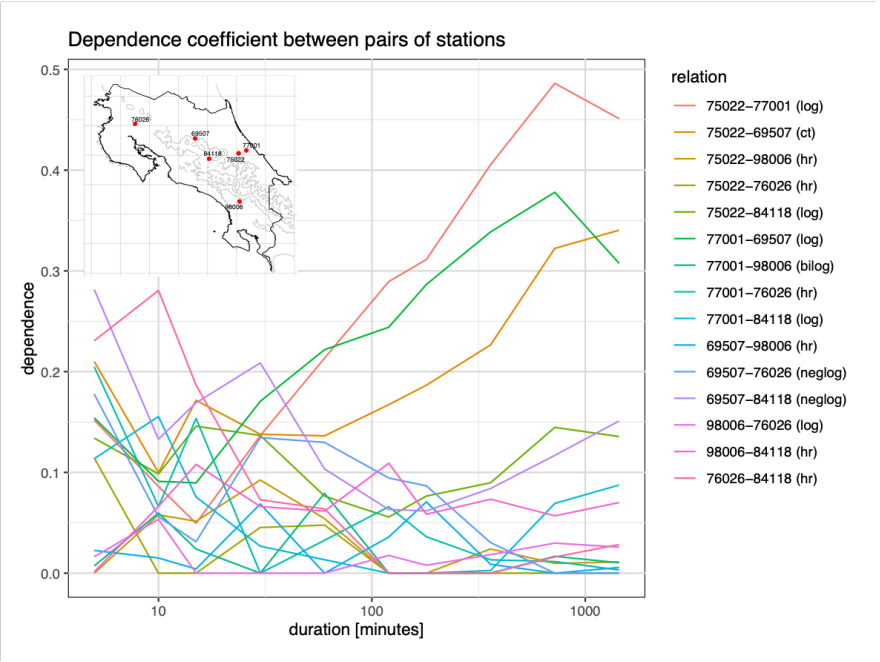


Figure 4.1: Measure of the dependence $2 - V_{x_1,x_2}(1,1)$ for rainfall intensities for each pair of station (x_1, x_2). The confidence bands are not shown for clarity, but only the pairs (77001–75022), (69507–75022) and (77001–69507) have confidence intervals that exclude the value 0.

the other pairs show no significant dependence. Even station 84118 does not have a strong dependence with those stations, while still being relatively close. However, it is protected from the Caribbean coast by multiple mountains, which might explain this type of behaviour. The dependence structure for low durations is more chaotic and hardly interpretable.

Some measures of χ and $\bar{\chi}$ for the pairs 75022–77001 and 75022–98006 are shown in Figure C.1 (Appendix C). The extremal coefficients start to be significantly different from 0 for the pair 75022–77001 only for very large durations, which correspond to large scale rainfall events. This is coherent with the results obtained with the parametric models above.

4.2.2 Extending the extremal dependence measure to any location

In this section we attempt to extend the previous bivariate model to accommodate the heterogeneous behaviour of extremal dependence across Costa Rica. In Section 4.2.1, we fitted multiple bivariate distributions for a number of stations which were specifically studied. By fixing one station and changing the other station, the extremal coefficient and the parameters of the bivariate model change with the location. However, this behaviour also depends on the station that is fixed, hinting at non-stationarity of the dependence structure of extreme rainfall. A first approach consists in incorporating spatial covariates into the model parameters. As an example, assume that the station 69507 is fixed, and consider the behaviour of the extremal dependence with respect to this station across Costa Rica. One could use the Hüsler–Reiss model defined in Section 4.1 with bivariate distribution

$$\mathbb{P}\{Z(x_1) \leq z_1, Z(x_2) \leq z_2\} = \exp\left\{-\frac{1}{z_1}\Phi\left(r^{-1} + \frac{r}{2}\log\frac{z_2}{z_1}\right) - \frac{1}{z_2}\Phi\left(r^{-1} + \frac{r}{2}\log\frac{z_1}{z_2}\right)\right\}, \tag{4.2.1}$$

but with a location-varying parameter $r \equiv r_{x_1,x_2}$. $Z(x_1)$ and $Z(x_2)$ are the transformed variables at locations x_1 and x_2 with unit Fréchet margins, following Equation (4.1.13). One

can for example define r with spatial covariates, such as

$$r_{x_1,x_2} = \exp(X_{x_1,x_2}^T \beta), \tag{4.2.2}$$

where X_{x_1,x_2} corresponds to covariates such as the longitude, latitude and elevation, and β a vector of parameters that needs to be estimated. One can use the tools described in Section 5.1.3 to fit the model. Here x_1 is fixed and only x_2 changes, so that the model is well-suited for a detailed study of the potential dependence from one station with at least some historical measurements to any other location. However, it does not generalize to any other location (of reference).

One can also use a more complex structure for the parameter r, using spline functions or spline tensors to add cross effects between the different covariates, for example. In our particular case, we decided to approximate r with a deep artificial neural network, taking as input the longitude, latitude, elevation and durations, and which outputs the parameter r. The architecture is shown in Figure F.2 (Appendix F). The loss function corresponds to the negative log-likelihood based on the bivariate density of the Hüsler–Reiss model, and we used a stochastic gradient descent approach with batch size $B = 2000$ to speed up the computation and avoid being stuck at local maxima.

Figure 4.2 shows the extremal coefficient for the model with station of reference 69507, for 1440-minute durations. During the training, durations were separated, and the model did not learn to approximate the extremal dependence of two stations with different durations. One can observe a clear pattern of dependence with stations facing the Caribbean coast and with similar elevations. This includes the stations 77001 and 75022 that were studied before, and for which the extremal dependence appeared to be strong for high durations. This model allows for change of the dependence structure across the map, as for example locations in the southern part of Costa Rica with similar altitude still have a low dependence with station 69507, as shown by station 98034. For the Hüsler–Reiss model, the extremal coefficient is $2\Phi(1/r)$. The bivariate joint survival probability,

$$\mathbb{P}\{Z(x_1) > z, Z(x_2) > z\} = 1 - 2\exp(-1/z) + \exp\{-V(1,1)/z\}, \tag{4.2.3}$$

can be used to assess if extreme events at two locations are likely to occur at the same time. This quantity is shown in Figure 4.3 for various return periods and the four stations shown in green in Figure 4.2. Station 69620 exhibits high dependence with station 69507 compared to the others, with a dependence that increases as the duration grows. For most stations, the model provided a reasonable approximation for the joint survival probability, though it struggles for the independent case, as illustrated with station 98034. The fit tends to deteriorate for lower durations, for which the confidence bands do not manage to envelop the empirical estimates of the joint probability. A comparison can be made with an empirical estimator for the extremal coefficient $V(1,1)$. Properties and estimators of the extremal coefficient $V(1,1)$ can be found in Schlather and Tawn [2003] and Ferreira [2018]. In the following, we will use the estimator

$$\hat{V}(1,1) = \left[\frac{1}{N}\sum_{i=1}^{N} \min\{z_i(x_1)^{-1}, z_i(x_2)^{-1}\}\right]^{-1}, \tag{4.2.4}$$

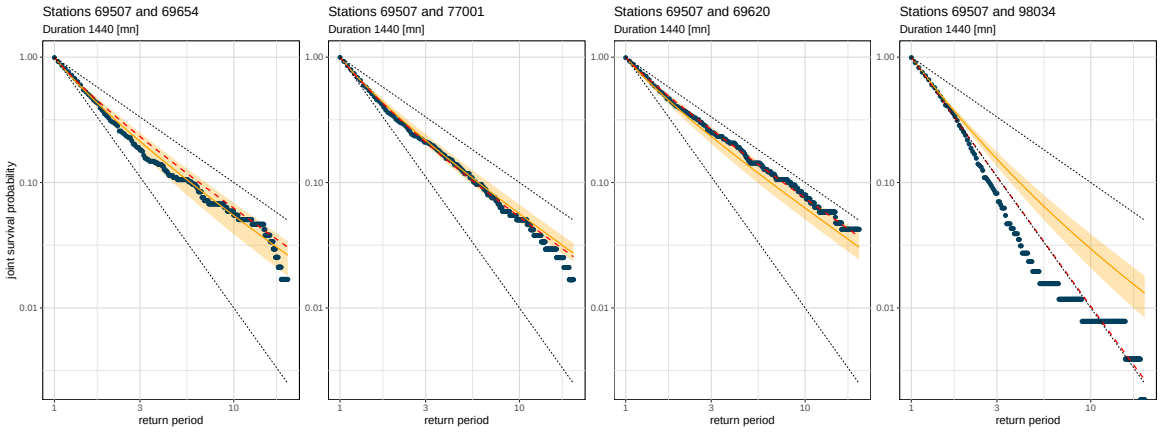
which follows from the fact that $1/\max\{Z(x_1), Z(x_2)\} \sim \text{Exp}\{V(1,1)\}$ (see Ribatet [2009] for more details). The empirical estimates are shown as dotted red lines in Figure 4.3, and generally agree with the fitted extremal coefficient. When two stations exhibit total independence (like the pair 69507–98034), the empirical estimate generally performs better.

The dependence structure also varies depending on the duration. Figure 4.2 shows the difference of extremal coefficients from the largest duration (1440 minutes) to the lowest (5 minutes) across Costa Rica (again with station 69507 as reference). The same pattern as before occurs. In particular, the stations with high dependence for duration 1440 minutes have a lower dependence for smaller durations, and this phenomena is reversed for locations with independent behaviour. This agrees with the observations from Section 4.2.1, and especially Figure 4.1.

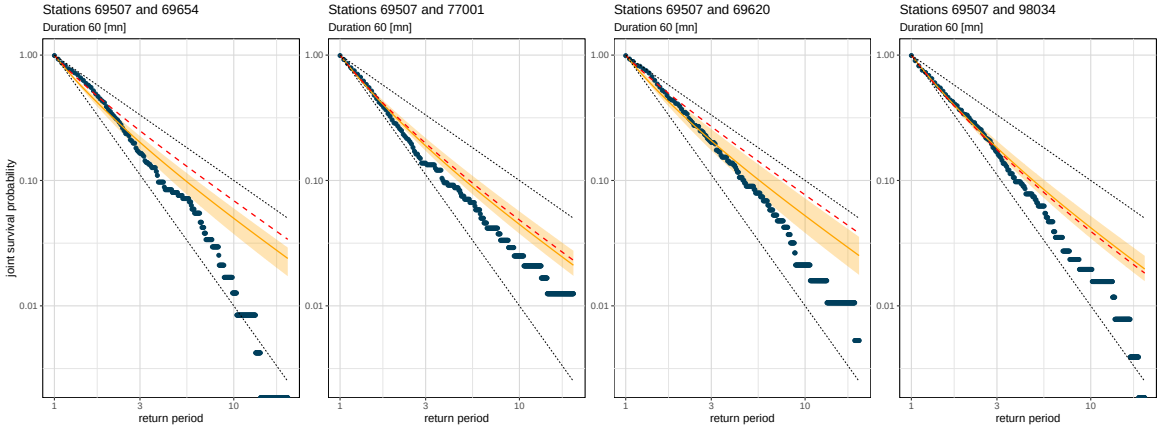
Estimates and standard errors of the extremal coefficient for any location and with respect to stations 69507, 77001, 75022, 84118, 98006 and 76026 are shown in Figures C.2 and C.3 (Appendix C). Stations 77001 and 75022 show a similar pattern with respect to the behaviour of the extremal coefficient as station 69507 presented here, which shows the overall dependence of the stations facing the Caribbean coast. The other three stations (that face the Pacific coast) have weaker extremal dependence, although it is still significant.

The modelization of extreme precipitation through the lens of bivariate extreme value theory allows to extend the univariate GEV framework to estimate asymptotic dependence or independence between any two stations. By fitting some of the bivariate families described in Section 4.1 on each pair of station among the six studied in details, we managed to discover interesting patterns with regards of extreme rainfall, which includes the presence of a strong asymptotic dependence at long durations between the stations facing the Caribbean coast. This observation is confirmed by the use of an extended bivariate model, which incorporates the variability of extremal dependence through the use of a conditional neural network. The fitted extremal coefficients coincide with their empirical measures, as highlighted by the estimates of the joint survival probability. However, this approach struggles to model total independence. Stations facing the Pacific coast show asymptotic dependence between themselves, revealing a Pacific/Caribbean demarcation. This will be highlighted in more detail in the clustering analysis of Section 5.2.1 (Chapter 5).

Even if this approach allows to quantify to which extent some high precipitation levels observed at a particular station can also be observed at any other location in Costa Rica, it nevertheless involves to fit one model per precipitation gauge. An extension to allow for any two location was undertaken, but did not lead to any useful results. This motivates the models presented in the next chapter, which offer a better understanding of the behaviour of extremal dependence across Costa Rica.



(a)



(b)

Figure 4.3: Joint survival probability plots of the extended bivariate model for four pairs of stations related to station 69507. The dark points show the empirical estimates for the joint probability, the yellow curve the estimates from the extended bivariate model, and the dotted red one the empirical estimate based on equation (4.2.4). The upper and lower dotted grey lines respectively correspond to the cases $V(1,1) = 1$ (total dependence) and $V(1,1) = 2$ (independence).

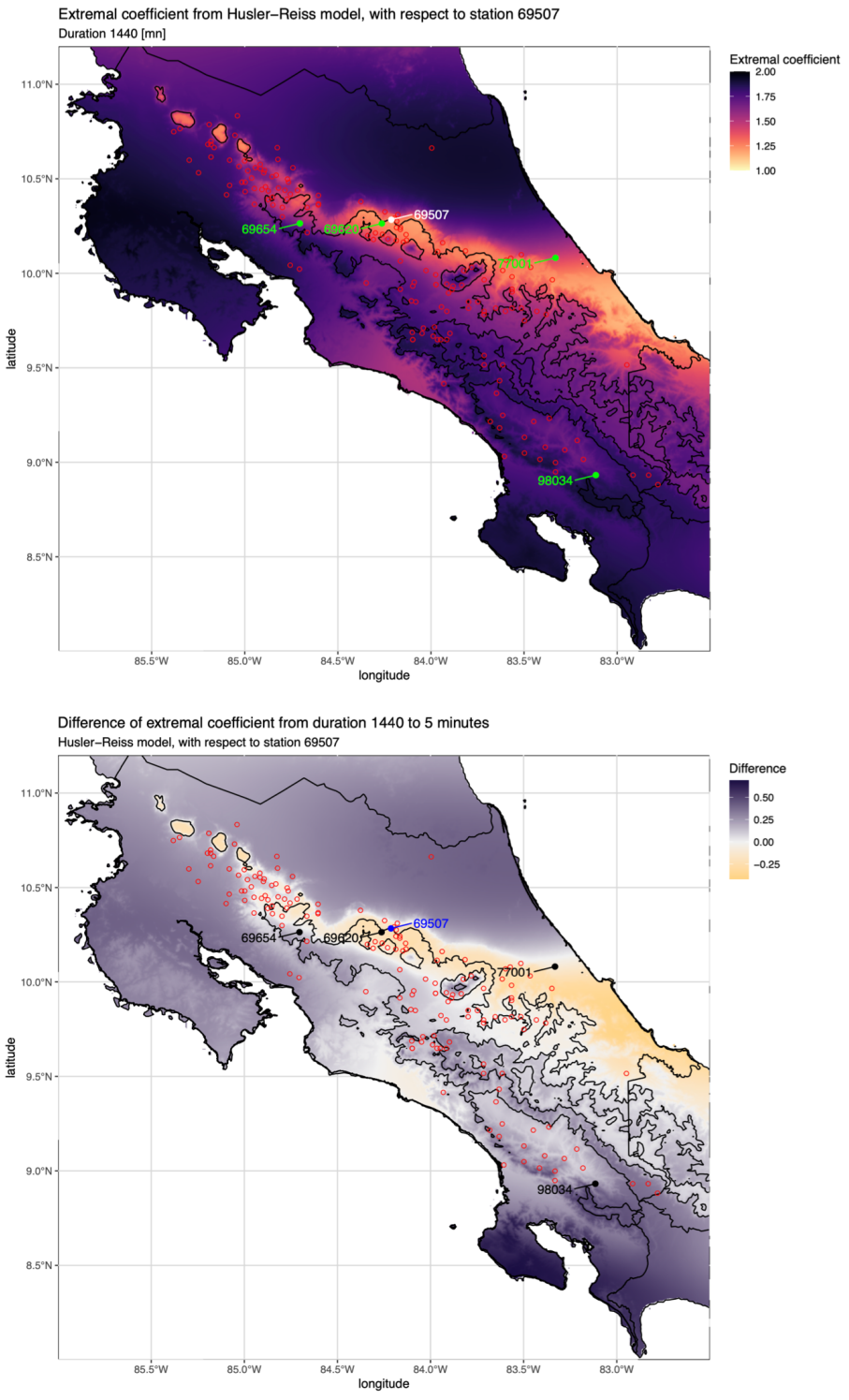


Figure 4.2: (top) Extremal coefficient from Husler-Reiss model with varying spatially-parameter learned with an artificial neural network. The station of reference is 69507, and the stations in green are the ones used to compute the joint survival probability of Figure 4.3. (bottom) Difference of value of the extremal coefficient from the largest duration (1440 minutes) to the lowest (5 minutes) across Costa Rica, with station 69507 as reference. Negative values correspond to the case where the dependence increases as the duration increases, and positive ones to the opposite case.

Chapter 5

Multivariate analysis with max-stable processes

The multivariate extreme value theory is not a direct extension of its univariate counterpart, as one needs to define what is an extreme in a multivariate framework. Pioneered by de Haan [1984], de Haan and Pickands [1986] and Resnick [1987], max-stable processes provide a tractable family to generalize multivariate extreme value theory for spatial extremes. This approach is particularly interesting when the dimension is high, as is the case when many stations are studied. In this chapter, we give an overview of the theory of max-stable processes and apply it to our study of extreme precipitation in Costa Rica. A major difficulty that may arise is the presence of complex weather systems and local effects that could alter the performance of max-stable models.

A first step will be dedicated to the identification of relevant regions based on measure of asymptotic dependence. This serves two purposes: identifying stations that share common patterns with respect to extreme precipitation, and modeling the behaviour of asymptotic dependence with max-stable processes within each region. The latter could produce problems at the boundaries of the regions and difficulties to estimate the uncertainty, as two steps are required.

Another approach discussed at the end of this chapter is to use non-stationary max-stable processes, which could directly incorporate the variability of the dependence structure through the model parameters. This approach will then be compared to the extended bivariate model presented in Chapter 4.

5.1 Theory

In the following, we review max-stable processes, especially in the context of spatial extremes. A stochastic process $\{Z(x), x \in \mathcal{X}\}$ is said to be max-stable if there exist sequences $A_{N_x} > 0$ and B_{N_x} for $N \geq 1$ and $x \in \mathcal{X}$, such that if $Z^{(1)}(x), \ldots, Z^{(N)}(x)$ are N independent copies of the process, and if one defines $\{Z^*(x), x \in \mathcal{X}\}$ as

$$Z^*(x) = \frac{\max_{1 \leq i \leq N} Z^{(i)}(x) - B_{N_x}}{A_{N_x}}, \quad x \in \mathcal{X}, \tag{5.1.1}$$

then $\{Z^*(x), x \in \mathcal{X}\}$ is identical in distribution to $\{Z(x), x \in \mathcal{X}\}$. In this setting, $\mathcal{X}$ is an arbitrary index set, possibly infinite, and can be defined as the set of locations for a spatial stochastic process. On the other hand, if $\mathcal{X}$ is finite, we get a multivariate extreme value

distribution, and for the case $|\mathcal{X}| = 2$, we recover the setting shown in Section 4.1. Assume now that the max-stable process has standard Fréchet margins:

$$\mathbb{P}(Z(x) \leq z) = \exp(-1/z), \quad x \in \mathcal{X}. \tag{5.1.2}$$

In this case, $A_{N_x} = N$ and $B_{N_x} = 0, N \geq 1, x \in \mathcal{X}$. Then, the process $\{\max_{1 \leq i \leq N} Z^{(i)}(x), x \in \mathcal{X}\}$ has the same distribution as $\{NZ^{(i)}(x), x \in \mathcal{X}\}$, and in the finite-dimensional case, one has

$$\mathbb{P}\left[\bigcap_{i=1}^{D}\{Z(x_i) \leq nz_i\}\right]^n = \mathbb{P}\left[\bigcap_{i=1}^{D}\{Z(x_i) \leq z_i\}\right] = \exp\{-V_D(z_1, \ldots, z_D)\}, \tag{5.1.3}$$

with V the so-called exponent measure, as defined for $D = 2$ in Section 4.1.

de Haan [1984] developed the spectral representation for max-stable processes and a generic method to construct them. Let $\{(\xi_i, s_i), i \geq 1\}$ denote the points of a Poisson process on $(0, \infty) \times \mathcal{S}$ with intensity measure $\xi^{-2}\,\mathrm{d}\xi \times \nu(\mathrm{d}s)$, for an arbitrary measurable set $\mathcal{S}$ and positive measure ν on $\mathcal{S}$. If $\{f(s, x), s \in \mathcal{S}, x \in \mathcal{X}\}$ is a non-negative function such that

$$\int_{\mathcal{S}} f(s, x)\nu(\mathrm{d}s) = 1, \quad x \in \mathcal{X}, \tag{5.1.4}$$

then the process Z^* defined by

$$Z^*(x) = \max_i \xi_i f(s_i, x), \quad x \in \mathcal{X} \tag{5.1.5}$$

is max-stable with standard Fréchet margins. As highlighted by Smith [1990], the above characterization can be given a rainfall-storm interpretation, where S refers to the space of storm centres with spatial distribution defined by the measure ν. The function f represents the shape of the storm and ξ_i its magnitude, so that the amount of rainfall at position x from a storm centred at location s_i with magnitude ξ_i is $\xi_i f(s_i, x)$. The process (5.1.5) has standard Fréchet margins. To see this, note that

$$Z^*(x) \leq z \iff \max_i \xi_i f(s_i, x) \leq z \iff \{(\xi, s) \in (0, \infty) \times S : \xi f(s, x) > z\} = \emptyset, \tag{5.1.6}$$

so that, using (5.1.4),

$$\mathbb{P}(Z^*(x) \leq z) = \exp\left[-\int_S \left(\int_{z/f(s,x)} \xi^{-2}\,\mathrm{d}\xi\right)\nu(\mathrm{d}s)\right] \tag{5.1.7}$$

$$= \exp\left[-\int_S f(s, x)z^{-1}\nu(\mathrm{d}s)\right] \tag{5.1.8}$$

$$= \exp(-1/z). \tag{5.1.9}$$

For multiple locations $x_1, \ldots, x_k \in \mathcal{S}$, one has

$$\mathbb{P}(Z^*(x_1) \leq z_1, \ldots, Z^*(x_k) \leq z_k) = \exp\left[-\int_S \max_{i=1,\ldots,k}\left\{\frac{f(s, x_i)}{z_i}\right\}\nu(\mathrm{d}s)\right]. \tag{5.1.10}$$

Smith [1990] provides examples of max-stable processes linked to different multivariate extreme value families, including the mixed, logistic and bilogistic models. In particular, assume that the space of locations $\mathcal{X}$ consists of two locations x_1 and x_2, and that

$$f(s, x) = \begin{cases} (1 - \alpha)s^{-\alpha}, & x = x_1 \\ (1 - \alpha)(1 - s)^{-\alpha}, & x = x_2, \end{cases} \tag{5.1.11}$$

with $0 < \alpha < 1$ and $\mathcal{S} = [0, 1]$, so that $\mathcal{S}$ can be interpreted as the set of possible locations on the segment linking x_1 and x_2. In this case,

$$\int_0^1 \max\left\{\frac{f(s, x_1)}{z_1}, \frac{f(s, x_2)}{z_2}\right\} ds = \left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{\alpha}, \tag{5.1.12}$$

and we get the logistic model of Tawn [1988].

5.1.1 The Gaussian value process

As in Smith [1990], one can define the function f as a multivariate normal density with zero mean and variance Σ:

$$f(s,x) = f_0(s-x) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left\{\frac{1}{2}(s-x)^T\Sigma^{-1}(s-x)\right\}. \tag{5.1.13}$$

The joint distribution for two locations x_1 and x_2 is then given by

$$\mathbb{P}\{Z^*(x_1)\leq z_1, Z^*(x_2)\leq z_2\} = \exp\left\{-\frac{1}{z_1}\Phi\left(\frac{a}{2}+\frac{1}{a}\log\frac{z_2}{z_1}\right)-\frac{1}{z_2}\Phi\left(\frac{a}{2}+\frac{1}{a}\log\frac{z_1}{z_2}\right)\right\} \tag{5.1.14}$$

with Φ the standard normal cumulative distribution function and a the Mahalanobis distance

$$a = \sqrt{(x_1-x_2)^T\Sigma^{-1}(x_1-x_2)}. \tag{5.1.15}$$

As $a\to 0^+$, we get perfect dependence, and independence as $a\to\infty$. The covariance matrix Σ defines the shape of the storms. When $\Sigma=\gamma\mathbb{I}_D$ with $\gamma>0$, the process is said to be isotropic, with a dependence function that scales equally in every direction $\mathcal{X}$. Otherwise, the process is called anisotropic. As shown in Blanchet and Davison [2011], one can transform the original space $\mathcal{X}$ into a climate space $\tilde{\mathcal{X}}$ for which the original process is now isotropic. Indeed, using the eigendecomposition $\Sigma = U\Lambda U^T$ with a diagonal matrix Λ and rotation matrix U, the squared Mahalanobis becomes

$$\begin{aligned}
a^2 &= (x_1-x_2)^T\Sigma^{-1}(x_1-x_2)\\
&= (x_1-x_2)^T(\Lambda^{-1/2}U)^T(\Lambda^{-1/2}U)(x_1-x_2)\\
&= \frac{1}{\lambda_1}(x_1-x_2)^T(\lambda_1^{-1/2}\Lambda^{-1/2}U)^T(\lambda_1^{-1/2}\Lambda^{-1/2}U)(x_1-x_2)\\
&= \frac{1}{\lambda_1}\{L(x_1-x_2)\}^T\{L(x_1-x_2)\},
\end{aligned}$$

where λ_1 is the first element of Λ, and $L=\lambda_1^{-1/2}\Lambda^{-1/2}U$. The transformation is thus defined with this matrix L, and provides a better way to compare the empirical extremal coefficient with those predicted by the model. Recall that it is defined as $\theta_N = V_D(1,\ldots,1)$, for V_D the exponent measure given by Equation 5.1.3. The characterization of the Gaussian value process yields

$$\theta_{x_1x_2} = 2\Phi\left(a/2\right). \tag{5.1.16}$$

This model offers numerous possibilities to define the space X and the structure of the covariance matrix Σ, and serves as a good starting point in several applications, such as the estimation of extreme snow depth [Blanchet and Davison, 2011] and the study of extreme rainfall in Australia [Saunders et al., 2019]. A recurring problem is that the location in $\mathcal{X}$ can exhibit different dependence structures, so that the max-stable process defined before might be an unrealistic option. Saunders et al. [2019] proposed a regionalization approach, where the stations are clustered using a hierarchical method and a distance measure based on the F-madogram. For each region, an anisotropic Smith model is fitted. This method showed convincing results and a good partition of the space into regions with similar dependence structures. This approach contains two separate steps, and so the uncertainty from the estimation of the clusters can not be directly relayed to the fitting step of the max-stable processes. One can however use bootstrapping to overcome this issue.

Another approach is to split the space into distinct homogeneous climatic regions and to specify this into the covariance matrix Σ, as shown in Blanchet and Davison [2011]. The behaviour of extreme events at the boundary between the regions can however be problematic. Nonetheless, this method yields reasonable results, especially for large regions with homogeneous dependency.

5.1.2 The Schlather model and its extension

Schlather [2002] introduced another characterization of max-stable process, with

$$Z^*(x) = \max_i \left[\xi_i \max\{0, Y_i(x)\}\right], \quad x \in \mathcal{X}, \tag{5.1.17}$$

where the $Y_i(\cdot)$ are independent and identically distributed copies of a stationary process $Y(\cdot)$ on $\mathcal{X}$ with $\mathbb{E}[\max\{0, Y(x)\}] = 1$. $\{\xi_i, i \geq 1\}$ are, as before, realisations of a Poisson process on $\mathbb{R}_+^*$ with intensity measure $\xi^{-2}d\xi$. As for the Smith model, the margins are standard Fréchet, and taking $Y_i(\cdot)$ as a standard Gaussian process with correlation function $\rho(h)$, the bivariate distribution is

$$\mathbb{P}\{Z^*(x_1) \leq z_1, Z^*(x_2) \leq z_2\} = \exp\left\{-\frac{1}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)\left(1 + \sqrt{1 - 2(\rho(h)+1)\frac{z_1 z_2}{(z_1+z_2)^2}}\right)\right\}, \tag{5.1.18}$$

with h the Euclidian distance $\|x_1 - x_2\|$. Valid parametric families for the correlation function $\rho(h)$ include the Cauchy, with $\rho(h) = [1+(h/c_2)^2]^{-\nu}$ ($c_2 > 0, \nu > 0$), the powered exponential, with $\rho(h) = \exp[-(h/c_2)^\nu]$ ($c_2 > 0$, $0 < \nu \leq 2$), the Whittle-Matérn or the Bessel. A sill c_1 and nugget effect μ can be added to these correlation functions,

$$\tilde{\rho}(h) = \begin{cases} \mu + c_1, & h = 0 \\ c_1 \rho(h), & h > 0 \end{cases} \tag{5.1.19}$$

with the constraint $\mu = 1 - c_1$ in order to have $\tilde{\rho}(0) = 1$. These correlation functions are isotropic. One can use the transformation $\rho(x_1, x_2) = \rho\{a(x_1, x_2)\}$ to get the general anisotropic case, where $a(x_1, x_2)$ is the Mahalanobis distance (5.1.15) with respect to locations x_1 and x_2. The extremal coefficient is

$$\theta_{x_1 x_2} = 1 + \left\{\frac{1 - \rho(h)}{2}\right\}^{1/2}. \tag{5.1.20}$$

Following Nikoloulopoulos et al. [2009] and Opitz [2013], one generalization of this model, the extremal t model, is defined by

$$Z^*(x) = \max_i \left[\xi_i c_\nu \max\left\{0, Y_i(x)\right\}^\nu\right] \tag{5.1.21}$$

with $c_\nu = 2^{1-\nu/2}\pi^{1/2}[\Gamma\{(\nu+1)/2\}]^{-1}$ and $\nu > 0$. The $Y_i(x)$ are realisations of a standard Gaussian process with correlation function $\rho(x_1, x_2)$ and $\Gamma(\cdot)$ is the gamma function. This model encompasses a wide range of models, from the Brown–Resnick process (Brown and Resnick [1977], Kabluchko et al. [2009]) as $\nu \to \infty$, to the Schlather process when $\nu = 1$, and is a great alternative to the Smith and Schlather processes. The bivariate distribution is given by

$$\mathbb{P}\{Z^*(x_1) \leq z_1, Z^*(x_2) \leq z_2\} = \exp\left\{-\frac{1}{z_1}T_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right] - \frac{1}{z_2}T_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right]\right\}, \tag{5.1.22}$$

where $T_\nu(\cdot)$ is the Student t cumulative distribution function with ν degrees of freedom, and

$$r_{x_1 x_2}(t) = \frac{t^{1/\nu} - \rho(x_1, x_2)}{(\nu+1)^{-1/2}[1 - \rho(x_1, x_2)^2]^{1/2}}.$$

5.1.3 Inference

Estimates for model parameters obtained using maximum likelihood require the joint density of the max-stable process. However, this density is typically not analytically known when the number of locations is greater than 2, and because of the explosion of the number of terms in the full likelihood, it is usually preferable to use a pairwise likelihood. Since the bivariate densities are known, the pairwise-likelihood is given by

$$l_p(\mathbf{z};\theta) = \sum_{k=1}^{N} \sum_{(i,j)\in\mathcal{P}_k} \log f(z_{k,i}, z_{k,j};\theta), \tag{5.1.23}$$

where θ are the parameters of the model, $\mathbf{z}$ is the set of all observations, $\mathcal{P}_k$ is the set of pairs of locations (i,j) for which the process was observed with realizations $(z_{k,i}, z_{k,j})$, for $k = 1,\dots,N$, and f is the joint bivariate density (see Section E.1, Appendix E). As the number N of samples increases and under suitable regularity conditions, the estimator $\hat{\theta}$ maximizing 5.1.23 has the limiting normal distribution

$$N^{1/2}(\hat{\theta}-\theta) \sim \mathcal{N}_p\{0, N^{-1}H(\hat{\theta})^{-1}J(\hat{\theta})H(\hat{\theta})^{-1}\}, \quad N\to\infty, \tag{5.1.24}$$

with

$$H(\theta) = N\int \frac{\partial^2 \log f(z;\theta)}{\partial\theta\partial\theta^T} g(z)\,\mathrm{d}z$$
$$J(\theta) = N\int \frac{\partial \log f(z;\theta)}{\partial\theta}\frac{\partial \log f(z;\theta)}{\partial\theta^T} g(z)\,\mathrm{d}z,$$

and $g(z)$ the true density of z. If our model is correctly specified, then $H(\theta) = -J(\theta)$. The uncertainty can be assessed with plug-in estimates for H and J, as shown in Padoan et al. [2009] and Varin and Vidoni [2005]. Due to potential local maxima, one can use a profile likelihood method to successively update the parameters until convergence, as shown in Blanchet and Davison [2011].

5.2 Application: multivariate analysis of extreme rainfall

In the following we extend the bivariate analysis of Section 4.2.1 through the use of max-stable processes. The study of bivariate extreme rainfall showed the stations had different extremal dependences depending on their locations, and the duration of the measurements. In particular, the stations facing the Caribbean coast have a dependence that increases as the duration grows, and have a very low extremal dependence with the three other stations studied. The aim of this section is to extend the measure of dependence of extreme rainfall to any location in Costa Rica.

As shown in Section 4.1.2, the extremal coefficient can be a good summary of the extremal dependence between two stations, and is given by $V_2(1,1)$, where V_2 is the exponent measure. A natural extension to the D-dimensional case is $V_D(1,\dots,1)$, with V_D the exponent measure for D locations. The parametric models given in Sections 5.1.1 and 5.1.2 provide an analytic form for the extremal coefficient. One can also be interested in non-parametric estimation of the extremal coefficient, in order to avoid assumptions about the distribution of the data. The variogram and semi-variogram introduced by Cressie [1993] are widely used to measure the dependence for spatial data, but their use is limited in the context of extremal data, as

they might not be defined due to the heavy tail characterization of the data. Cooley et al. [2006] used the F-madogram as a measure of extremal dependence. It is defined as

$$\nu_F(x_1, x_2) = \frac{1}{2}\mathbb{E}\left[|F\{Z^*(x_1)\} - F\{Z^*(x_2)\}|\right], \tag{5.2.1}$$

where $Z^*(\cdot)$ is a stationary max-stable process with standard Fréchet margins, with $F(z) = \exp(-1/z)$. An estimator for (5.2.1) is given by

$$\hat{\nu}_F(x_1, x_2) = \frac{1}{2N}\sum_{i=1}^{N}\left|\hat{F}_1\{z_i^*(x_1)\} - \hat{F}_2\{z_i^*(x_2)\}\right|, \tag{5.2.2}$$

with $z_i^*(x_1)$ and $z_i^*(x_1)$ realizations of the max-stable process at locations x_1 and x_2, and $\hat{F}_k$ an empirical estimator of the margins, such as

$$\hat{F}_k(z) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{z_i^*(x_k) \leq z\}, \quad k = 1, 2. \tag{5.2.3}$$

The F-madogram takes values in $[0, 1/6]$, with 0 corresponding to perfect dependence, and $1/6$ to independence. The extremal coefficient can be derived from the F-madogram as

$$\theta_{x_1 x_2} = \frac{1 + 2\nu_F(x_1, x_2)}{1 - 2\nu_F(x_1, x_2)}. \tag{5.2.4}$$

An extension of the F-madogram introduced by Naveau et al. [2009], the λ-madogram, fully characterizes the dependence through $\mathbb{P}\{Z^*(x_1) \leq z_1, Z^*(x_2) \leq z_2\}$, for any $z_1, z_2 \in \mathbb{R}_{>0}$.

5.2.1 A regionalization approach

Figure 5.1 shows estimates of the extremal dependence between each pair of stations, computed with F-madogram. The dependence differs depending on the duration. For stations less than 200km apart, high durations tend to have more independence than low durations., but this is reversed for stations more than 250km apart. This includes pairs of stations from the north and the south Pacific coast. A closer look at the extremal dependence behaviour for each basin showed that the stations from basin 69 tend to approach asymptotic independence for long distances, and that the stations from basin 98 exhibit dependent behaviour of extremes that does not vary much with the distance, and only slightly with the duration (with higher dependence for long durations). These results highlight the presence of multiple weather systems and potentially local effects that might be caused by topography.

In order to account for these differences, we borrow the approach proposed by Saunders et al. [2019], and use hierarchical clustering to group the stations, with a distance measure based on the F-madogram between two stations x_1 and x_2 is defined as $d(x_1, x_2) = \hat{\nu}_F(x_1, x_2)$ and ranging from 0 to $1/6$. The use of hierarchical clustering may turn out to be more informative than approaches such as K-medoids and K-means, as it provides a natural tree-like interpretation of dependence among the stations. In particular, the approach starts by assigning one cluster to each station, and then successively aggregates the clusters based on a linkage criterion that measures the dissimilarity between clusters. The two clusters that have the smallest dissimilarity are merged. To extend this partitioning of the stations to any location on the domain of study, Saunders et al. [2019] proposed to use a weighted k-nearest neighbours classifier. The results of this approach applied to our study of costarican extreme rainfalls are shown in Figure 5.2. Hierarchical clustering does not provide a generic way to

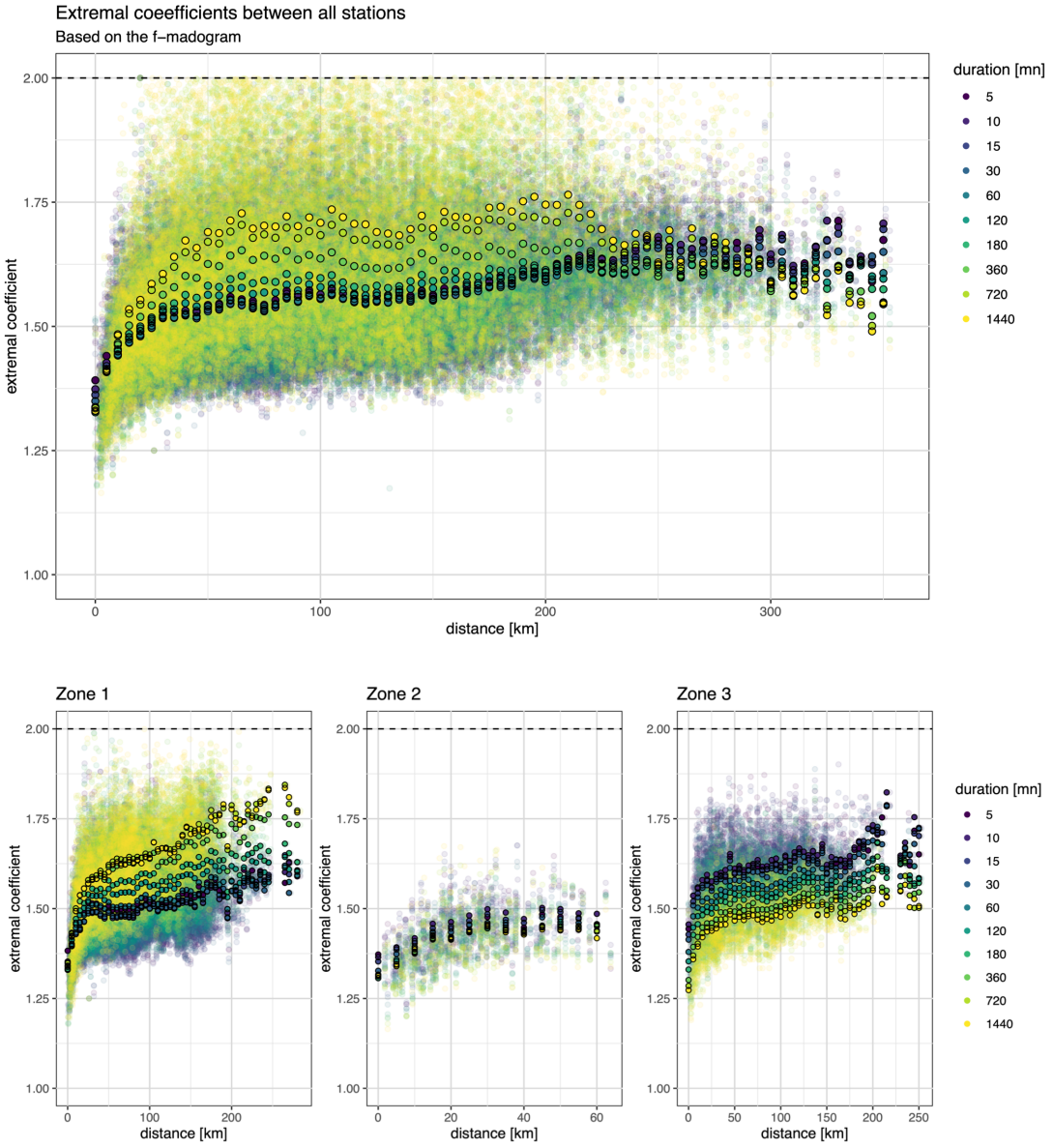


Figure 5.1: Estimates of the extremal dependence between each pair of stations, for all durations, with 5km bins for clarity. The top panel shows these estimates for all 160 stations, while the panels at the bottom show them for three different zones (given by the clustering approach, see Figure 5.2). Overall, higher durations tend to exhibit more asymptotic extremal independence as the distance increases, until 200km. From then, the dependence increases (for duration 360 to 1440 minutes). This includes pairs of stations with the north and south of Costa Rica respectively, with basins 76/69 (north) and 98 (south).

select the number of clusters for the partitioning, which is often based on other desired properties. In our case, we selected the number of groups based on the behaviour of the extremal coefficient with the distance and the durations, with clusters ideally showing homogeneous properties. This led us to consider three groups of stations, as highlighted in Figure 5.2. For the classification step, we used the k-NN classifier (see Hastie et al. [2004]) with $k = 2$ chosen by cross validation, using the longitude, latitude and elevation of each station. The distance is computed across all durations, as the attribution of stations among clusters might oth-

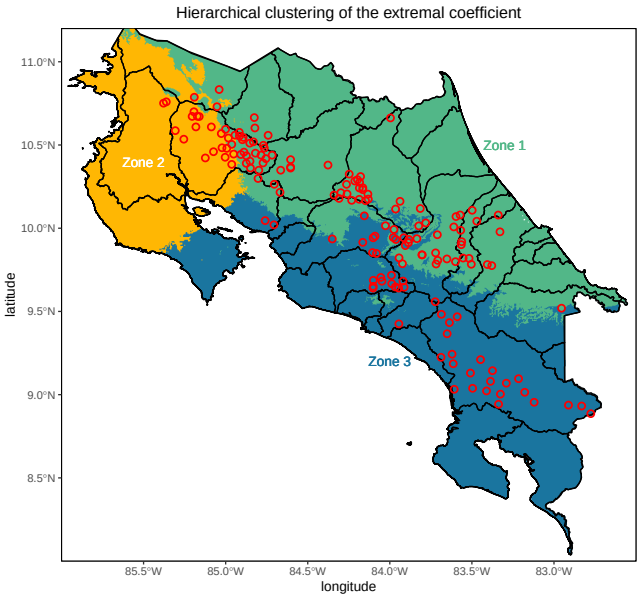


Figure 5.2: Clusters found with hierarchical clustering. The red circles correspond to the stations, and the colors to one of the three clusters. The classification is done with k-nearest neighbours, with $k = 2$ based on cross validation. The black delimitations correspond to the drainage basins.

erwise change depending on the duration, which would make the max-stable models hardly interpretable. As shown in Figures D.1, D.2 and D.3 (Appendix D), the stations in the southernmost regions of the Pacific coast of Costa Rica have large dissimilarities with the rest of the stations for low durations. As the duration increases, all the Pacific coast is included in this cluster. The stations in the center of the country and those close to San Jose have extremal dependence behaviours that are similar to those of the Guanacaste and Alajuela provinces (in the north of Costa Rica). Appendix D provides a comparison of the regions found with hierarchical clustering, based on another distance matrix, computed with the Tail pairwise dependence matrix (TPDM). The approach based on F-madogram gives more interpretable regions, and so we decided to keep it in the following.

A closer look at the extremal coefficient for each of the three regions reveals different behaviours (Figure 5.1, lower panels). The first region (consisting of precipitation gauges from basin 69 and those on the Caribbean coast) is characterized by a dependence that increases as the distance increases, at a rate that differs depending on the duration. Indeed, higher durations tend to move toward independence at a faster rate than lower durations, that roughly plateau at a value of 1.5. The second region (north of Costa Rica and the province of Guanacaste) sees higher durations that tend to be more dependent than lower durations. As this region is relatively small compared to the two others, asymptotic extremal dependence at large distances can not be measured. The third region (south Pacific coast) is also characterized with extreme rainfall for high durations that are more dependent than for lower durations (unlike the first region), but the rate toward asymptotic independence as the distance increases is similar for all durations. Lastly, it should be mentioned that even for very close stations, asymptotic dependence is not perfectly achieved, and the extremal dependence typically ranges from 1.2 to 1.5.

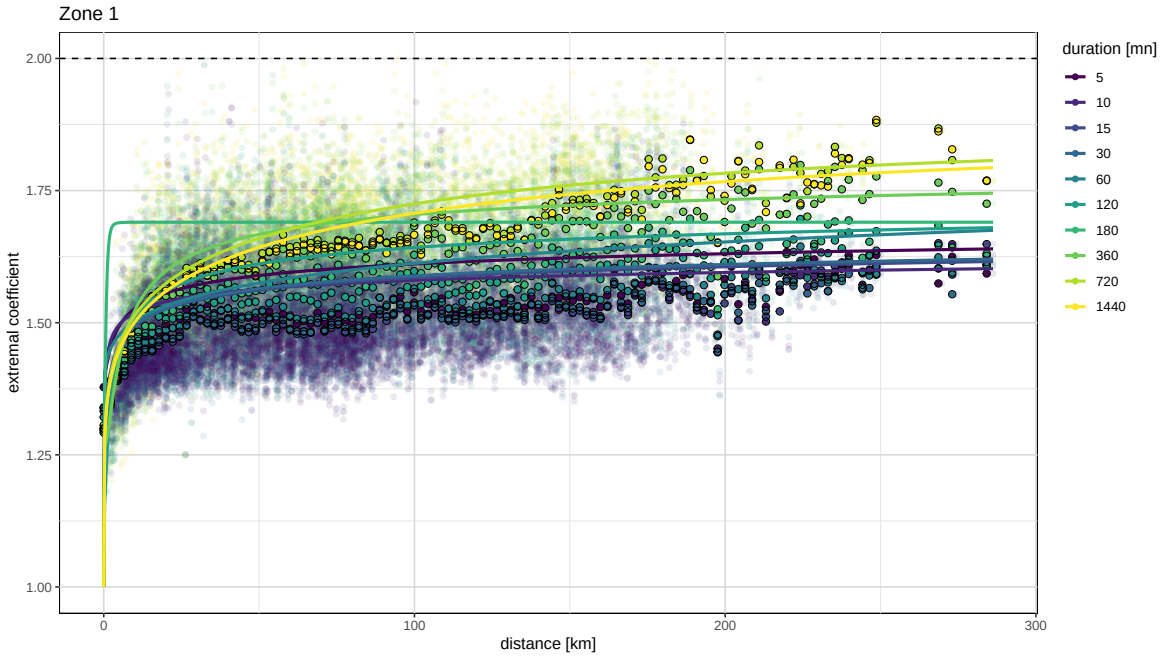


Figure 5.3: Empirical extremal coefficients along with the one predicted by the max-stable process. Points correspond to the binned empirical estimates of the extremal coefficients.

Figure 5.3 shows the fitted values of a t-extremal max-stable process (with Whittle-Matern correlation function) for the first cluster (zone 1, that includes basin 69 and basins facing the Caribbean coast), for each duration. The estimation procedure failed to converge for durations between 360 and 30 minutes, and in this case the Cauchy correlation function was used. The max-stable model seems to fit the data relatively well for high durations, but the fitted extremal coefficient slightly diverges from its empirical counterpart for low durations.

5.2.2 Modelling dependence of extremes with non-stationary processes

In this section, we extend the regionalization approach by fitting a unique non-stationary max-stable process for all locations. More specifically, we borrow the approach of Huser and Genton [2016] and merge it with a conditional density network (CDN), as in Sections 3.3.2 and 4.2.2. The objective is to have a model that can take into account the presence of multiple local effects and weather systems, with a dependence structure that is non-stationary. More specifically, we use the extremal t model with bivariate distribution (E.1.1) and correlation function

$$\rho(x_1, x_2) = |\Omega_{x_1}|^{1/4} |\Omega_{x_2}|^{1/4} \left| \frac{\Omega_{x_1} + \Omega_{x_2}}{2} \right|^{-1/2} R\left(Q^{1/2}_{x_1,x_2}\right), \tag{5.2.5}$$

and

$$R(h) = \exp(-h^{\alpha}), \quad h \geq 0,\ \alpha \in]0, 2]; \tag{5.2.6}$$

α controls the roughness of the random fields. Non-stationarity emerges as the covariance matrix Ω_x changes depending on the location x. In order to combine the covariance at two different sites x_1 and x_2, one can use the approach of Paciorek and Schervish [2006] to define the quadratic form

$$Q_{x_1,x_2} = (x_1 - x_2)^T \left(\frac{\Omega_{x_1} + \Omega_{x_2}}{2} \right)^{-1} (x_1 - x_2). \tag{5.2.7}$$

Huser and Genton [2016] suggested to model non-stationarity through the covariance matrix Ω_x as

$$\Omega_x = \begin{pmatrix} s_{11}(x) & s_{12}(x) \\ s_{21}(x) & s_{22}(x) \end{pmatrix} \in \mathbb{R}^{2\times 2}, \tag{5.2.8}$$

with

$$\log s_{ii}(x) = \boldsymbol{X}_{s_{ii}}^T(x)\beta_{s_{ii}}, \quad i = 1, 2, \tag{5.2.9}$$

$$s_{ij}(x) = \delta(x)\{s_{ii}(x)s_{jj}(x)\}^{1/2} = s_{ji}(x), \quad i \neq j, \tag{5.2.10}$$

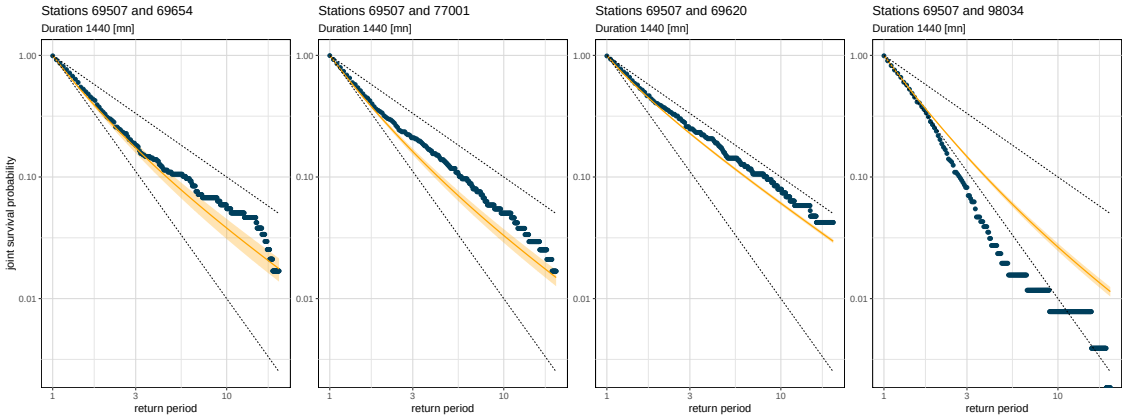
$$\delta(x) = -2\log\left[\{\boldsymbol{X}_\delta^T(x)\beta_\delta\}^{-1} - 1\right] - 1 \in (-1, 1). \tag{5.2.11}$$

The constraint on $\delta(x)$ ensures that the matrix is invertible and positive definite, and the matrices $\boldsymbol{X}_\delta, \boldsymbol{X}_{s_{11}}$ and $\boldsymbol{X}_{s_{22}}$ are the vector of covariates, typically the longitude and latitude of the precipitation gauges. The parameters $\beta_{s_{11}}, \beta_{s_{22}}$ and β_δ are estimated by maximizing the log pairwise likelihood, as in Section 5.1.3. Due to potential non-linear relationships between the covariates and the dependence structure, we decided to use the approach of Sections 3.3.2 and 4.2.2 and use a conditional density network. The models takes as input the longitude, latitude, altitude and duration, and provides as output the entries $s_{11}(x), s_{22}(x)$ and $\delta(x)$ to construct the covariance matrix Ω_x. The architecture is shown in Figure F.3 (Appendix F), and consists of three hidden layers of sizes 100, 20, 20 and 10, with the sigmoid activation function. Batch normalization layers [Ioffe and Szegedy, 2015] were used to properly control and accelerate training. As highlighted in Huser and Genton [2016], the use of a single smoothness parameter α can limit the flexibility of the model, especially for regions with different characteristics (such as the altitude). They proposed to use a mixture of max-stable processes where

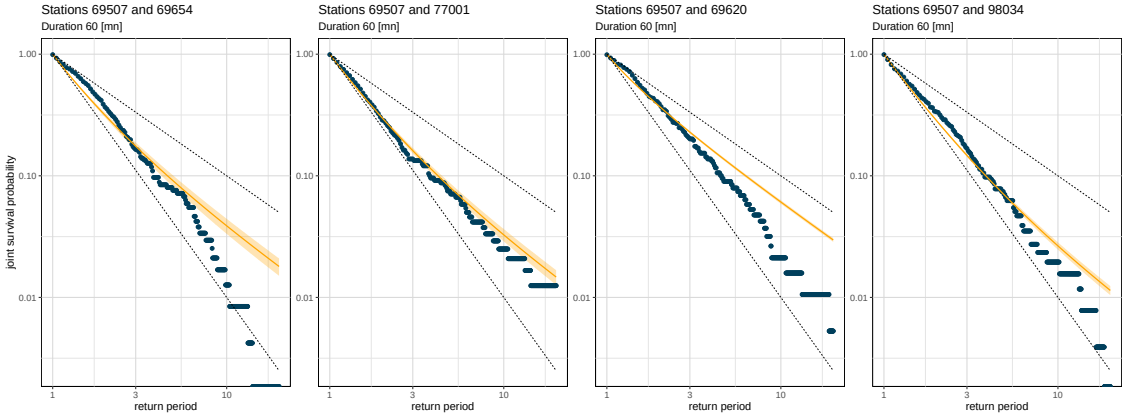
$$\rho(x_1, x_2) = \frac{a(x_1)a(x_2)\rho^1(x_1, x_2) + \{1 - a(x_1)\}\{1 - a(x_2)\}\rho^2(x_1, x_2)}{\sqrt{[a(x_1)^2 + \{1 - a(x_1)\}^2][a(x_2)^2 + \{1 - a(x_2)\}^2]}}, \tag{5.2.12}$$

and with correlation functions ρ^1 and ρ^2 having different smoothness parameters. The weights $a(x)$ can be modelled with covariates, for example via a CDN, taking as input longitude, latitude, altitude and duration. We chose smoothness parameters with values $\alpha = 0.5$ and $\alpha = 2$ for the model mixture. The factor $a(x)$ takes highest values for regions at mid-altitude, between 700 and 1600m, near the province of San José (Figure E.3 Appendix E). This corresponds to a model with lower smoothness, and better models to possible high granularity of the terrain. High and low altitude regions are characterized by a higher smoothness parameter. However, one should mention that these regions have very few representative precipitation gauges in our dataset, and may be strongly distant.

Figure 5.4 shows the estimated joint survival probability, as in Section 4.2.2. Compared to the extended bivariate approach, the confidence are narrower and struggles to contain the empirical estimates. The fit is better for shorted return periods, and the estimates of the extremal coefficient agree with those obtained with the bivariate model. For pairs of stations with asymptotic independence, this approach get better results. Estimates of the extremal coefficient for any location and with respect to stations 69507, 77001, 75022, 84118, 98006 and 76026 are shown in Figure E.4 (Appendix E), along with the contour from the covariance matrix, centered at each one of the six precipitation gauges. Stations located on the Pacific coast tend to have a dependency structure that extends along the coast, while for the Caribbean coast ellipses are more circular. The extremal coefficient might not have a circular shape with the distance in this case, as the weight function $a(\cdot)$ depends on the location and thus impacts the correlation function.



(a)



(b)

Figure 5.4: Joint survival probability plots of the non-stationary max-stable model for four pairs of stations related to station 69507. The dark points show the empirical estimates for the joint probability and the yellow curve the estimates from the non-stationary max-stable model. The upper and lower dotted grey lines correspond to the cases $V(1,1) = 1$ (total dependence) and $V(1,1) = 2$ (independence).

5.3 Comparison of the bivariate and max stable models

In this section, we compare the extended bivariate and non-stationary max-stable approaches. From a practical standpoint, the non-stationary max-stable model is better suited for a global understanding of the extremal dependence of extreme rainfall. Indeed, as the model learns to map the location of the precipitation gauge with its covariance matrix, the model can estimate $V_{x_1,x_2}(1,1)$ for any two locations x_1 and x_2. The extended bivariate approach models the extremal coefficient with respect to a reference station. Hence, one model per precipitation gauge needs to be fitted (excluding the bootstrapping process to construct the confidence intervals), which might be computationally expensive. The max-stable approach model the dependency structure with respect to each location, which makes it an attractive tool for questions of interpretability as well. Moreover, the definition of the stochastic process (5.1.21) can be used for simulations. This is particularly interesting when one wants to estimate the distribution of monthly maxima for a subset of stations. An analytical form for the bivariate case is known (see Figure 5.4), but the extension for more than two stations might require to perform simulation from the max-stable process, for which more details can be found in Schlather [2002], Dombry et al. [2012] and Dombry et al. [2016]. However, this aspect is not covered in this report other than the bivariate case.

In order to assess the performance of the model, we use the continuous rank probability score (CRPS), which finds its use in measuring the accuracy of a probability forecast model. As both bivariate and max-stable models provide confidence intervals for the extremal coefficient, this measure might provide a better comparison. The reader may refer to Gneiting and Raftery [2007] for an introduction to strictly proper scoring rules and the CRPS, and Taillardat et al. [2019] for an application to extreme event evaluation. Briefly, it is defined as

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \left[F(y) - \mathbb{1}_{(y \geq x)}\right]^2 \mathrm{d}y, \tag{5.3.1}$$

where x corresponds to the observed value and F to the distribution function. If F is the normal cumulative distribution function with mean μ and variance σ^2, then

$$\mathrm{CRPS}(F, x) = \sigma \left[2\phi\left(\frac{x-\mu}{\sigma}\right) + \frac{x-\mu}{\sigma}\left\{2\Phi\left(\frac{x-\mu}{\sigma}\right) - 1\right\} - \frac{1}{\sqrt{\pi}}\right], \tag{5.3.2}$$

with ϕ and Φ the normal density and cumulative distribution functions respectively. As noted in Grimit et al. [2006], the CRPS can be empirically estimated from a sample $y_1, \ldots, y_N \sim F$ by

$$\mathrm{CRPS}(\hat{F}_N, x) = \frac{1}{N}\sum_{i=1}^{N} |y_i - x| - \frac{1}{2N^2}\sum_{i,j=1}^{N} |x_i - x_j|, \tag{5.3.3}$$

with $\hat{F}_N$ the empirical cumulative distribution function based on samples $y_1, \ldots, y_N$. Proper scoring rules and CRPS can be computed using the `R` package `scoringRules` [Jordan et al., 2019]. Like in Section 3.3.3, one can extend the CRPS to compare to models with the CRPS index (CRPSI). If $\mathcal{M}$ and $\mathcal{R}$ are two models ($\mathcal{R}$ being the reference one), then

$$CRPSI(F, x) = \begin{cases} 1 - \frac{CRPS^{\mathcal{M}}(F,x)}{CRPS^{\mathcal{R}}(F,x)}, & CRPS^{\mathcal{M}}(F, x) \leq CRPS^{\mathcal{R}}(F, x), \\ -1 - \frac{CRPS^{\mathcal{R}}(F,x)}{CRPS^{\mathcal{M}}(F,x)}, & CRPS^{\mathcal{M}}(F, x) > CRPS^{\mathcal{R}}(F, x). \end{cases} \tag{5.3.4}$$

For our comparison, we let $\tilde{\theta}_{x_1,x_2} = \tilde{V}_{x_1,x_2}(1,1)$ be the estimated extremal coefficient predicted by the model with distribution $F^{\theta}_{x_1,x_2}$, and $\hat{\theta}_{x_1,x_2} = \hat{V}_{x_1,x_2}(1,1)$ be the empirical extremal coefficient computed from the data with equation (4.2.4). Figure 5.5 shows the average CRPS index (across the pairs of stations from the testing set) of the extended bivariate model against the non-stationary max-stable process, depending on the distance between the stations and the duration. Positive CRPSI values indicate an improvement of the bivariate compared to the max-stable model, while red ones an improvement of the max-stable from the bivariate approach. The bivariate approach is globally preferred for durations above 60 minutes, especially to model the extremal dependence of stations that are far apart. For durations smaller than 60 minutes, the results are less interpretable, and the max-stable process can find its use for long distances as well.

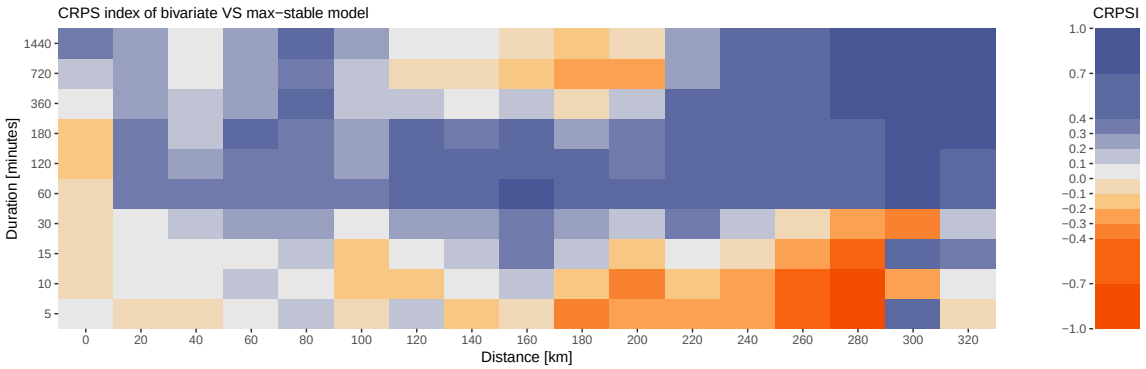


Figure 5.5: CRPS index of the extended bivariate model against the non-stationary max-stable process. Positive values of CRPS index (in blue) indicate improvements of the bivariate model compared to the non-stationary max-stable process, while negative ones (red) an improvement of the max-stable to the bivariate model.

Chapter 6

Conclusion

The study of extreme rainfall in Costa Rica is a real challenge, both by the diversity of the geographical landscapes that make up the country, and by the availability of rainfall records. The characterization of rainfall is done through intensity-duration-frequency (IDF) curves, which make it possible to summarize the risks of extreme precipitation for different return levels and durations. On the other hand, they are of great importance for water resource management, drainage basin systems design and extreme rainfall events.

Extreme value theory (EVT) is particularly suitable for the construction of such curves, because it is interested in modeling the distribution and the asymptotic behavior of random value maxima. In this report, we focused on the block maxima approach, and we first studied six stations in Costa Rica with different characteristics, whether in terms of location, altitude or basin. The EVT study revealed significant differences in return levels, depending on months and durations. The Caribbean coast is characterized by 20-year return levels with a small variation through the months, but whose seasonality varies according to the duration. Return levels for long durations, linked to extreme precipitation that extends over time, tend to peak during the months of December to February, while short durations peak between May and October. The effect called Little summer has also been observed on the IDF curves, but is more a phenomenon linked to stations located on the Pacific coast, which is marked by a strong seasonality, especially the region of Guanacaste. In order to extend the construction of IDF curves to the whole territory, two approaches have been implemented: one based on the k-nearest neighbours algorithm, and the other on a conditional neural network (CDN). The method based on CDN is advantageous for modeling complex interactions between altitude, location, duration and monthly variables, and for estimating level returns for long periods. Although these univariate approaches make it possible to construct IDF curves for any location in Costa Rica, they may suffer from the lack of stations available in certain regions of the country.

Thus, a multivariate study of extreme rainfall was conducted in order to estimate the probabilities that several stations would see significant monthly rainfall levels simultaneously. Initially, a bivariate approach was used and extended to the entire territory. The latter revealed interesting asymptotic dependence phenomena, such as the appearance of a dependency pattern at the foot of the mountains of the Caribbean coast, in which the dependence increases as the duration prolongs. A similar phenomenon but to a lesser extent is observed for the Pacific coast. This approach makes it possible to model a dependency that can greatly vary in space, but has the disadvantage of requiring a model for each station, which can prove to be computationally expensive.

The second multivariate approach is based on max-stable processes, whose direct integration with the theory of extreme values and the diversity of the classes of models constituting it make it a convincing tool. Given the potential presence of local climatic phenomena, the max-stable process like that of Smith or Schlather struggles to correctly model the dependency. A clustering analysis based on the F-madogram has highlighted regions that tend to see extreme events simultaneously, which could prove useful for the implementation of risk anticipation measures for certain areas of the country.

Additionally, we used a non-stationary max-stable process with the particularity of being combined with a deep neural network to model dependency across Costa Rica. Although this approach does not perform as well as the extended bivariate approach for long durations, it nevertheless holds up for short ones. In addition, this method benefits from better interpretability and the possibility of modeling the risks between several stations through simulations, unlike the bivariate approach. However, this aspect has not been fully covered in this report, and could be investigated in future studies.

Bibliography

Alfaro, E. (2002). Some characteristics of the precipitation annual cycle in central america and their relationships with its surrounding tropical oceans. *Tóp. Meteorol. Ocean.* 7: 88–103.

Alfaro, E., Chourio, X., Muñoz, and Mason, S. (2017). Improved seasonal prediction skill of rainfall for the Primera season in Central America. *International Journal of Climatology* 38 (2), doi:10.1002/joc.5366.

Amador, J. (1998). A climatic feature of the tropical Americas: The trade wind easterly jet. *Top. Meteor. Oceanogr.* 5: 91–102.

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical Modeling and Analysis of Spatial Data*, *101*. doi:10.1201/9780203487808.

Beirlant, J., Goegebeurm, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, pagination: 522.

Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019). Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences* 116: 15849–15854.

Bengio, Y., Simard, P. and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5: 157–166, doi:10.1109/72.279181.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29: 1165–1188, doi:10.1214/aos/1013699998.

Bentzien, S. and Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society* 140: 1924–1934, doi:https://doi.org/10.1002/qj.2284.

Birnbaum, Z. W. and Tingey, F. H. (1951). One-Sided Confidence Contours for Probability Distribution Functions. *The Annals of Mathematical Statistics* 22: 592–596, doi:10.1214/aoms/1177729550.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics* 5: 1699–1725, doi:10.1214/11-AOAS464.

Braun, H. (1980). A Simple Method for Testing Goodness of Fit in the Presence of Nuisance Parameters. *Journal of the Royal Statistical Society. Series B (Methodological)* 42: 53–63.

Brown, B. M. and Resnick, S. I. (1977). Extreme Values of Independent Stochastic Processes. *Journal of Applied Probability* 14: 732–739.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference. *Sociological Methods & Research* 33: 261–304.

Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes* 24: 673–685, doi:https://doi.org/10.1002/hyp.7506.

Castillo, R. and Amador, J. A. (2020). Precipitation and Temperature in Costa Rica at the End of the Century Based on NEX–GDDP Projected Scenarios. *Atmosphere* 11, doi:10.3390/atmos11121323.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Coles, S. and Dixon, M. J. (1999). Likelihood-Based Inference for Extreme Value Models. *Extremes* 2: 5–23.

Cooley, D., Naveau, P. and Poncet, P. (2006). *Variograms for spatial max-stable random fields*. New York, NY: Springer New York. 373–390, doi:10.1007/0-387-36062-X_17.

Cooley, D. and Thibaud, E. (2018). Decompositions of Dependence for High-Dimensional Extremes. doi:10.1093/biomet/asz028.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley series in probability and mathematical statistics. New York: J. Wiley Sons, revised edition ed.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley series in probability and statistics. J. Wiley Sons.

Davison, A. C. (2019). *Risk, rare events and extremes, MATH-447 course*. EPFL.

de Haan, L. (1984). A Spectral Representation for Max-stable Processes. *The Annals of Probability* 12: 1194–1204, doi:10.1214/aop/1176993148.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. doi:10.1007/0-387-34471-3.

de Haan, L. and Pickands, J. (1986). Stationary min-stable stochastic processes. *Probability Theory and Related Fields* 72: 477–492, doi:https://doi.org/10.1007/BF00344716.

Dombry, C., Engelke, S. and Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika* 103(2): 303–317.

Dombry, C., Éyi-Minko, F. and Ribatet, M. (2012). Conditional simulation of max-stable processes. *Biometrika* 100: 111–124.

Dudani, S. A. (1976). The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6: 325–327.

Dyrrdal, A., Lenkoski, A., Thorarinsdottir, T. and Stordal, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics* 26: 89–106, doi:10.1002/env.2301.

El Adlouni, S., Ouarda, T., Zhang, X., Roy, R. and Bobée, B. (2007). Generalized Maximum Likelihood Estimators for the Nonstationary Generalized Extreme Value Model. *Water Resources Research* 43(3), doi:10.1029/2005WR004545.

Ferreira, M. (2018). Heuristic tools for the estimation of the extremal index: A comparison of methods. *REVSTAT Statistical Journal* 16: 115–136.

Galambos, J. (1975). Order Statistics of Samples from Multivariate Distributions. *Journal of the American Statistical Association* 70: 674–680.

GFDRR (2011). Vulnerability, Risk Reduction, and Adaptation to Climate Change – Costa Rica. Tech. rep., WB, UNDP, UNISDR, Global Facility for Disaster Reduction and Recovery.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks 9: 249–256.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102: 359–378, doi:10.1198/016214506000001437.

Gonfiantini, R., Roche, M., Olivry, J., Fontes, J. and Zuppi, G. (2001). The altitude effect on the isotopic composition of tropical rains. *Chemical Geology - CHEM GEOL* 181: 147–167, doi:10.1016/S0009-2541(01)00279-0.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press, `http://www.deeplearningbook.org`.

Grimit, E. P., Gneiting, T., Berrocal, V. J. and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society* 132: 2925–2942, doi:https://doi.org/10.1256/qj.05.235.

Gumbel, E. (1960). Distributions des valeurs extremes en plusieurs dimensions. *Publications de l'Institut de Statistique de l'Université de Paris* 9: 171–173.

Hastenrath, S. (1967). Rainfall distribution and regime in Central America. *Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie B* 15: 201–241.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Math. Intell.* 27: 83–85, doi:10.1007/BF02985802.

Huser, R. and Genton, M. (2016). Non-Stationary Dependence Structures for Spatial Extremes. *Journal of Agricultural, Biological, and Environmental Statistics* 21: 470–491, doi:10.1007/s13253-016-0247-4.

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

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv* abs/1502.03167.

Jaklič, A., Šajn, L., Derganc, G. and Peer, P. (2015). Automatic digitization of pluviograph strip charts. *Meteorological Applications* 23, doi:10.1002/met.1522.

Jiang, Y., Cooley, D. and Wehner, M. (2020). Principal Component Analysis for Extremes and Application to US Precipitation. *Journal of Climate* 33: 6441–6451, doi:10.1175/JCLI-D-19-0413.1.

Jordan, A., Krüger, F. and Lerch, S. (2019). Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software* 90: 1–37, doi:10.18637/jss.v090.i12.

Kabluchko, Z., Schlather, M. and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *The Annals of Probability* 37, doi:10.1214/09-aop455.

Karnauskas, K., Seager, R., Giannini, A. and Busalacchi, A. (2013). A simple mechanism for the climatological midsummer drought along the Pacific coast of Central America. *Atmósfera* 26: 261–281, doi:10.1016/S0187-6236(13)71075-0.

Khaliq, N., Ouarda, T., Ondo, J.-C., Gachon, P. and Bobée, B. (2006). Frequency Analysis of a Sequence of Dependent and/or Non-Stationary Hydro-Meteorological Observations: A Review. *Journal of Hydrology* 329: 534–552, doi:10.1016/j.jhydrol.2006.03.004.

Kharin, V. V. and Zwiers, F. W. (2005). Estimating Extremes in Transient Climate Change Simulations. *Journal of Climate* 18: 1156–1173.

Khosravi, A., Nahavandi, S., Creighton, D. and Atiya, A. F. (2011). Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances. *IEEE Transactions on Neural Networks* 22: 1341–1356, doi:10.1109/TNN.2011.2162110.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Koutsoyiannis, D., Kozonis, D. and Manetas, A. (1998). A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology* 206: 118–135, doi:10.1016/S0022-1694(98)00097-3.

Leadbetter, M., Lindgren, G. and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer, New York, NY, doi:https://doi.org/10.1007/978-1-4612-5449-2.

Lehmann, E., Phatak, A., Soltyk, S., Chia, J., Lau, R. and Palmer, M. (2013). Bayesian hierarchical modelling of rainfall extremes : 2806–2812.

Magaña, V., Amador, J. A. and Medina, S. (1999). The Midsummer Drought over Mexico and Central America. *Journal of Climate* 12: 1577–1588, doi:https://doi.org/10.1175/1520-0442(1999)012<1577:TMDOMA>2.0.CO;2.

Naveau, P., Guillou, A., Cooley, D. and Diebolt, J. (2009). Modelling pairwise dependence of maxima in space. *Biometrika* 96: 1–17.

Nikoloulopoulos, A., Joe, H. and Li, H. (2009). Extreme value properties of multivariate t copulas. *Extremes* 12: 129–148, doi:10.1007/s10687-008-0072-4.

Olsson, J., Södling, J., Berg, P., Wern, L. and Eronn, A. (2019). Short-duration rainfall extremes in Sweden: A regional analysis. *Hydrology Research* 50 (3): 945–960, doi:10.2166/nh.2019.073.

Opitz, T. (2013). Extremal t processes: Elliptical domain of attraction and a spectral representation. *Journal of Multivariate Analysis* 122: 409–413, doi:10.1016/j.jmva.2013.08.008.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial Modelling Using a New Class of Nonstationary Covariance Functions. *Environmetrics* 17 (5): 483–506.

Padoan, S. A., Ribatet, M. and Sisson, S. A. (2009). Likelihood-Based Inference for Max-Stable Processes. *Journal of the American Statistical Association* 105: 263–277, doi:10.1198/jasa.2009.tm08577.

Peter, M., Ulbrich, U. and Rust, H. (2017). A spatial and seasonal climatology of extreme precipitation return-levels: A case study. *Spatial Statistics* 34, doi:10.1016/j.spasta.2017.11.007.

Resnick, S. (2002). Hidden Regular Variation, Second Order Regular Variation and Asymptotic Independence. *Extremes* 5: 303–336, doi:10.1023/A:1025148622954.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*.

Ribatet, M. (2009). A User's Guide to the `SpatialExtremes` Package.

Rothfuss, J., Ferreira, F., Walther, S. and Ulrich, M. (2019). Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks. *ArXiv* abs/1903.00954.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, doi:10.1017/CBO9780511755453.

Sansom, J. (1987). Digitizing pluviographs. *Journal of Hydrology (New Zealand)* 26: 197–209.

Saunders, K. R., Stephenson, A. G. and Karoly, D. J. (2019). A Regionalisation Approach for Rainfall based on Extremal Dependence. *arXiv: Applications* .

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Texts in Statistical Science, pagination: 512.

Schlather, M. (2002). Models for Stationary Max-Stable Random Fields. *Extremes* 5: 33–44, doi:10.1023/A:1020977924878.

Schlather, M. and Tawn, J. A. (2003). A Dependence Measure for Multivariate and Spatial Extreme Values: Properties and Inference. *Biometrika* 90: 139–156.

Schmidhuber, J. (2014). Deep Learning in Neural Networks: An Overview. *CoRR* abs/1404.7828.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* 72: 67–90, doi:10.1093/biomet/72.1.67.

Smith, R. L. (1990). Max-stable processes and spatial extremes (Unpublished manuscript) .

Song, L., Chen, M., Gao, F., Cheng, C., Chen, M., Yang, L. and Wang, Y. (2019). Elevation Influence on Rainfall and a Parameterization Algorithm in the Beijing Area. *Journal of Meteorological Research* 33: 1143–1156, doi:10.1007/s13351-019-9072-3.

Stedinger, J. and Foufoula-Georgiou, E. (1993). Frequency analysis of extreme events. *Handbook of Hydrology* 18.

Stephenson, A. and Tawn, J. (2005). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika* 92: 213–227, doi:10.1093/biomet/92.1.213.

Taillardat, M., Fougères, A.-L., Naveau, P. and De Fondeville, R. (2019). Extreme events evaluation using CRPS distributions. *arXiv: Methodology* Working paper or preprint.

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

Telgarsky, M. (2015). Representation Benefits of Deep Feedforward Networks. *CoRR* abs/1509.08101.

Ulrich, J., Jurado, O., Peter, M., Scheibel, M. and Rust, H. (2020). Estimating idf curves consistently over durations with spatial covariates. *Water* 12(11): 3119, doi:10.3390/w12113119.

Van de Vyver, H. (2012). Spatial regression models for extreme precipitation in Belgium. *Water Resources Research* 48, doi:https://doi.org/10.1029/2011WR011707.

Varin, C. and Vidoni, P. (2005). A Note on Composite Likelihood Inference and Model Selection. *Biometrika* 92: 519–528.

Vuille, M. (2011). *Climate Variability and High Altitude Temperature and Precipitation*. 153–156, doi:10.1007/978-90-481-2642-2_66.

Wang, X., Zwiers, F. and Swail, V. (2004). North atlantic ocean wave climate change scenarios for the twenty-first century. *Journal of Climate* 17: 2368–2383, doi:10.1175/1520-0442(2004)017<2368:NAOWCC>2.0.CO;2.

Wei, H., Fan, G., Zhou, D., Ni, C., Li, X., Wang, Y., Liu, Y. and Huang, X. (2008). Preliminary analysis on the relationships between Tibetan Plateau NDVI change and its surface heat source and precipitation of China. *Science in China Series D Earth Sciences* 51: 677–685, doi:10.1007/s11430-008-0063-y.

Youngman, B. D. (2020). `evgam`: An `R` package for Generalized Additive Extreme Value Models. *arXiv: Computation* .

Appendices

Appendix A

Data exploratory analysis

Figure A.1 shows Costa Rica's elevation profile and the precipitation gauges, along with their location on the map. The distribution of station altitudes does not exactly coincide with that of Costa Rica, as altitudes between 500m and 1000m are strongly represented among the stations, while large costarican plains constitute a wide part of the country. The six stations (69507, 75022, 76026, 77001, 84118, and 98006) were chosen to represent a variety of locations, while sharing similar characteristics. For example, stations 98006 and 75022 have a very similar elevation, but a different drainage basin. This allowed us to identify useful covariates for univariate and multivariate analysis of extremes, such as adding longitude and latitude in addition to altitude.

Figure A.2 shows the total number of observations for each month, duration and station, across the entire dataset. The period from December to February is marked by fewer observations available, either due to missing values or because no precipitation was recorded.

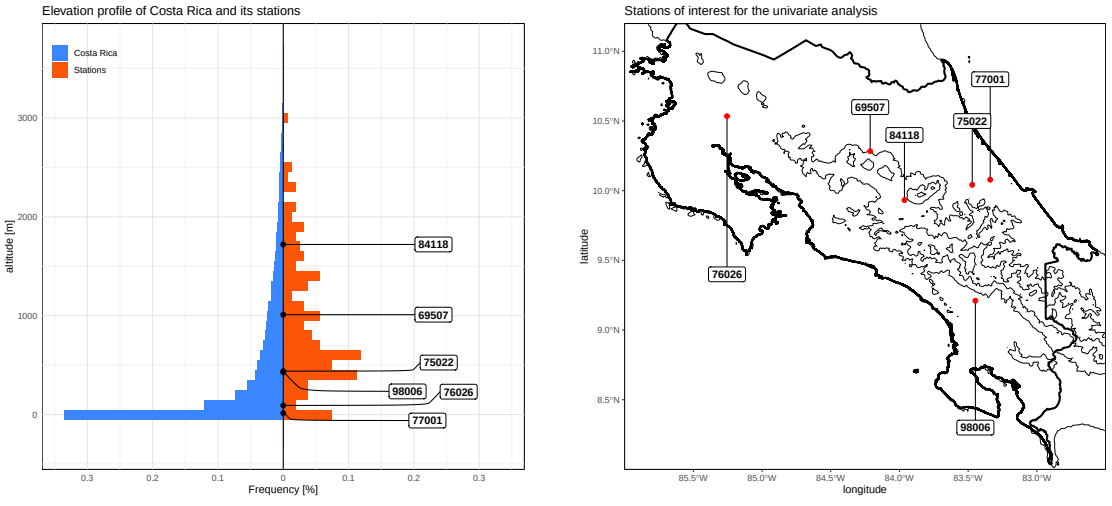


Figure A.1: Elevation profile of Costa Rica and the stations. The elevations range from 0m (sea level) to over 3700m, and the highest station is located at 3652m. A large part of Costa Rica has an elevation below 500m, while most hydrological stations are at elevation 400–1000m. The right panel shows the six stations studied in details during the univariate analysis, along with 1000m elevation contours (black lines).

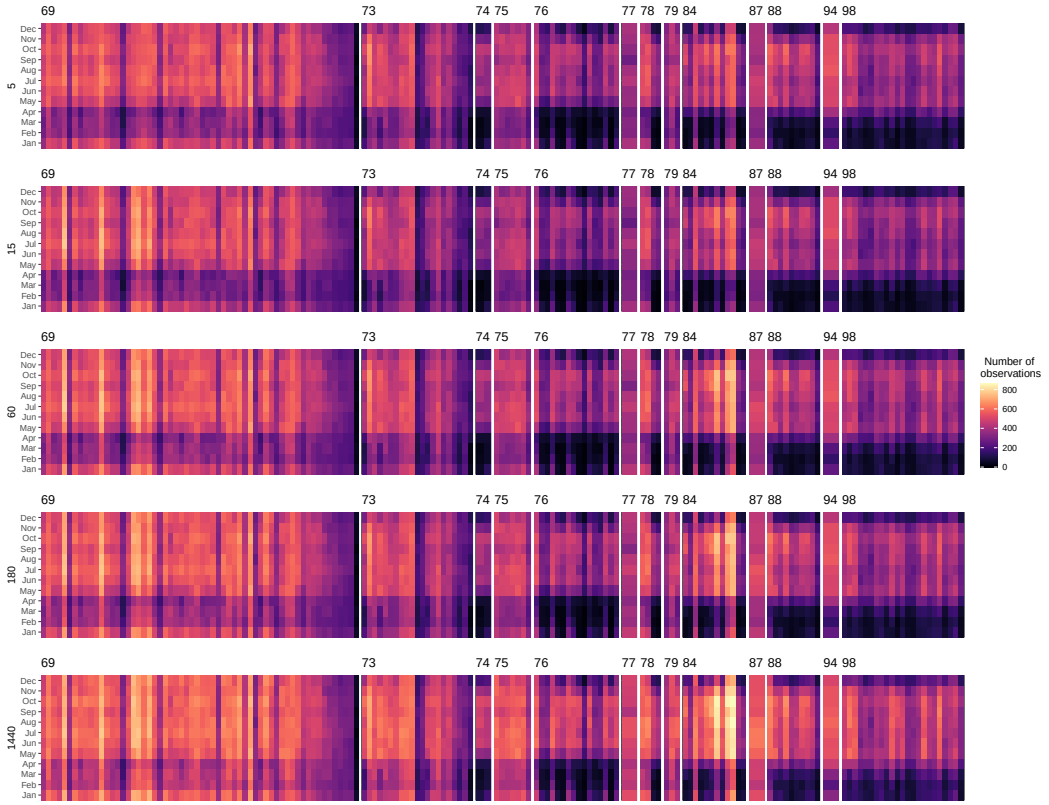


Figure A.2: Total number of observations across the month, stations and durations. Darker regions are linked with fewer observations. Stations are grouped here by drainage basin.

Appendix B

Univariate analysis

This section groups estimated IDF curves of the six stations studied in detailed through this report, along with maps of the 20-year return levels derived from the k-NN and CDN approaches (Figures B.9 and B.10).

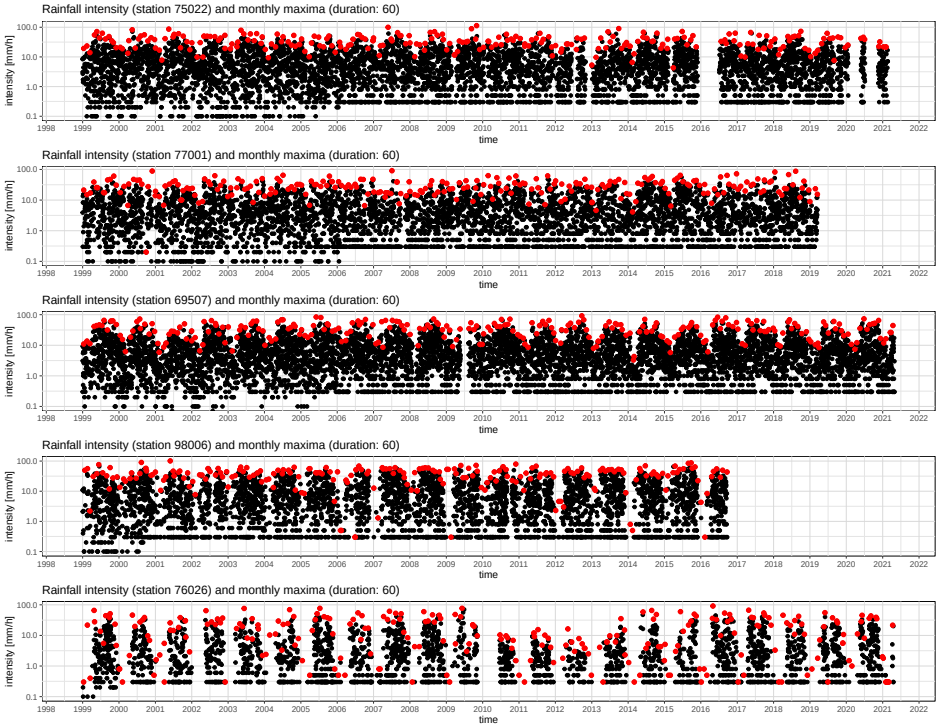
Figure B.1 shows the daily rainfall intensities of several stations, and highlights the set of monthly maximum intensities. Seasonality is clearly visible, and some stations (98006 and 77001 for example) have not provided any measurement for more than three years.

Figures B.2 to B.7 present estimated IDF curves for January, April, July and October, for the six stations studied in Section 3.2. These curves were obtained based on the model selection described in Section 3.2, for which the QQplot are also given hereafter. In the following, we provide a brief comparison of the IDF curves. The stations 69507, 75022 and 77001 have similar IDF curves (both from a quantitative and qualitative point of view), although they have different elevation (respectively 1010, 438 and 15m). The 20-year return levels reach their highest values between December and January for long durations (720 to 1440 minutes), and between May and October for lower durations. Station 77001 has the weakest seasonality effect among the six stations, and its highest 20-year return levels are all reached between November and January. The QQplots show that models fit the data reasonably well for all durations, except 5 minutes, for which confidence bands (not shown in the QQplots for clarity) exclude the diagonal. It should be noted that these three stations are on the Caribbean side of Costa Rica.

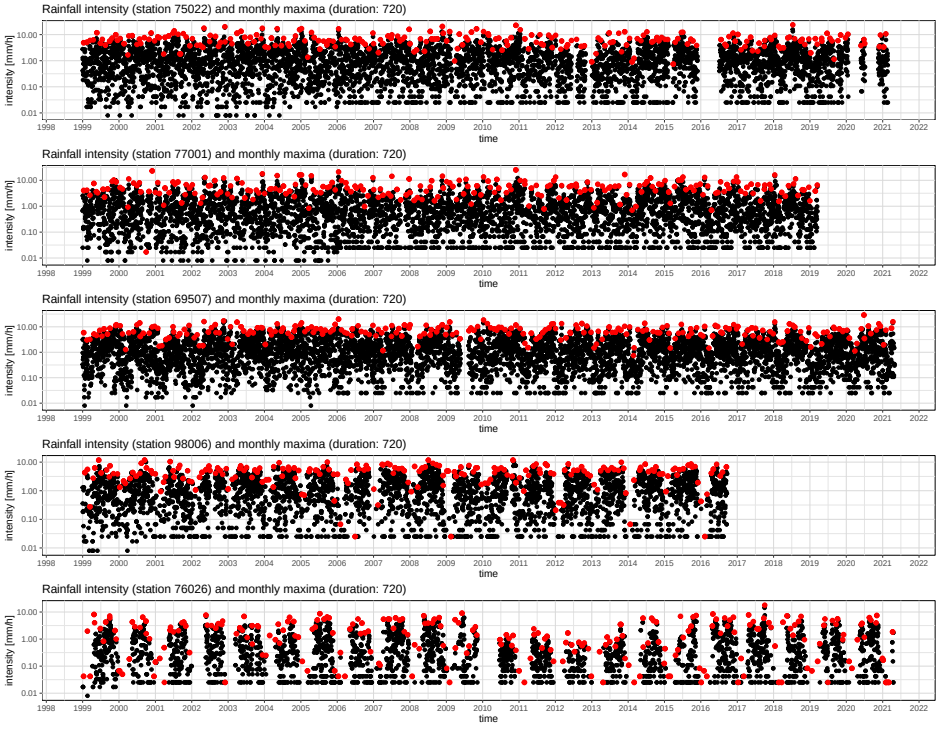
Stations 76026, 84118 and 98006 have a stronger seasonality; especially station 76026 (in the province of Guanacaste), which has the lowest and highest return levels among all stations. Estimated IDF curves show that this station can reach 20-year return levels near 1000mm/hr for the smallest durations between June and September. However, the fit for this station is questionable, especially for 5 to 15 minutes durations. Stations 98006 and 84118 have similar IDF curves, characterized by almost constant return levels between May and October (except the Little summer effect that occurs only for station 84118 in this case).

Figure B.8 presents the expected quantile score index between each pair of models described in Chapter 3: the reference model (with a separate model fitted for each station), the k-NN and CDN approaches. The approach involving the conditional density network shows improvements for large quantiles (typically 0.999) compared to the reference and the k-NN models.

Lastly, QQplots for the k-NN and CDN models are illustrated by Figures B.11 and B.12 respectively, and show that the two approaches are adequate to model monthly maximum intensities.

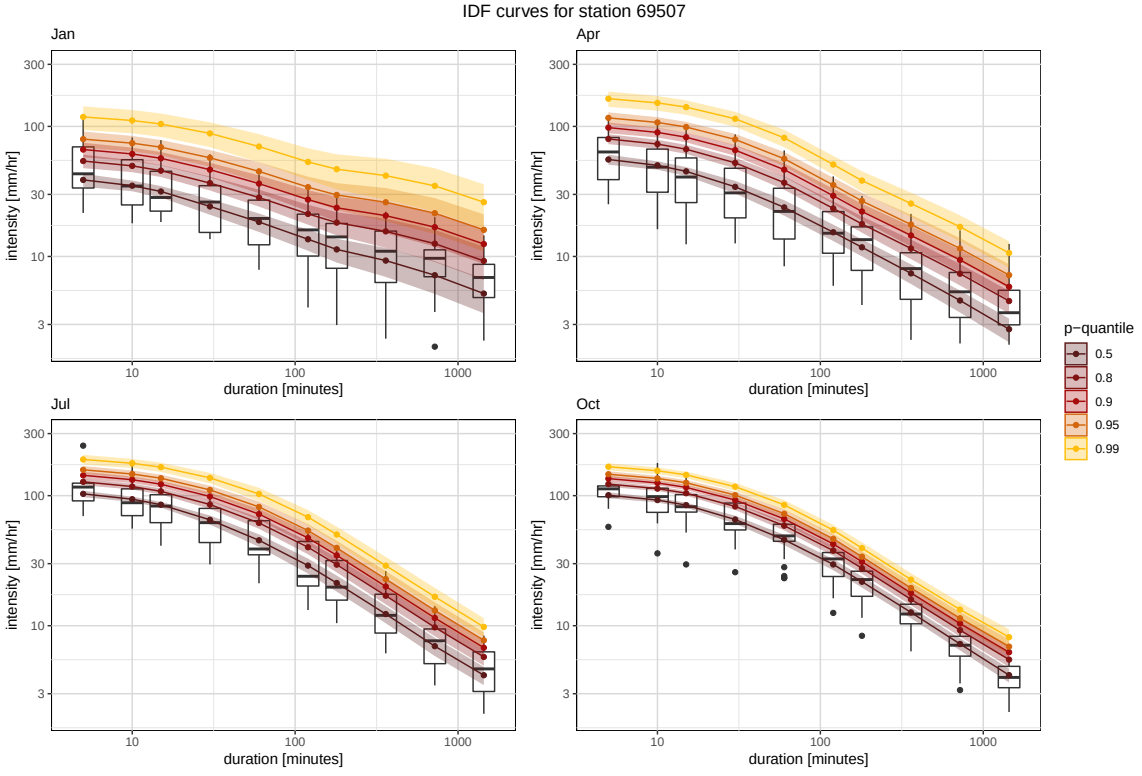


(a)

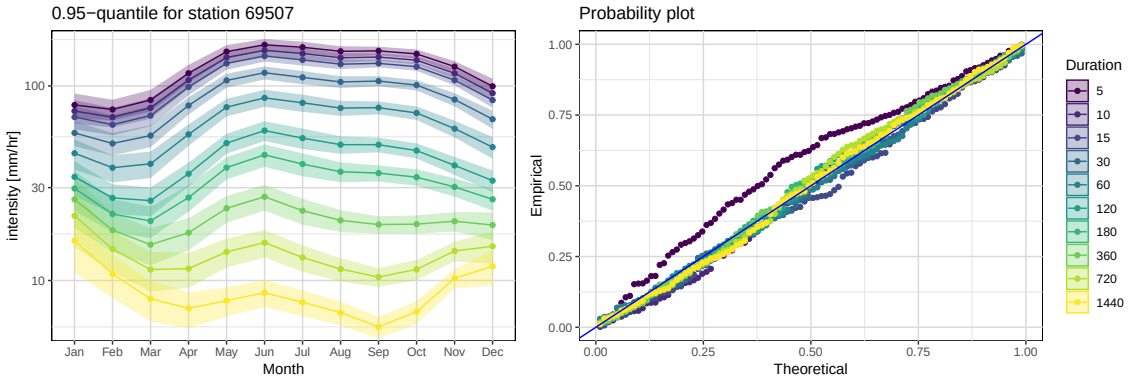


(b)

Figure B.1: Rainfall intensities for durations 60 and 720 minutes, for 5 different stations. Black points correspond to daily measurements, and red ones to the monthly maxima.

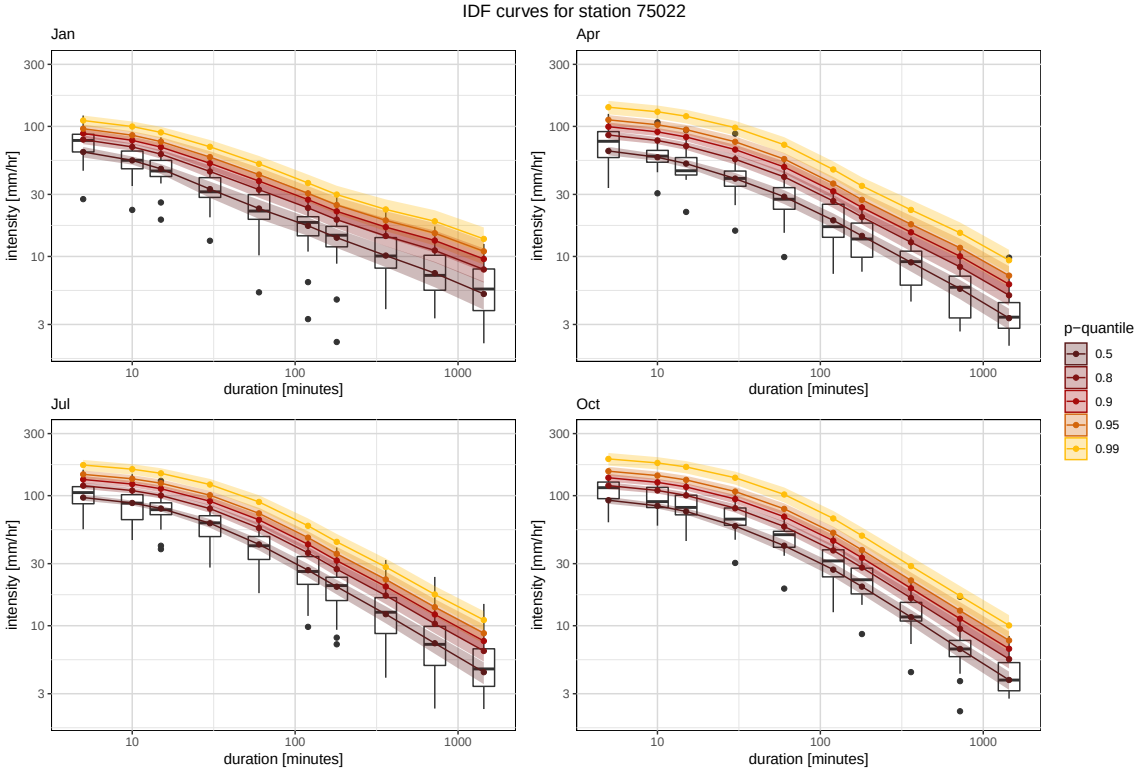


(a)

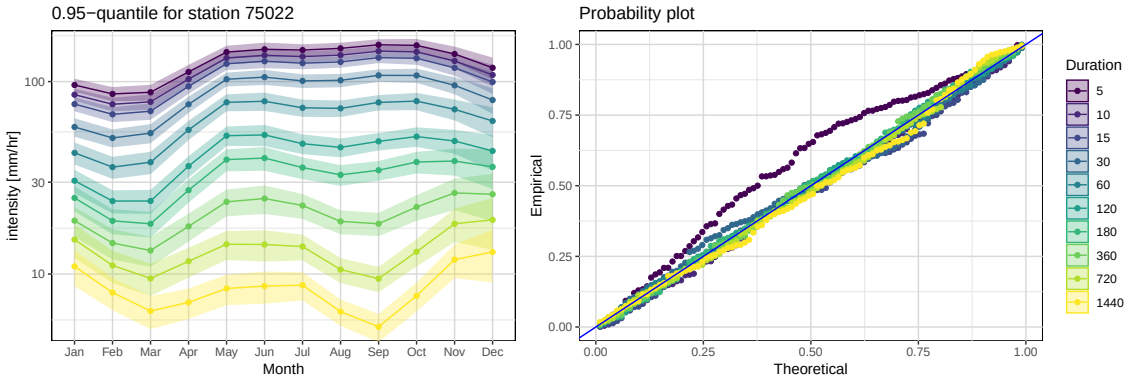


(b)

Figure B.2: IDF curves for station 69507, for January, April, July and October (a). 20-year return levels across months and durations are shown in the bottom (b) left panel, and reveal different patterns depending across durations. QQplot for the GEV model fitted in Section 3.2 is shown in the bottom (b) right panel.

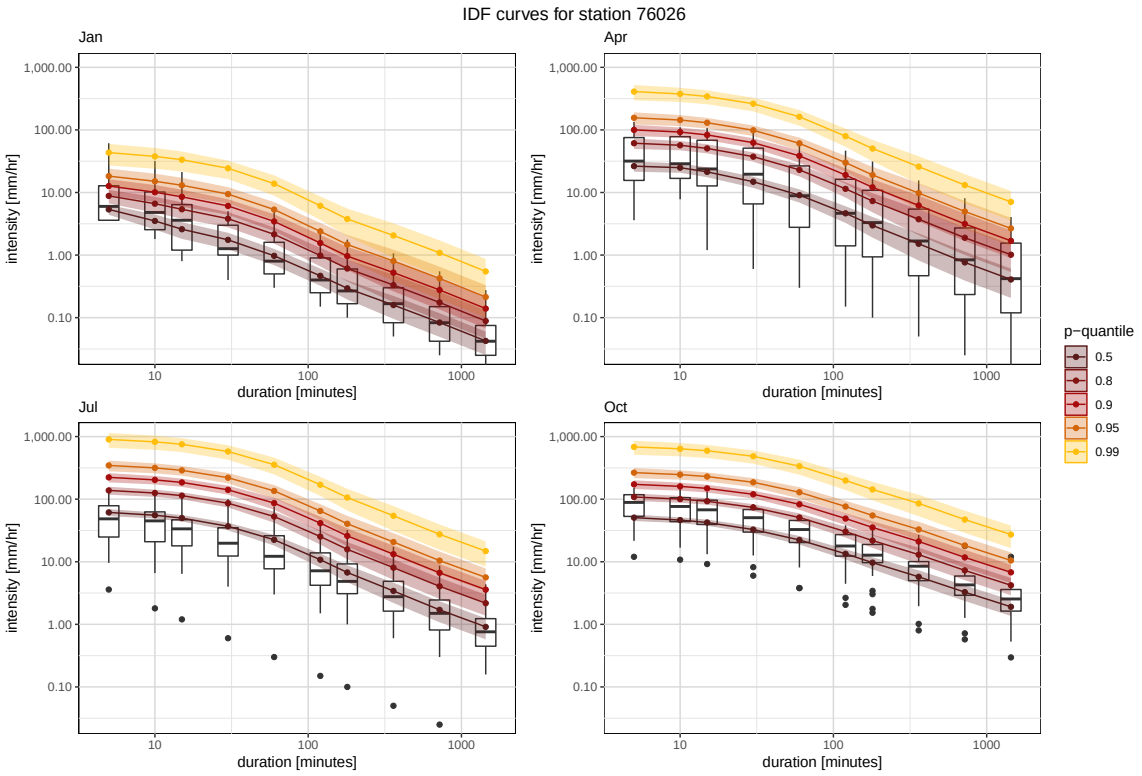


(a)

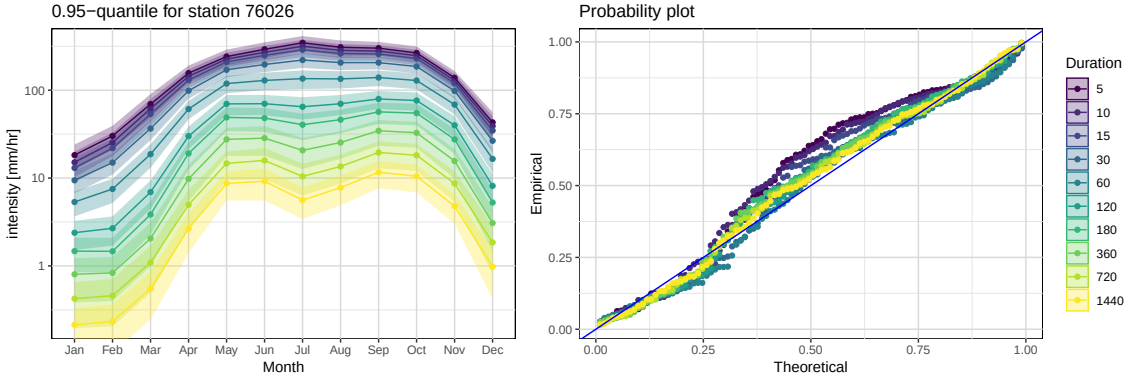


(b)

Figure B.3: IDF curves for station 75022, for January, April, July and October (a). 20-year return levels across months and durations are shown in the bottom (b) left panel, and reveal different patterns depending across durations. QQplot for the GEV model fitted in Section 3.2 is shown in the bottom (b) right panel.

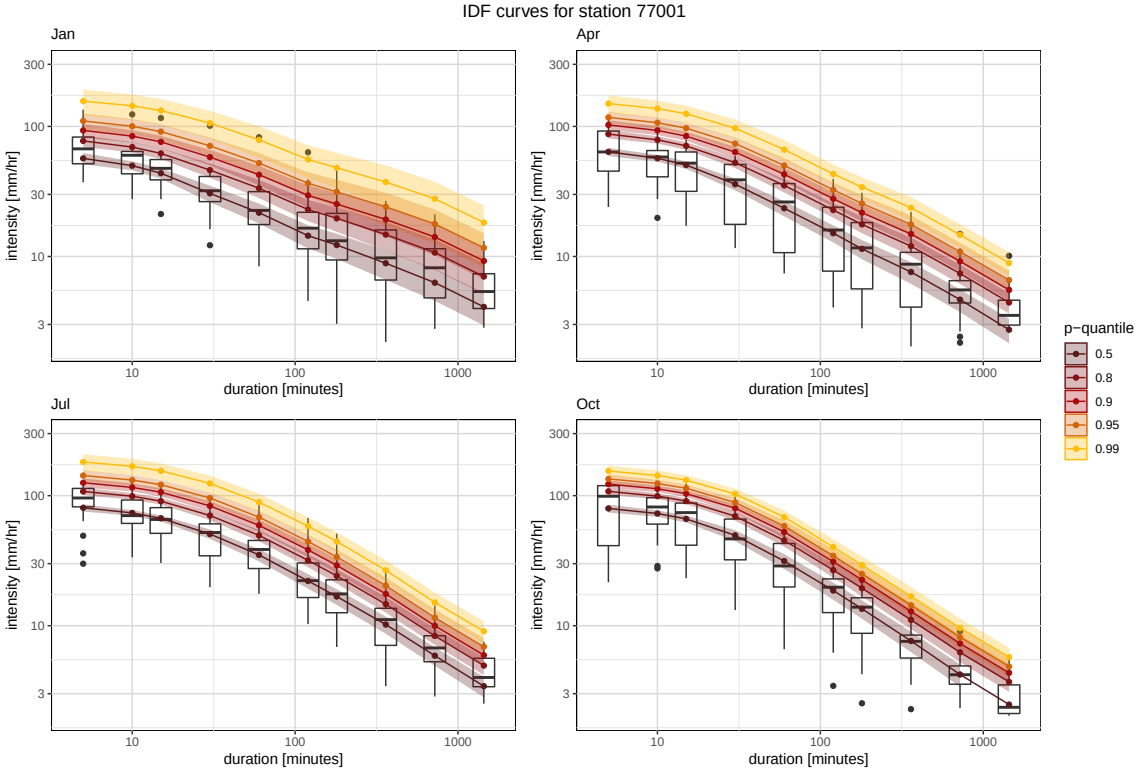


(a)

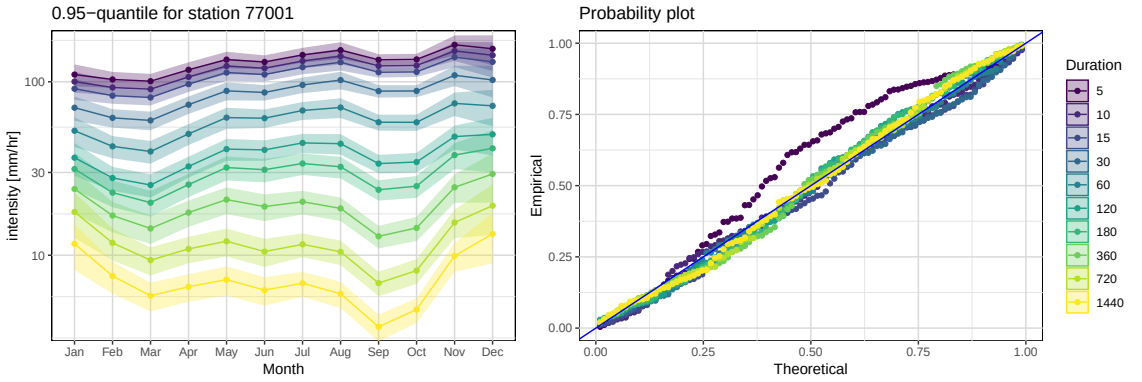


(b)

Figure B.4: IDF curves for station 76026, for January, April, July and October (a). 20-year return levels across months and durations are shown in the bottom (b) left panel, and reveal different patterns depending across durations. QQplot for the GEV model fitted in Section 3.2 is shown in the bottom (b) right panel.

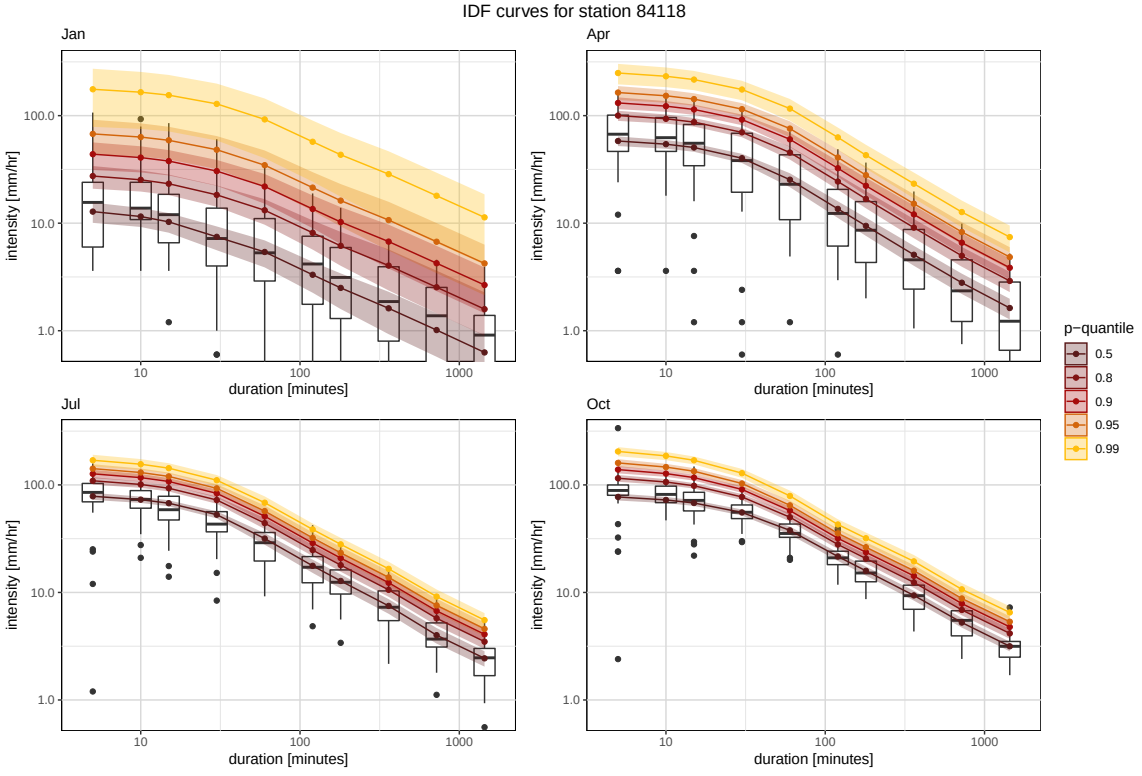


(a)

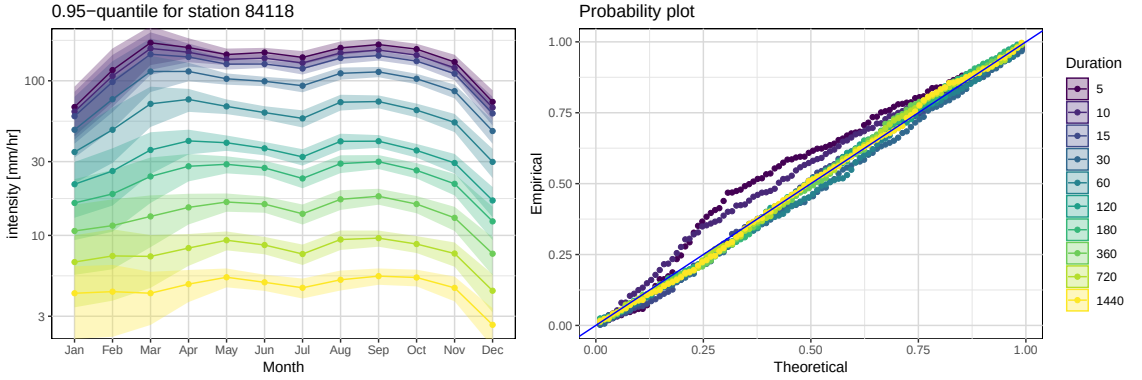


(b)

Figure B.5: IDF curves for station 77001, for January, April, July and October (a). 20-year return levels across months and durations are shown in the bottom (b) left panel, and reveal different patterns depending across durations. QQplot for the GEV model fitted in Section 3.2 is shown in the bottom (b) right panel.

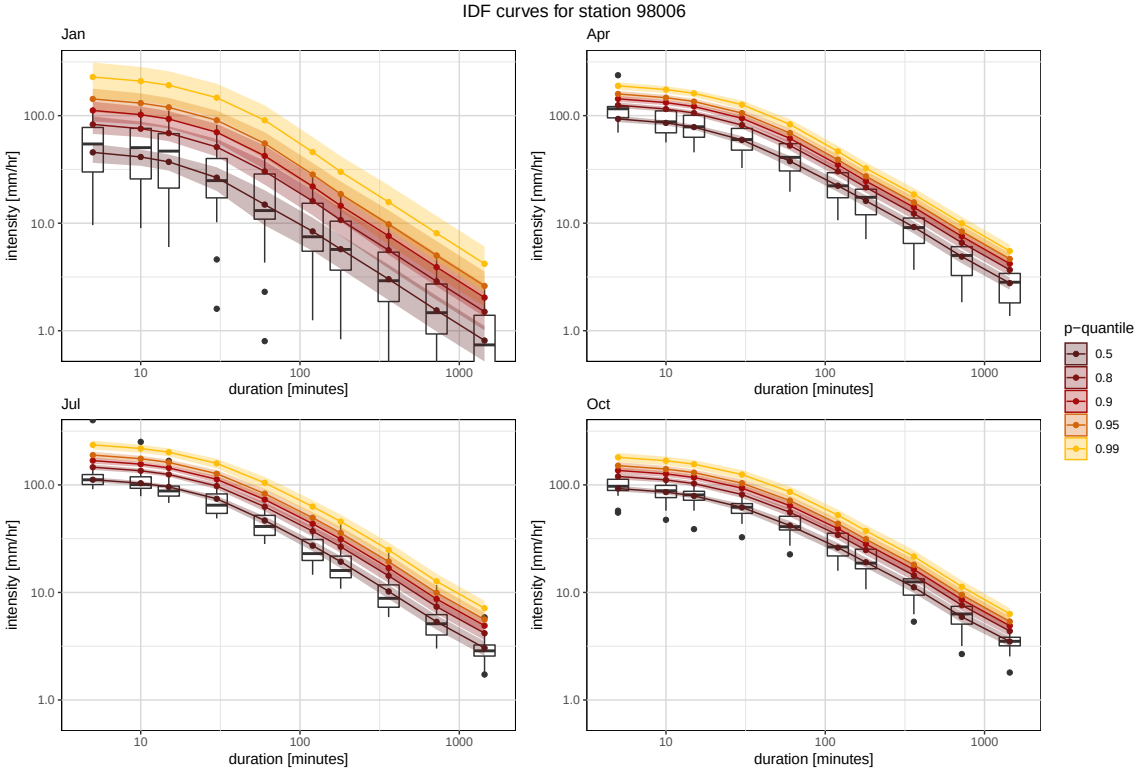


(a)

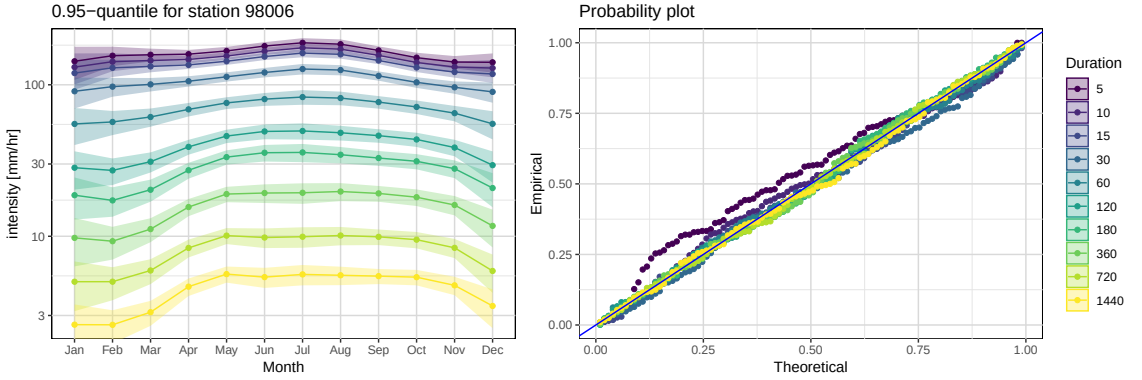


(b)

Figure B.6: IDF curves for station 84118, for January, April, July and October (a). 20-year return levels across months and durations are shown in the bottom (b) left panel, and reveal different patterns depending across durations. QQplot for the GEV model fitted in Section 3.2 is shown in the bottom (b) right panel.



(a)



(b)

Figure B.7: IDF curves for station 98006, for January, April, July and October (a). 20-year return levels across months and durations are shown in the bottom (b) left panel, and reveal different patterns depending across durations. QQplot for the GEV model fitted in Section 3.2 is shown in the bottom (b) right panel.

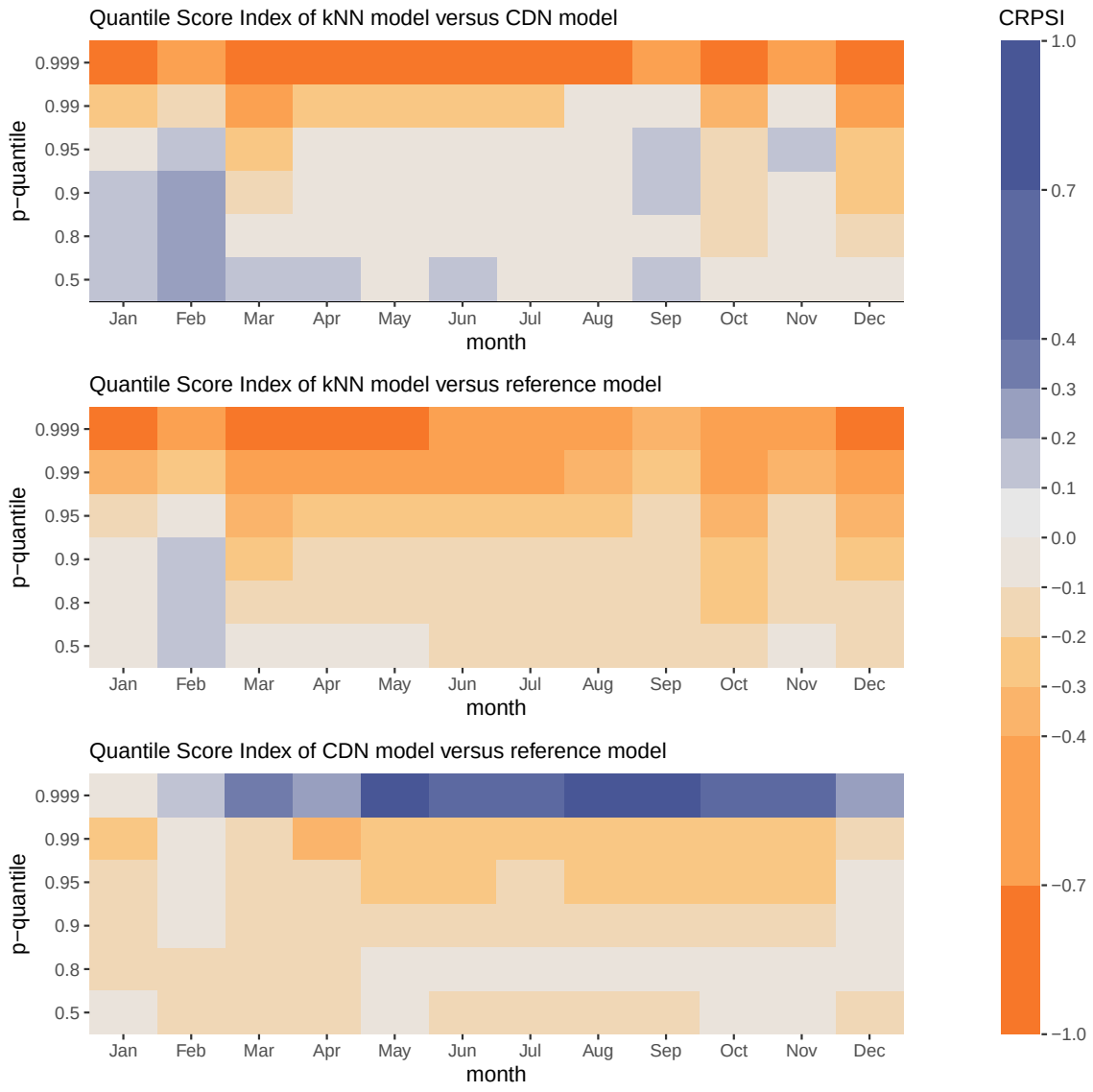


Figure B.8: Expected Quantile Score Index (QSI) for several months and quantiles, for the reference, k-NN and CDN models. The reference model correspond to the seasonal GEV model from Section 3.2. The CDN-based approach performs better than the k-NN based one, especially for high quantiles. This is also the case compared to the reference model, although only for the quantile 0.999.

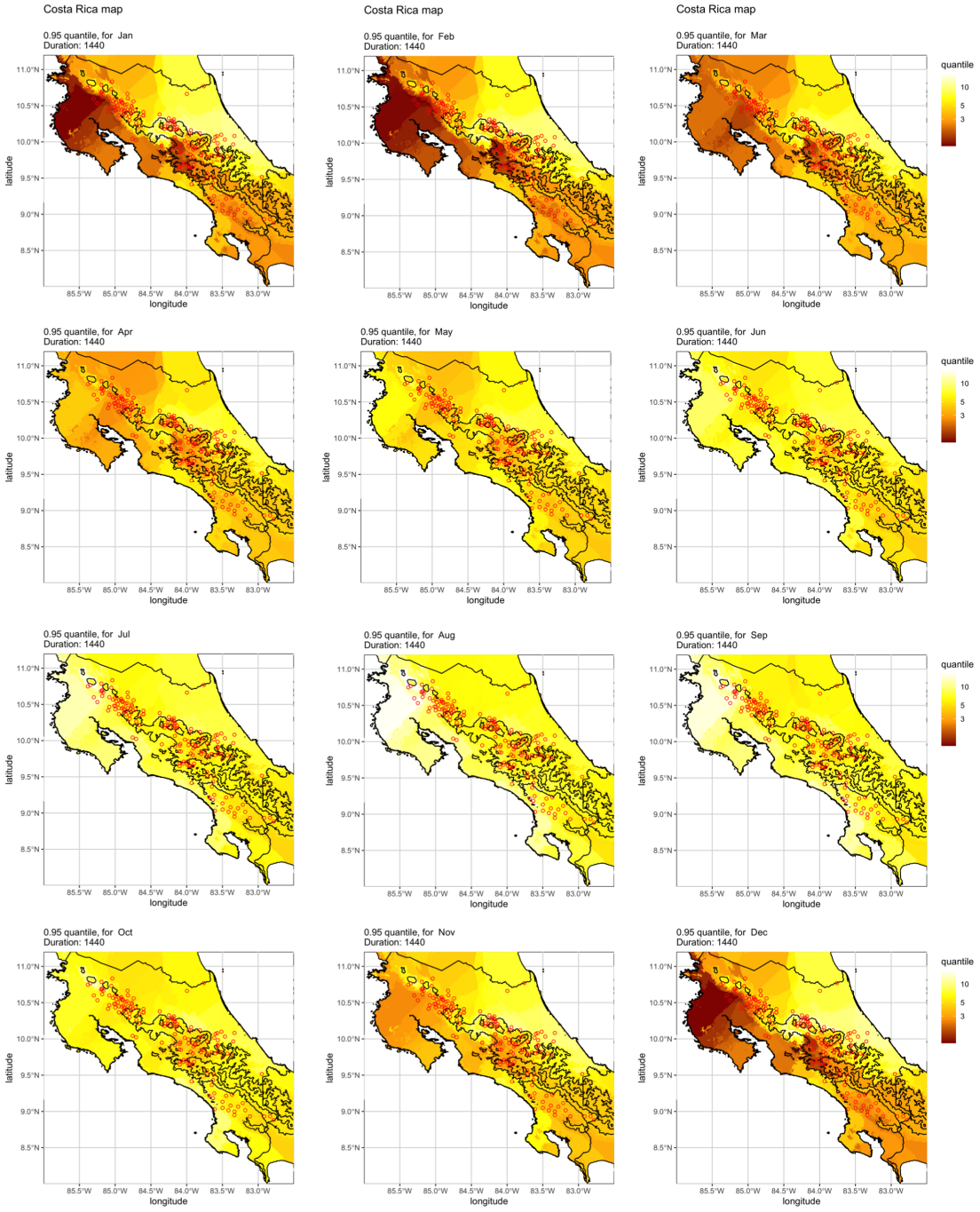
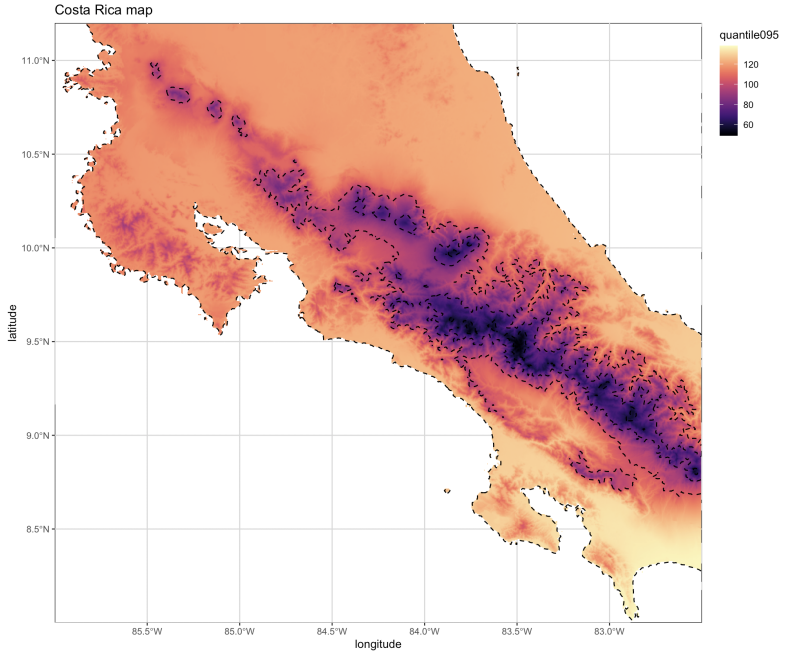
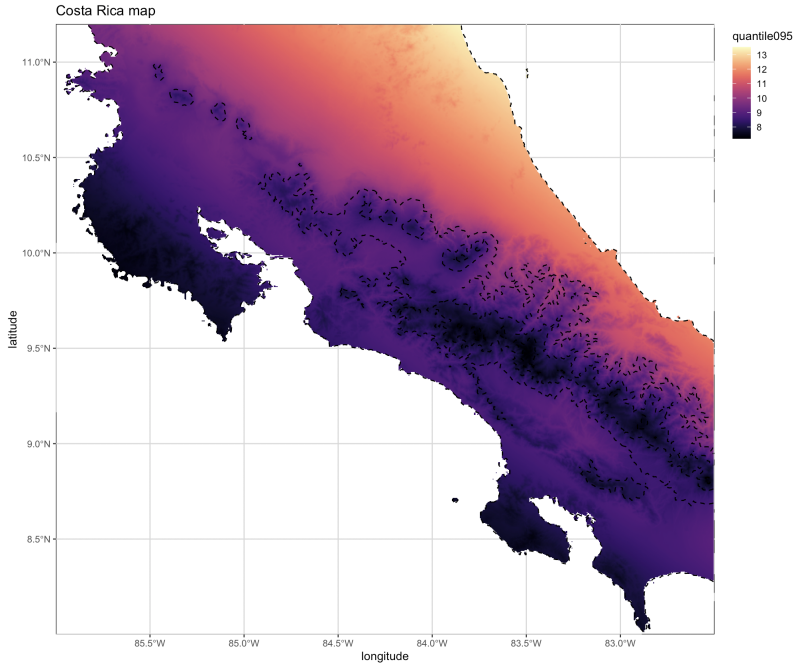


Figure B.9: Extension of the 20-year return levels over Costa Rica with the k-NN approach, for duration 1440 minutes. The province of Guanacaste clearly has the highest seasonality, with the highest and lowest return levels observed over the entire country. The center of Costa Rica (north of the province of San José) has lower return levels compared to any other location, for all months. Lastly, the province of Limón (Caribbean coast) has weak seasonality, with return levels that are almost identical over the year.

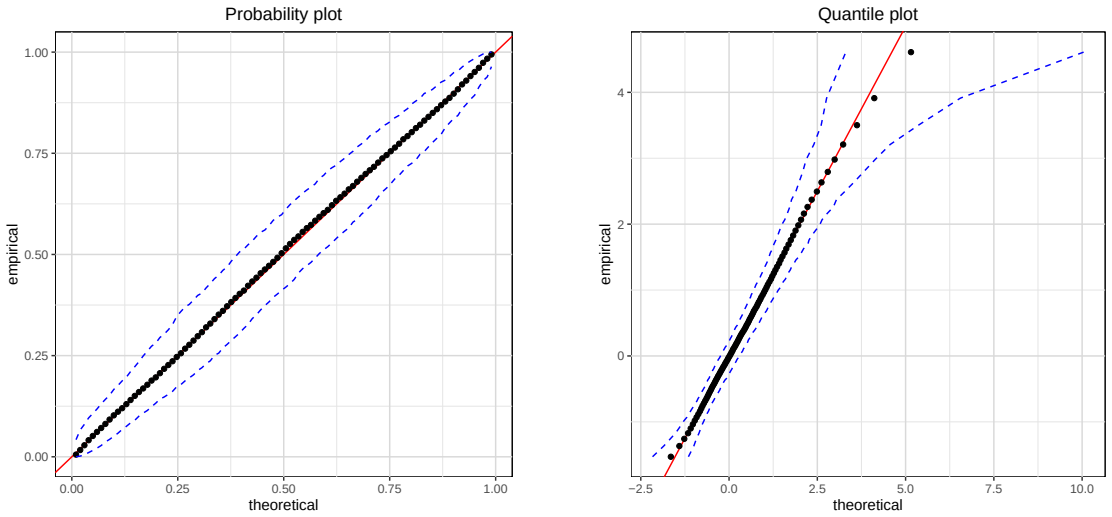


(a)



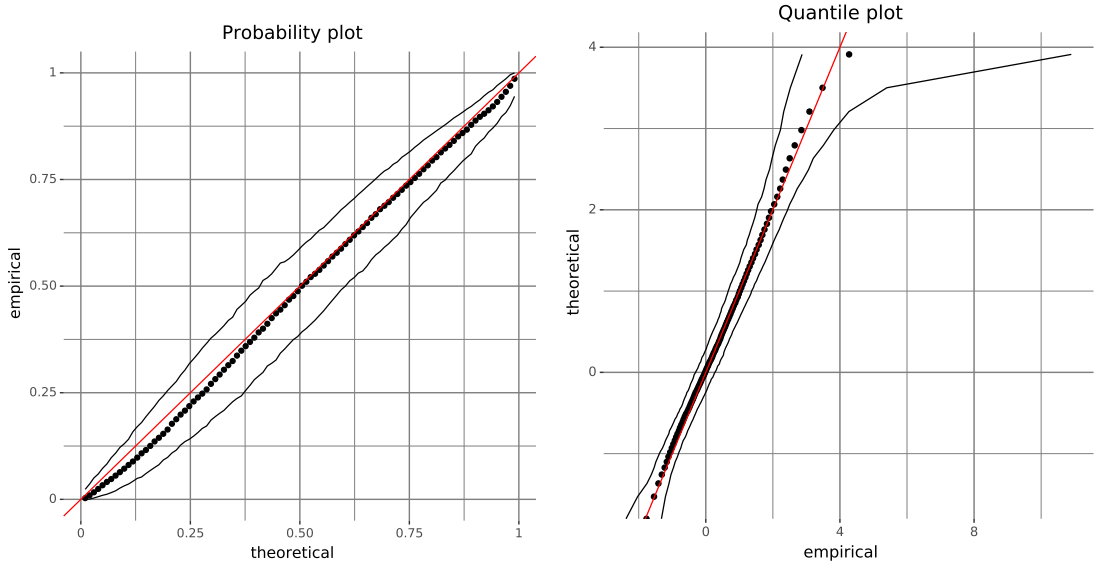
(b)

Figure B.10: 20-year return level plots for duration 60 (a) and 1440 (b) minutes with GEV parameters predicted by a neural network, taking as input the longitude, latitude, altitude, duration and month. Months were transformed with a Fourier basis. Lower durations seem to depend more heavily on the altitude compared to higher durations. A clear seasonality is present, and for high durations, a demarcation between the Pacific and Caribbean coasts. Return levels for the Caribbean coast increase as we go from south to north along this coast.



(a) (b)

Figure B.11: Probability and quantile plots for the univariate k-NN model. Blue dotted lines correspond to the 95% confidence bands obtained with bootstrapping.



(a) (b)

Figure B.12: Probability and quantile plots for the univariate CDN model. Black lines correspond to the 95% confidence bands obtained with bootstrapping.

Appendix C

Bivariate analysis

This section shows additional figures regarding the bivariate analysis of extremes. Figure C.1 shows estimates of χ and $\bar{\chi}$ for the pairs 75022–77001 and 75022–98006. Dependence emerges only for the former one, as the estimates deviate significantly from 0.

Figures C.2 and C.3 show the mean and standard errors of the bivariate model extended with a conditional neural network, for 1440 minute durations. Yellow regions indicate higher extremal dependence with respect to the location of reference, that is shown in white. Uncertainty is large at locations with few precipitation gauge around.

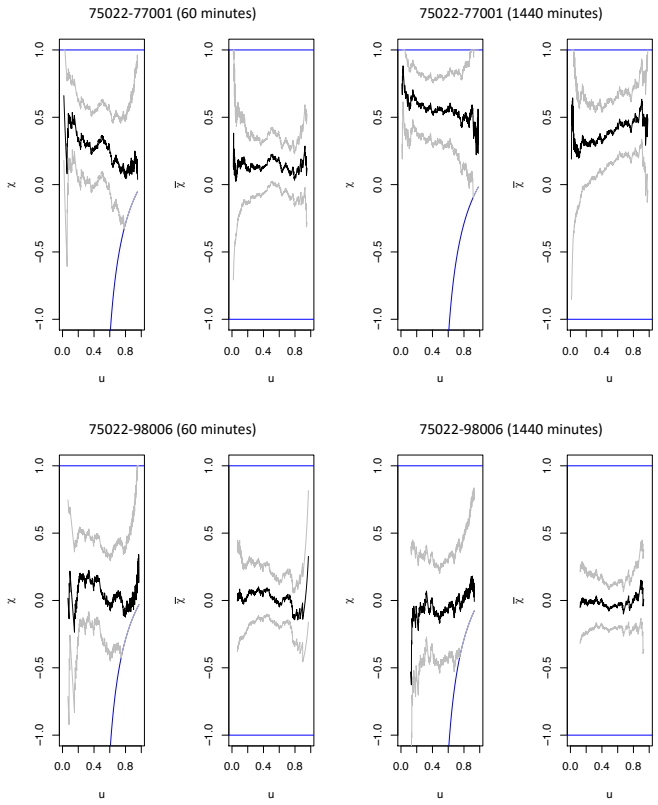


Figure C.1: χ and $\bar{\chi}$ plots for the pairs 75022–77001 and 75022–98006, for duration 60 and 1440 minutes. Grey lines correspond to approximate 95% confidence intervals, and the blue lines to the upper and lower bounds. As highlighted, dependence emerges among these four cases only for high durations and the pair of stations 75022–77001.

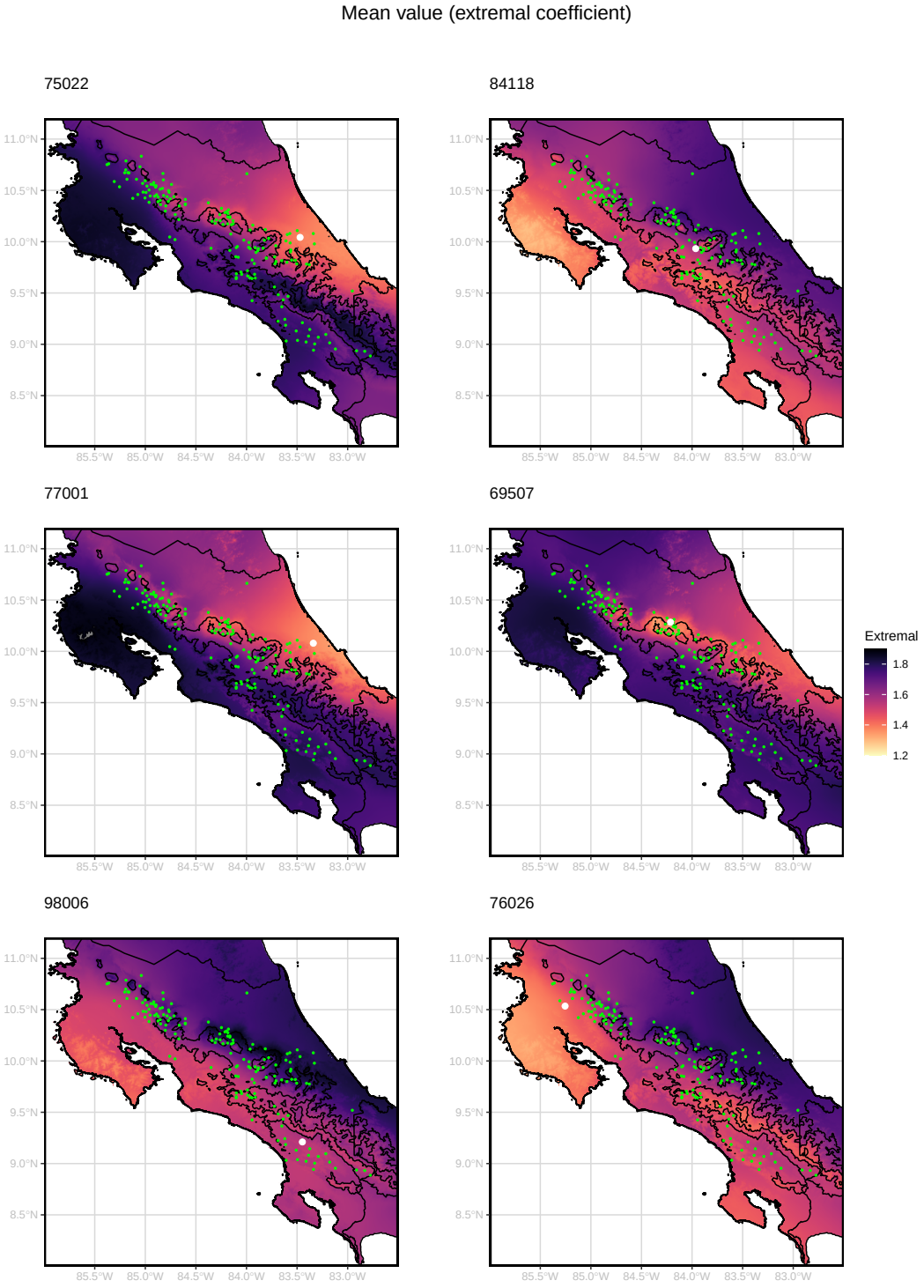


Figure C.2: Mean extremal coefficient from the extended bivariate approach (and duration 1440 minutes), with the six stations taken as reference precipitation gauge. Green points show the other locations, and the white point the station of reference. One can distinguish overall a clear Pacific/Caribbean demarcation with respect to extremal dependence (as highlighted as well by the clustering analysis). Stations 84118 and 98006 do not have a strong extremal dependence with its neighbouring stations, unlike the other ones.

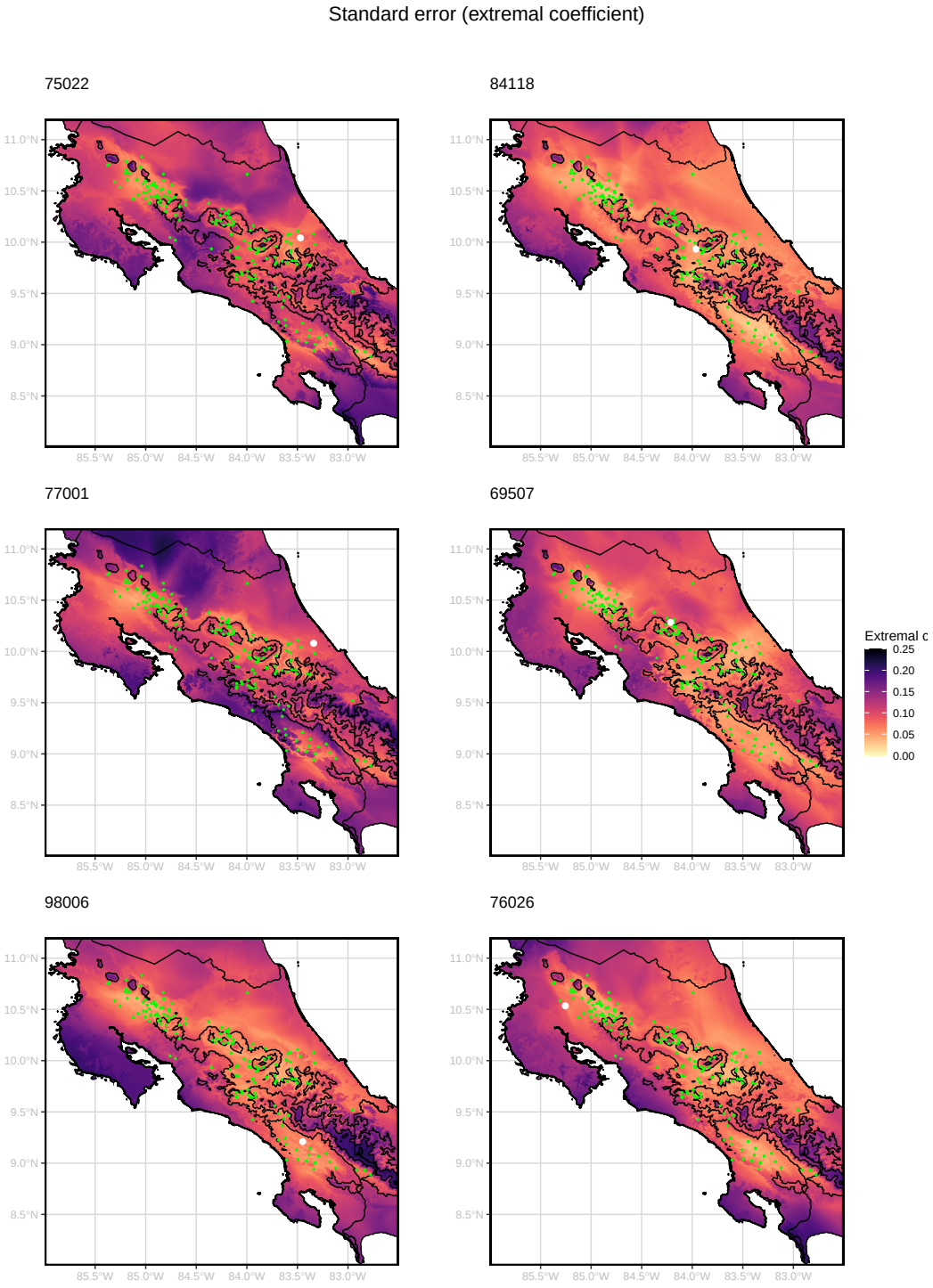


Figure C.3: Standard errors of the estimated extremal coefficient from the extended bivariate approach (and duration 1440 minutes), with the six stations taken as reference precipitation gauge. Green points show the other locations, and the white point the station of reference. As expected, regions with very few stations have a variance that increases (in dark purple). This includes most of the Guanacaste province, the north Caribbean coast and the highly mountainous part of Costa Rica. For these regions, the variance may be so high that the two bounds of the extremal coefficient are included.

Appendix D

Clustering analysis

In section 5.2.1, we used a hierarchical clustering approach based on the extremal dependence, estimated with the F-madogram. In this section, we compare the results based on another distance matrix, the Tail pairwise dependence matrix (TPDM) [Cooley and Thibaud, 2018], whose symmetry and positive definiteness make it an attractive tool. Jiang et al. [2020] used it in conjunction with a principal component analysis to analyze U.S. extreme precipitations. Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a p-dimensional regularly varying random vector with index $\alpha = 2$ and angular measure H. The TPDM Σ_X is defined as

$$(\Sigma_X)_{i,j} = \int_{\mathbb{S}} w_i w_j \, \mathrm{d}H(w), \quad i, j = 1, \ldots, p, \tag{D.0.1}$$

with $w_i = X_i / \|\boldsymbol{X}\|$ and $\mathbb{S} = \{\boldsymbol{w} \in \mathbb{R}_+^p : \|\boldsymbol{w}\| = 1\}$. More details about regularly varying random variables can be found in Resnick [2002]. Let $\{\boldsymbol{x}_n\}_{n=1}^N$ be the set of observations in $\mathbb{R}^p$ (typically, p is the number of stations) transformed so that their marginal distributions are regularly varying with index $\alpha = 2$. As described in Jiang et al. [2020], this can be obtained with the transformation

$$\tilde{x}_{n,i} = \{-\log \hat{F}_i(x_{n,i})\}^{-1/2}, \tag{D.0.2}$$

where $\{x_{n,i};\, i = 1, \ldots, p;\, n = 1, \ldots, N\}$ are the original observations and $\{\hat{F}_i(\cdot)\}_{i=1}^p$ the set of empirical marginal cumulative distribution functions. Let $r_{n,ij} = \{\tilde{x}_{n,i}^2 + \tilde{x}_{n,j}^2\}^{1/2}$ and $(w_{n,i}, w_{n,j}) = (\tilde{x}_{n,i}, \tilde{x}_{n,j})/r_{n,ij}$. The entries of the TPDM can be estimated as

$$(\hat{\Sigma}_X)_{i,j} = \frac{2}{\tilde{n}_{i,j}} \sum_{k=1}^{\tilde{n}_{i,j}} w_{k,i} w_{k,j} \mathbb{1}_{(r_{n,ij} > r_{0,ij})}, \tag{D.0.3}$$

with a high threshold $r_{0,ij}$ that is exceeded $\tilde{n}_{i,j}$ times. This threshold may be set a priori, for example by taking the highest 5% values, or by a more careful analysis that would ensure the stability of the estimates above the chosen threshold. When the dimension p is high, this might be hard to implement in practice.

For our study, we considered $p = 160$ (the number of stations), and we adopted the 5% rule stated above to select the threshold. The distance matrix d for the hierarchical cluster is defined as $d(x_i, x_j) = -\log(\hat{\Sigma}_X)_{i,j}$. A comparison of the clusters found based on the F-madogram and the TPDM distances is shown in Figures D.1, D.2 and D.3. Overall, the approach based on the F-madogram defines regions that are more sensible for any multivariate extreme value analysis, as they tend to share similar topographical properties, unlike the TPDM-based approach, which finds dislocated regions across the map, especially for low to medium durations. When only two clusters are allowed, the demarcation is made between

the Pacific and Caribbean coasts, which is especially visible for high durations. When more clusters are allowed, one can observe more interesting patterns and relationships between the precipitation gauges. In particular, stations from the central region of San Jose share similar properties with stations from basin 76 or the northern part of basin 69. One should highlight that the basins shown on the map do not exactly match basins given in the dataset. Some stations from basin 69 and 76 are grouped in the same basin on the map, but not in the dataset, which is backed by the clustering analysis shown here, that clearly separates this basin into two parts. Overall, both approaches roughly agree on the regionalization for high durations.

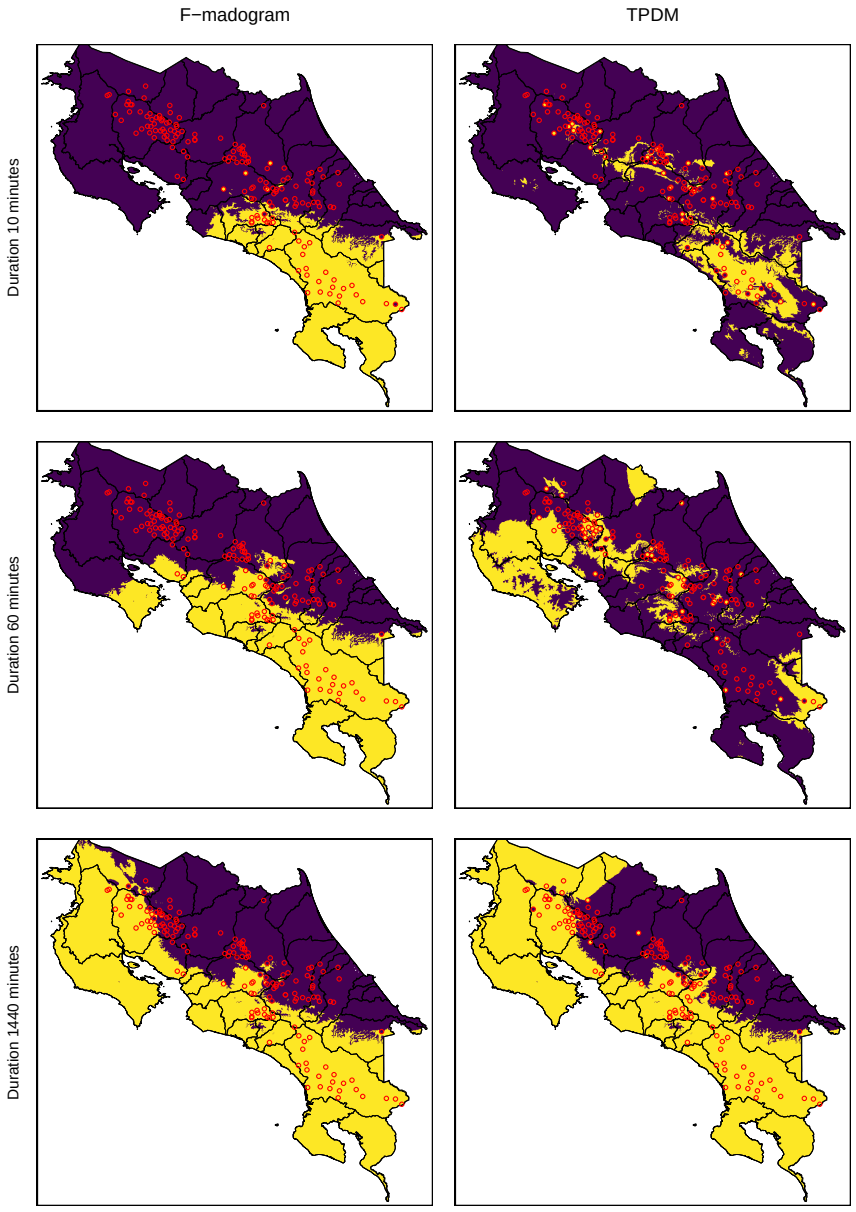


Figure D.1: Regionalization analysis with two clusters. The approach that uses the F-madogram to compute the distance matrix clearly delimits the Pacific and Caribbean basins, especially for high durations. This is visible for the TPDM approach only for large durations.

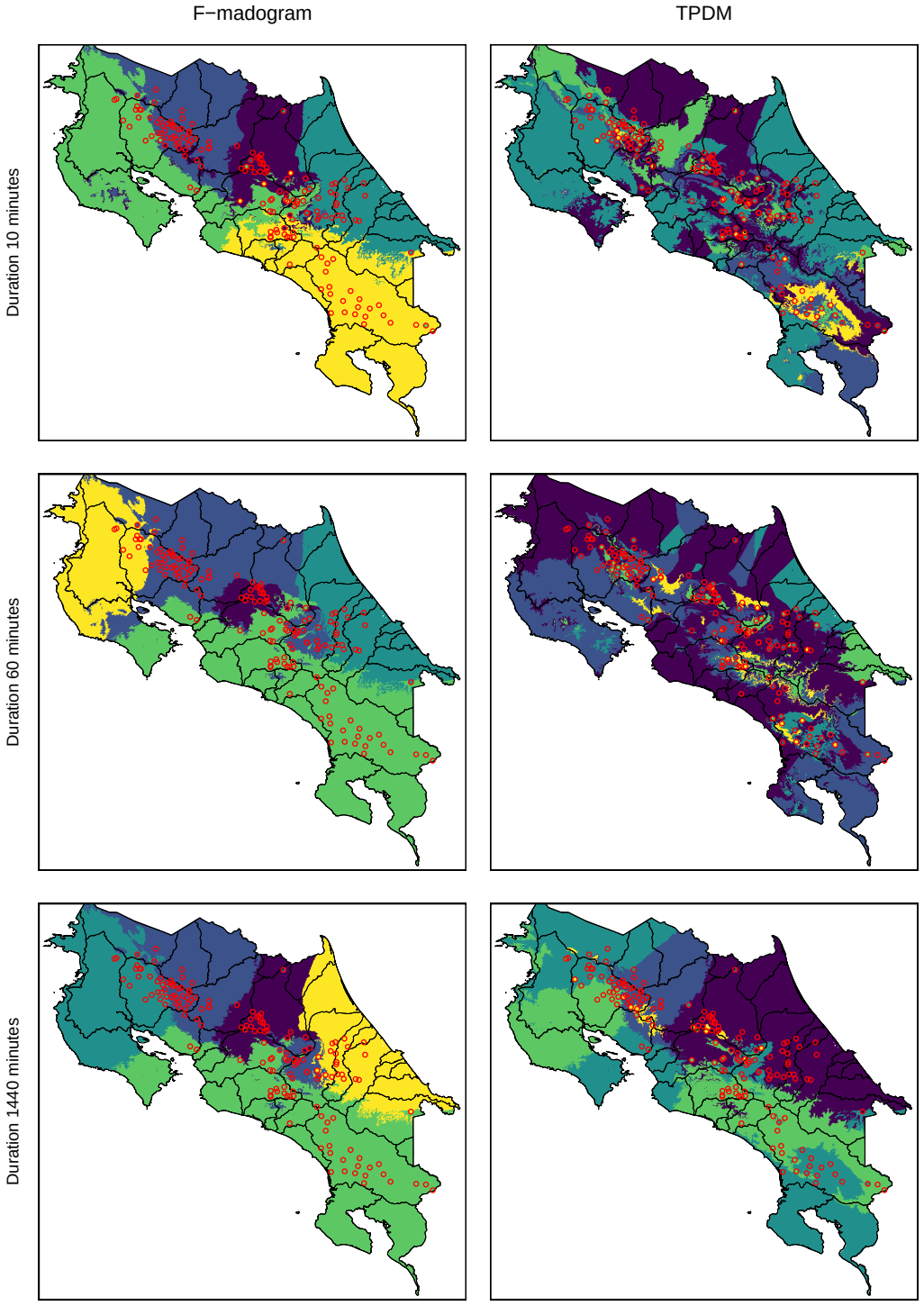


Figure D.2: Regionalization analysis with five clusters. The F-madogram approach still provides homogeneous regions as the number of clusters increases, while the TPDM approach struggles to provide a meaningful regionalization for low and medium durations.

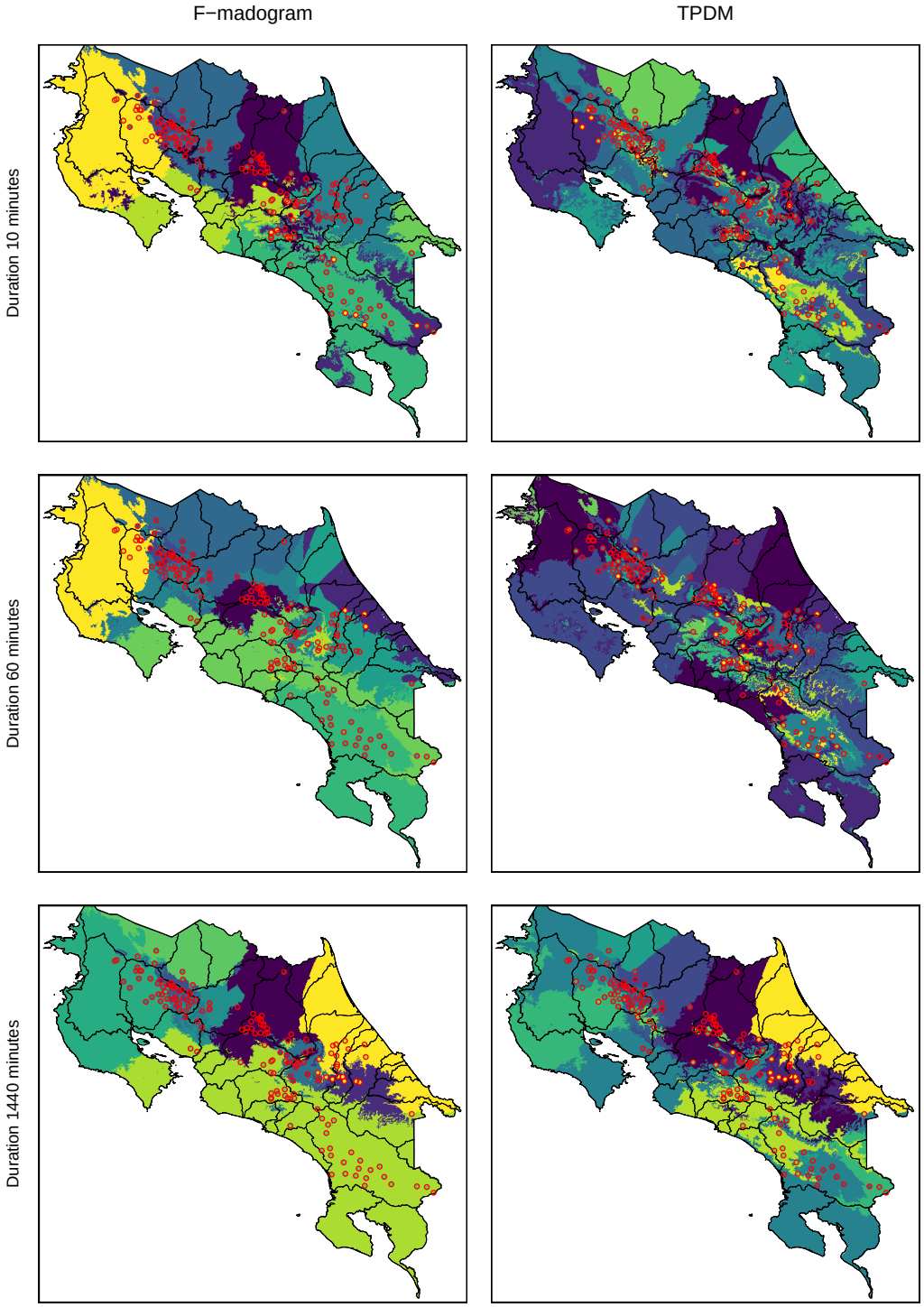


Figure D.3: Regionalization analysis with ten clusters. The F-madogram approach again provides homogeneous regions as the number of clusters increases, while the TPDM approach struggles to provide a meaningful regionalization for low and medium durations.

Appendix E

Non-stationary max-stable process

This section gathers some of the results from the max-stable approach described in Chapter 5. Figures E.1 and E.2 show the behaviour of the empirical extremal coefficient with the distance and the altitude, based on Equation (4.2.4), with respect to each one of the six stations. Overall, the altitude does not seem to fully explain the behaviour of extremal dependence. Asymptotic independence emerges as distance increases, but this effects is weak for most of the stations. Longer durations (720–1440 minutes) tend to exhibit more asymptotic dependence than the other durations for small distances, but this phenomenon is reversed as the distance grows.

Figure E.3 presents the mean elements of the covariance matrix Ω_x from Section 5.2.2, and Figure E.4 the ellipsoids generated by this covariance matrix. Stations located on the Pacific coast tend to have a dependence structure that stretches along the coast, while ellipses are more circular for the Caribbean coast. The extremal coefficient might not exhibit circular shape in this case, as the weight function $a(\cdot)$ depends on the location and impacts the correlation function.

In Section E.1, one can found the derivation of the bivariate density for the extremal t process that is used for the log likelihood function.

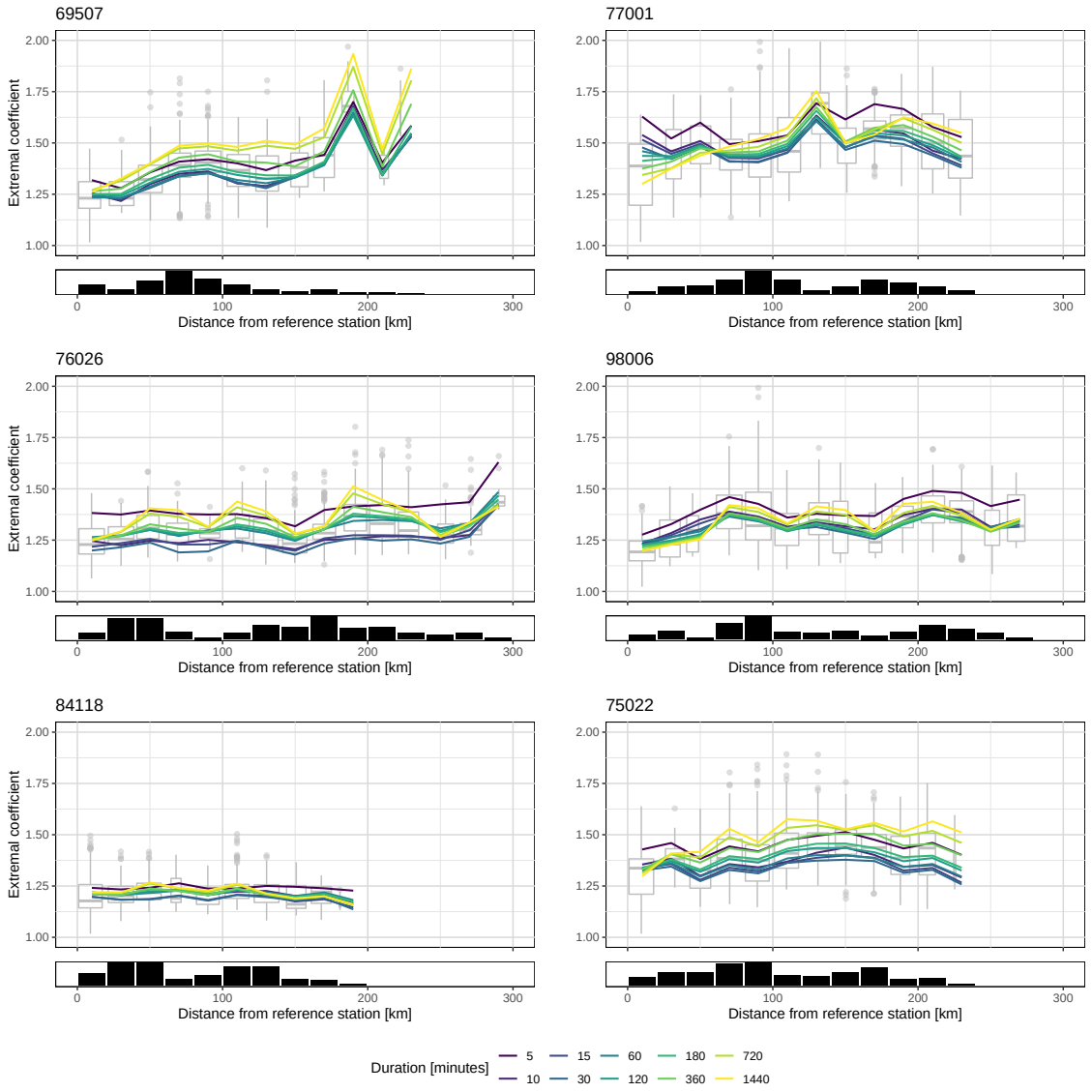


Figure E.1: Empirical extremal coefficient for the six stations of interest with respect to all the other stations, as function of the distance from the reference station (mentioned at the top of each panel). Colored lines correspond to the binned empirical estimates with bin size 20km, and boxplots are computed based on each empirical estimates. The panels with dark bars represent the frequency of stations within the distance range. The extremal coefficient slightly tends to move toward asymptotic independence as the distance increases, but this is not visible for station 84118. For very short distances, long durations tend to exhibit higher dependence than for shorter ones, but this phenomenon is reversed as the distance increases.

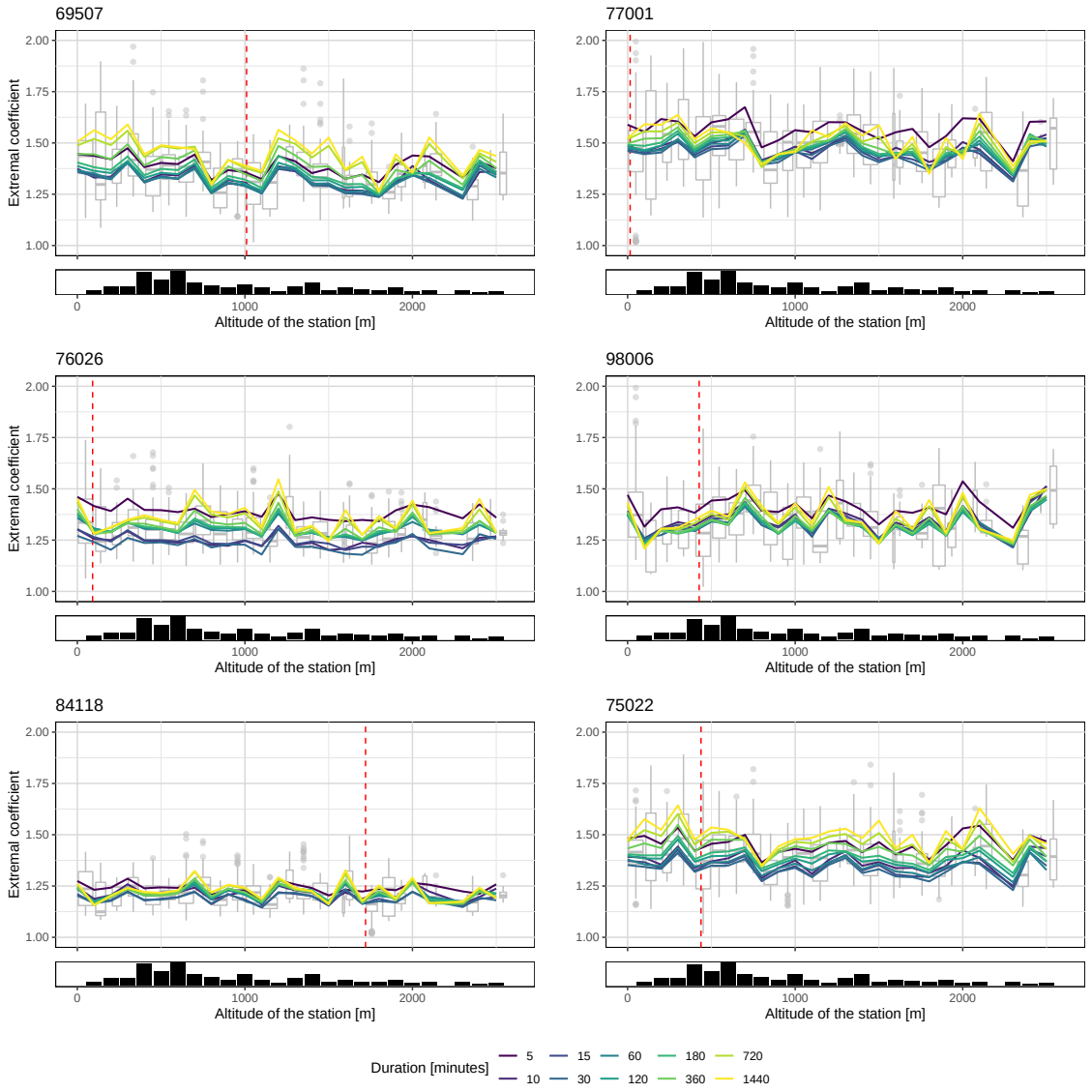


Figure E.2: Empirical extremal coefficient for the six stations of interest with respect to all the other stations, as function of the altitude. The altitude of the station of reference is represented by a vertical red line. Colored lines correspond to the binned empirical estimates with bin size 100m, and boxplots are computed based on each empirical estimates. The panels with dark bars represent the frequency of stations within the altitude range (which coincides for all panels with Figure A.1, Appendix A). The altitude does not seem to fully explain the behaviour of extremal dependence.

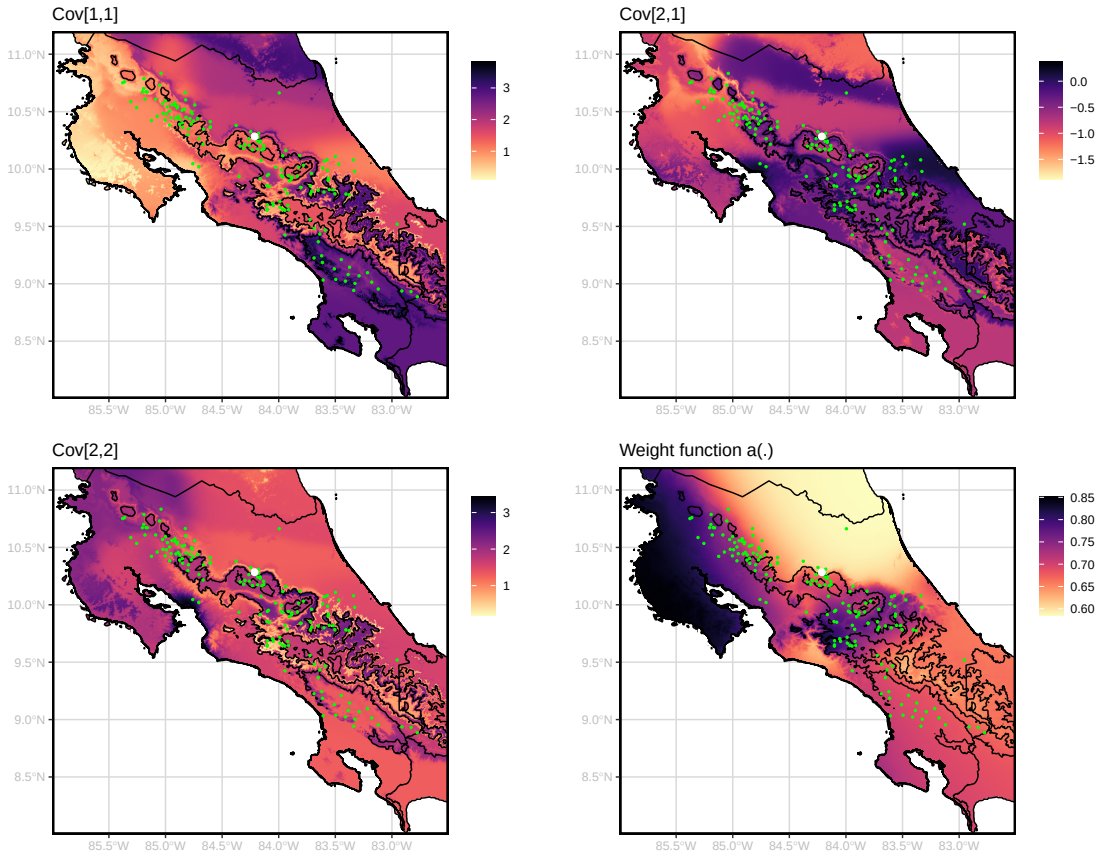


Figure E.3: Elements of the covariance matrix from the non-stationary max-stable process, depending on the location (for duration 1440 minutes). The covariance does not vary much from one duration to another. Green points show the other locations. The top left panel show the element $\Omega_{1,1}$, $\Omega_{1,2}$ for the top right and $\Omega_{2,2}$ for the bottom left. The bottom right panel show the function $a(\cdot)$ for all the locations. High values correspond to low smoothness and includes the provinces of Guanacaste and San José.

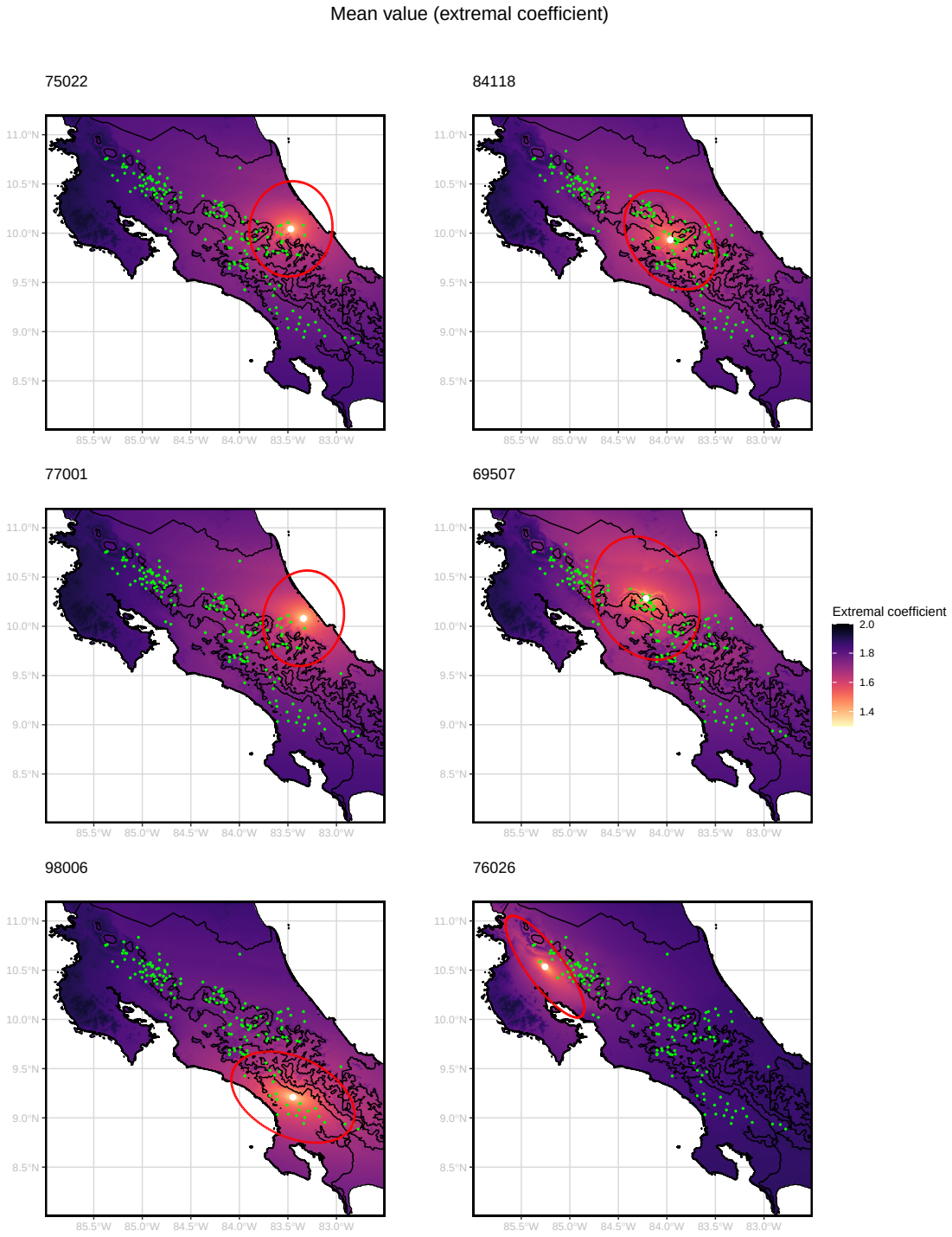


Figure E.4: Mean estimated extremal coefficient from the non-stationary max-stable process (with duration 1440 minutes), with the six stations taken as reference precipitation gauges. Green points show the other locations, and the white point the station of reference. Red circles represent the contour from the covariance matrix, centered at the precipitation gauge of reference.

E.1 Bivariate density function of the extremal t model

In the following, we give the derivation of the bivariate density for the extremal t process given in Section 5.2.2, which is necessary for model fitting and maximization of the log likelihood. As a reminder, the bivariate cumulative distribution for the extremal t process is

$$\mathbb{P}\{Z^*(x_1) \leq z_1, Z^*(x_2) \leq z_2\} = \exp\left\{-\frac{1}{z_1}T_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right] - \frac{1}{z_2}T_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right]\right\}, \quad \text{(E.1.1)}$$

where $T_\nu(\cdot)$ is the Student t cumulative distribution function with ν degrees of freedom, and

$$r_{x_1x_2}(y) = \frac{y^{1/\nu} - \rho(x_1, x_2)}{(\nu+1)^{-1/2}[1-\rho(x_1,x_2)^2]^{1/2}}. \quad \text{(E.1.2)}$$

In the following, we use the notation $K = (\nu+1)^{-1/2}[1-\rho(x_1,x_2)^2]^{1/2}$ and let V be the exponent measure

$$V(z_1, z_2) = \frac{1}{z_1}T_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right] + \frac{1}{z_2}T_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right]. \quad \text{(E.1.3)}$$

The bivariate density f is given by

$$f(z_1, z_2) = \left[\frac{\partial V(z_1,z_2)}{\partial z_1}\frac{\partial V(z_1,z_2)}{\partial z_2} - \frac{\partial^2 V(z_1,z_2)}{\partial z_1 \partial z_2}\right]\exp\{-V(z_1,z_2)\}. \quad \text{(E.1.4)}$$

Then,

$$\begin{aligned}\frac{\partial V(z_1,z_2)}{\partial z_1} =& -\frac{1}{{z_1}^2}T_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right] + \frac{1}{z_1}t_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right]\frac{\partial r(z_2/z_1)}{\partial z_1} \\ &+ \frac{1}{z_2}t_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right]\frac{\partial r(z_1/z_2)}{\partial z_1},\end{aligned} \quad \text{(E.1.5)}$$

$$\begin{aligned}\frac{\partial V(z_1,z_2)}{\partial z_2} =& -\frac{1}{{z_2}^2}T_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right] + \frac{1}{z_2}t_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right]\frac{\partial r(z_1/z_2)}{\partial z_2} \\ &+ \frac{1}{z_1}t_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right]\frac{\partial r(z_2/z_1)}{\partial z_2},\end{aligned} \quad \text{(E.1.6)}$$

with $t_{\nu+1}$ the Student t density function with $\nu+1$ degrees of freedom. The derivative of r with respect to y is

$$\frac{\partial r(y)}{\partial y} = y^{-1+1/\nu}(\nu K)^{-1}, \quad \text{(E.1.7)}$$

so that the missing partial derivatives of Equations (E.1.5) and (E.1.6) are

$$\begin{aligned}\frac{\partial r(z_2/z_1)}{\partial z_1} &= -\frac{z_2}{K\nu {z_1}^2}\left(\frac{z_2}{z_1}\right)^{-1+1/\nu}, \\ \frac{\partial r(z_1/z_2)}{\partial z_2} &= -\frac{z_1}{K\nu {z_2}^2}\left(\frac{z_1}{z_2}\right)^{-1+1/\nu}, \\ \frac{\partial r(z_1/z_2)}{\partial z_1} &= \frac{1}{K\nu z_2}\left(\frac{z_1}{z_2}\right)^{-1+1/\nu}, \\ \frac{\partial r(z_2/z_1)}{\partial z_2} &= \frac{1}{K\nu z_1}\left(\frac{z_2}{z_1}\right)^{-1+1/\nu}.\end{aligned} \quad \text{(E.1.8)}$$

In order to compute the bivariate density, one needs the second partial derivatives of V with respect to z_1 and z_2

$$\begin{aligned}\frac{\partial^2 V(z_1,z_2)}{\partial z_1 \partial z_2} =& -\frac{1}{{z_1}^2} t_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right]\frac{\partial r(z_2/z_1)}{\partial z_2} \\ & -\frac{1}{{z_2}^2} t_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right]\frac{\partial r(z_1/z_2)}{\partial z_1} \\ & +\frac{1}{z_1}\left\{\frac{\partial t_{\nu+1}[r(z_2/z_1)]}{\partial r(z_2/z_1)}\frac{\partial r(z_2/z_1)}{\partial z_1}\frac{\partial r(z_2/z_1)}{\partial z_2} + t_{\nu+1}\left[r\left(\frac{z_2}{z_1}\right)\right]\frac{\partial^2 r(z_2/z_1)}{\partial z_1 \partial z_2}\right\} \\ & +\frac{1}{z_2}\left\{\frac{\partial t_{\nu+1}[r\,(z_1/z_2)]}{\partial r(z_1/z_2)}\frac{\partial r(z_1/z_2)}{\partial z_2}\frac{\partial r(z_1/z_2)}{\partial z_1} + t_{\nu+1}\left[r\left(\frac{z_1}{z_2}\right)\right]\frac{\partial^2 r(z_1/z_2)}{\partial z_1 \partial z_2}\right\},\end{aligned} \tag{E.1.9}$$

with

$$\begin{aligned}\frac{\partial^2 r(z_1/z_2)}{\partial z_1 \partial z_2} &= -\frac{1}{K\nu^2 {z_2}^2}\left(\frac{z_1}{z_2}\right)^{-1+1/\nu}, \\ \frac{\partial^2 r(z_2/z_1)}{\partial z_1 \partial z_2} &= -\frac{1}{K\nu^2 {z_1}^2}\left(\frac{z_2}{z_1}\right)^{-1+1/\nu}.\end{aligned} \tag{E.1.10}$$

The final elements that are needed for the bivariate density are the Student t density and its derivative,

$$\begin{aligned}t_{\nu+1}(x) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{2\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \\ \frac{t_{\nu+1}(x)}{\partial x} &= \frac{x\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{2\pi}\Gamma(\nu/2)}\left(-\frac{\nu+1}{\nu}\right)\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+3}{2}}.\end{aligned} \tag{E.1.11}$$

Appendix F

Neural network architectures

The following section shows the architectures of the neural networks used through this report. Figure F.1 corresponds to the architecture used for the univariate GEV model. It consists of four hidden layers of sizes 100, 20, 20 and 10, the input (that consists of the longitude, latitude, altitude, duration and encodings of the month) and output being of dimension 8 and 3 respectively. The deep neural network is represented as a green box, and each layer as a white rectangle. Within each layer, the computations are described in Section 3.3.2:

$$a^k = l\left(W^k a^{k-1} + b^k\right) \in \mathbb{R}^{N_k}, \quad k = 1, \ldots, K, \tag{F.0.1}$$

with $l(\cdot)$ the sigmoid activation function applied component-wise. The parameters W^k, b^k, $k = 1, \ldots, K$ are optimized by minimizing the negative log-likelihood of the GEV distribution given in 3.1.5, using the Adam stochastic optimizer [Kingma and Ba, 2015]. The output of the feedforward ANN is

$$\begin{aligned} \mu &= a_1^K \in \mathbb{R}, \\ \sigma &= \exp(a_2^K) \in \mathbb{R}_+^*, \\ \xi &= 2h(a_3^K)/3 - 0.5 \in [-0.5, 1.0], \end{aligned} \tag{F.0.2}$$

with $h(x) = \{1 + \exp(-x)\}^{-1}$ the sigmoid function.

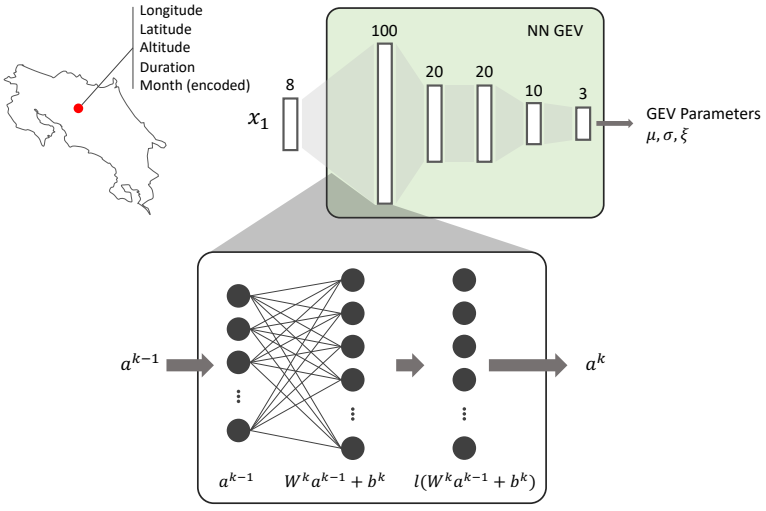


Figure F.1: Neural network architecture of GEV model. The lower box shows the computations in each layer.

Figures F.2 and F.3 show the architecture of the bivariate and max-stable models respectively. As for the GEV model, green and orange boxes correspond to deep neural networks that contains different layers. For the bivariate model, only the characteristics of the second station are used as input, as the station of reference is assumed to be fixed during the learning phase. This implies that one model needs to be trained for each precipitation gauge. The output of the model is the parameter r of the Hüsler–Reiss model.

The non-stationary max-stable process estimates the components of the covariance matrix Ω for each station x (based on their characteristics, such as the longitude and latitude), along the weight $a(x)$ that is also estimated using a neural network (shown as an orange box in Figure F.3). During the learning phase, pairs of stations (x_1, x_2) are sampled randomly from the dataset, and used to predict the correlation $\rho(x_1, x_2)$ that is then fed into the loss function, that is the negative log likelihood for the extremal t process (the bivariate density is presented in Section E.1 Appendix E).

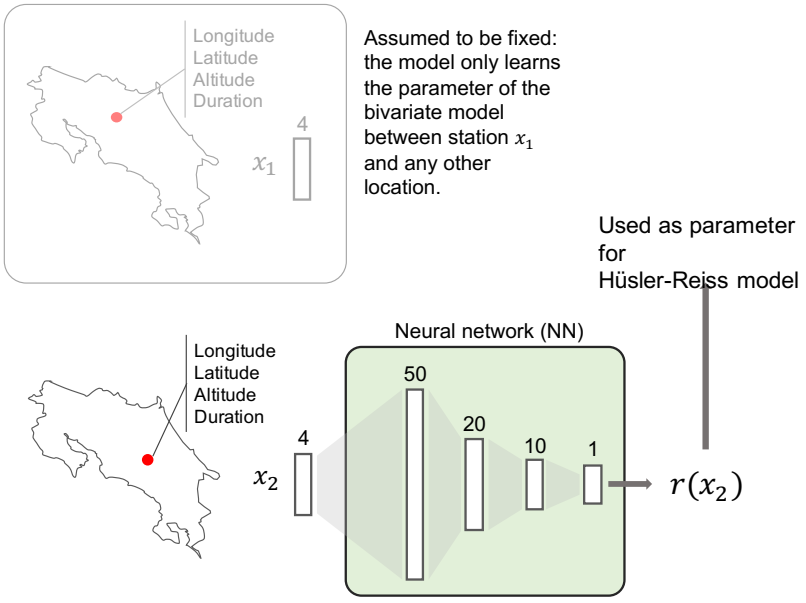


Figure F.2: Neural network architecture of the bivariate model.

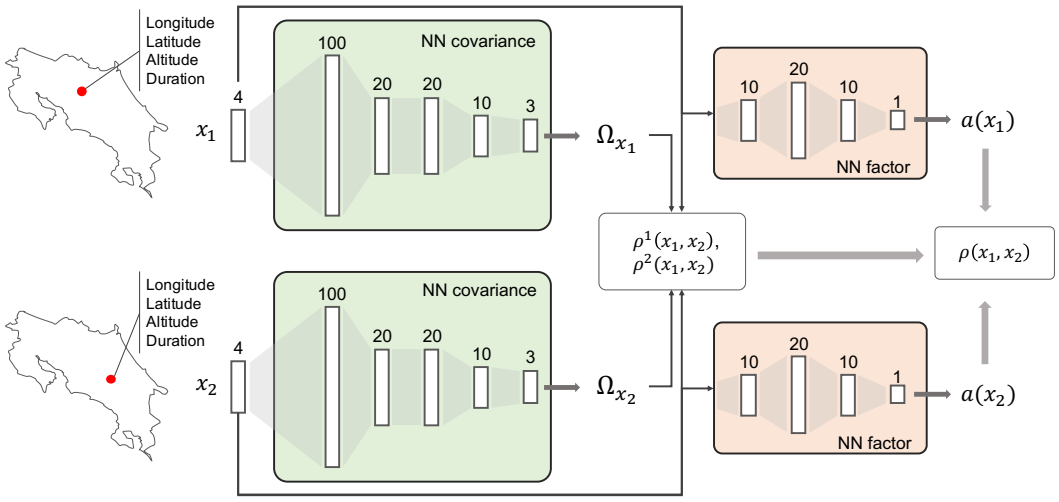


Figure F.3: Neural network architecture of the non-stationary max-stable approach.

List of Figures