UNIVERSITE CATHOLIQUE DE LOUVAIN

FACULTE DES SCIENCES

ECOLE DE STATISTIQUE, BIOSTATISTIQUE
ET SCIENCES ACTUARIELLES

Temporal analysis of the evolution of extreme values using climatological data

Promoteur : Johan Segers

[] Lecteurs : Anna Kiriliouk

Michel Crucifix

Mémoire présenté en vue de l'obtention du

Master en statistiques, orientation générale

par : **Antoine Pissoort**

Juin 2017

Abstract

This thesis aims to analyse extreme temperatures from Uccle and assess their nonstationarity. trend in the location parameter of the temperatures assessing the climate warming. After having proven it by hand of splines's deriatives that a correction for simultaneous intervals led to non-significant changes , we will do it by hand of the Extreme Value Theory (EVT) that there is indeed an upward trend in the location parameter of the temperatures assessing the climate warming. First of all, we will do introductory analysis of the trend... and we will discover that.. The analysis will focus on yearly maxima and hence we will go through with EVT by defining and presenting the usual methods such as GEV.

Regarding the computations, we took advantage of a high-level language (c++) to make our analysis efficient and also made use of parallel computing to decrease computation time for time consuming...

Keywords • Extreme Value Theory • block-maxima model • peaks-over-threshold method • trend analysis • Bayesian inference • Generalized Additive Models with splines smoothing • Neural Networks • nonstationary models • Markov Chain Monte Carlo • Hamiltonian Monte Carlo • Parallel computing • R package • Shiny application

Acknowledgements

I would first like to thank my thesis supervisor Johan Segers for all his help and his guidance during this whole year. The repeated appointments we have had

I also would like to thank the "Institut Royal de Météorologie" (IRM) of Belgium for his help and his guidance but also for his provided quality datasets.

A bit less usual, I would like to thank the open source community such as R or LaTeX can benefit. For me and for a number of student, it has been a nonnegligable source of ideas and the most efficient source of learning.
Finally, I want to thank my family and my friends, but also Bernadette for her support and all the time I have spent writing in her room for my thesis but also during my whole academic studies.

Contents

List of Figures

List of Tables

List of Abbreviations

For convenience, we place a list of all the abbreviations we will use in the text. However, these will always be defined in their first occurrence in the text.

DA . . . . . Domain of Attraction

df . . . . . (cumulative) distribution function

EVI . . . . . Extreme Value Index (ξ)

EVT . . . . . Extreme Value Theory

GEV . . . . . Generalized Extreme Value

GML . . . . . Generalized Maximum Likelihood

GPD . . . . . Generalized Pareto Distribution

MCMC . . . . . Marko Chain Monte Carlo

MH . . . . . Metropolis-Hastings (algorithm)

ML . . . . . Maximum Likelihood

MLE . . . . . Maximum Likelihood Estimator

NN . . . . . Neural Network

TN . . . . . Temperature miNimum

TX . . . . . Temperature maXimum

Introduction

Unlike its counterparts (see for example credit risk analysis, financial applications,...), the Extreme Value Theory (EVT) applied on broad environemental area such as meteorological data, has strong impacts on the people lives

An important question is still whether climate changes caused by anthropogenic activities will change the intensity and frequency of extreme events Milly et al. [2008].

The problem facing climate change evidence is the lack of past data to compare with her

For such an analysis, the number of parameters to take into account is considerable (and tend to infinity)

To make a parallel with Chaos Theory and the well-known butterfly effect which have strong applications in weather models

We anticipate that climate change to affect the extreme weather

[extremes in climate change p.347]

It has been proven that winter is becoming warmer in context of RC. (see naveau,...)

?

"The first myth about climate extremes, which has been purported by researchers in climatology or hydrology, among them prominent names, is that “extremes are defined as rare events” or similar. This myth is debunked by a simple bimodal PDF (Fig. 6.12a). The events sitting in the tails of that distribution are not rare" [Mudelsee, 2014, pp.257]

Until now, studies on climate extremes in Europe have usually had a strong national signature , or have had to make use of either a dataset with daily series from a very sparse network of meteorological stations (e.g. eight stations in Moberg et al. (2000)) or standardized data analysis performed by different researchers in different countries along the lines of agreed methodologies (e.g. Brazdil et al., 1996; Heino et al., 1999) Klein Tank and Wijngaard [2002]

Extrapolation !!!! See p154 [statistical analysis of extreme book]

Voir effet de l’ilot de chaleur –> urbanisation sur les tempés !

—> artificial warming on cities stations which were not(less) urbanized 100 years ago.

[In this thesis, efforts have been made to use power of (hyper)references into the text. While this not (yet ?...) usable in printed versions, the reader may feel more comfortable in a numeric version to more easily handle the vast amount of sections, equations, references, etc... and the links that are made

between them.]

We can summarize the research question of this thesis as the following :

-

In this part, we will make use of general methods to assess if there is indeed a trend in the maximum temperatures

There are two main approaches in EVT, the block-maxima and the peaks-over-threshold approach (see Section **2**) yielding to different extreme value distribution. The former aims at while the latter models the (...)

In Chapter **1** we will present the method of block-maxima and derive the Generalized Extreme Value (GEV) distributions. In Chapter **2** we will . In Chapter **3** In Chapter **4** we will ..

Finally, we notice that this thesis will concentrate on the block-maxima (GEV) methods (see Chapter **1**), i.e. to data relating to maxima over a period of 1 year. For example here in the introduction, or for the Bayesian analysis in Chapter **5**.

Part I

Theoretical Framework : Extreme Value Theory

Chapter 1

Method of Block Maxima

Contents

This chapter introduce the basics of EVT by considering the *block-maxima* approach. After defining useful concepts in Section **1.1** to introduce the emergence of this theory, we will get into the leading theorem of EVT in Section **1.2**. Section **1.3** will present some mathematical applications and Section **1.4** conditions of this theorem in order to visualize the implications of the extremal theorem and the characterizations of the underlying distributions. Section **1.5** will introduce key concepts of inference in EVT which will be presented in a general (and frequentist) way in Section **1.6**. Finally, Section **1.7** provides some tools to assess the accuracy of the fitted model.

This chapter is mostly based on Coles [2001, chap.3], Beirlant et al. [2006, chap.2] and Reiss and Thomas [2007, chap.1-4], and other relevant articles.

1.1 Preliminaries

In the following, a sequence of independent and identically distributed (iid) random variables are assumed and written in the form $\{X_n\}_{n\in\mathbb{N}}$. The X_i's share a common cumulative distribution function (df) F. The iid assumption will be relaxed in Chapter **3**.

Statistical Tools

Let $X_{(i)}$ denote the i-th ascending order statistic,

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}, \tag{1.1}$$

assuming n observations. One order statistic is of particular interest, the *maximum* $X_{(n)}$

$$X_{(n)} := \max_{1\leq i\leq n} X_i, \tag{1.2}$$

while the *minimum* $X_{(1)}$ can be defined it with respect to the maximum operator

$$X_{(1)} := \min_{1\leq i\leq n} X_i = -\max_{1\leq i\leq n}(-X_i). \tag{1.3}$$

This text will **focus on maxima** but it is important to keep in mind that the analysis made in the following can be extended to minima through relation (1.3).

Furthermore, we can retrieve the distribution of $X_{(n)}$. By definition,

$$\begin{aligned} \Pr\{X_{(n)} \leq x\} &= \Pr\{X_1 \leq x, \ldots, X_n \leq x\} \\ &\overset{(\perp\!\!\!\perp)}{=} \Pr\{X_1 \leq x\} \ldots \Pr\{X_n \leq x\} \\ &= F^n(x), \end{aligned} \tag{1.4}$$

where the independence $(\perp\!\!\!\perp)$ follows directly from the iid assumption of the sequence $\{X_i\}$.

First Definitions and Theorems : Motivations

Definition 1.1 (Distributions of same type)**.** *We say that two dfs G and G^* are of the* **same type** *if, for constants $a > 0$ and b we have*

$$G^*(az + b) = G(z), \qquad \forall z. \tag{1.5}$$

△

This means that the distributions only differ in location and scale. This concept will be useful later in the text to derive the three different families of extreme value distributions which come from other distributions that are of the *same type*.

Principles of stability : Amongst the principles about EVT that will be covered during this text, EVT will be highly influenced by the principles of *stability*. It states that a model should remain valid and consistent whatever choices are made on the structure of this model. For example, if we propose a

model for the annual maximum temperatures and another for the 5-year maximum temperatures, the two models should be mutually consistent since the 5-year maximum will be the maximum of 5 annual maxima. Similarly, in a Peaks-Over-Threshold setting (presented in Chapter 2), a model for exceedances over a high threshold should remain valid for exceedances of higher thresholds.

Definition 1.2 (Max-stability)**.** From Leadbetter et al. [1983] or Resnick [1987], *we say that a distribution* G *is* **max-stable** *if, for each* $n \in \mathbb{N}$ *we have*

$$G^n(a_n z + b_n) = G(z), \qquad n = 1, 2, \ldots, \tag{1.6}$$

for appropriate normalizing constants $a_n > 0$ *and* b_n*.* △

In other words, taking powers of G results only in a change of location and scale. This concept will be closely connected with the fundamental limit law for extreme values that we will present in the next Section. However, the power of max-stable processes is often used in a multivariate setting, whereas we will focus on univariate sequences. Refer for example to Ribatet et al. [2015] for an introduction on max-stable processes.

A fundamental concept of EVT is the concept of **degenerate** dfs. We recall that the df of a random variable is said to be *degenerate* if it assigns all probability to a single point. We illustrate this by the construction of the well-known *Central Limit Theorem* (CLT) that we will state below and which concerns the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. We know from the Weak Law of Large Numbers that $\bar{X}_n$ will converge *almost surely* to the true mean μ (see[1] Theorem **A.1**). and thus in distribution, that is to a non-random single point, i.e. to a *degenerate* distribution

$$\Pr\{\bar{X}_n \leq x\} = \begin{cases} 0, & x < \mu; \\ 1, & x \geq \mu. \end{cases}$$

This is not useful, in particular for inferential purposes.

For this reason, CLT aims at finding a non-degenerate limiting distribution for $\bar{X}_n$, after allowing for normalization by sequences of constants. We will state it in its most basic form :

Theorem 0 (Central Limit Theorem)**.** *Let* $\{X_i\}$ *be a sequence of* n *iid random variables with* $E(X_i^2) < \infty$*. Then, as* $n \to \infty$*,*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where $\mu = E(X_i)$ *and* $\sigma^2 = V(X) > 0$*.*

Then, by making a proper choice of some normalizing constants, μ and $\sqrt{n}$ (as location and scale parameters respectively), we find the non-degenerate normal distribution in the limit for the empirical mean $\bar{X}_n$ provided X_i has a nonzero variance and finite second moment.

With the same logic, we find for the distribution of maximum order statistics $X_{(n)}$

$$\lim_{n\to\infty} \Pr\{X_{(n)} \leq x\} = \lim_{n\to\infty} \Pr\{X_i \leq x\}^n = \begin{cases} 0, & F(x) < 1; \\ 1, & F(x) = 1, \end{cases} \tag{1.7}$$

[1] Appendix **A.2** can be useful for a relevant short review of main concepts of convergence.

which is also a degenerate distribution.

Whereas the CLT dealt with the sample mean, EVT also aims to find a non-degenerate distribution in the limit of the maximum $X_{(n)}$ by means of normalization.

1.2 Extremal Types Theorem : Extreme Value distributions

Introduced by Fisher and Tippett [1928], later revised by Gnedenko [1943] and streamlined by de Haan [1970], the *extremal types theorem* is important for its applications in EVT. Let $\{X_i\}$ be a sequence of iid random variables with df F. It states the following :

Theorem 1.1 (Extremal Types). *If there exist sequences of normalizing constants $a_n > 0$, $b_n \in \mathbb{R}$ and a non-degenerate limiting distribution G such that*

$$\lim_{n\to\infty} \Pr\Big\{a_n^{-1}(X_{(n)} - b_n) \leq z\Big\} = F^n(a_n z + b_n) = G(z), \qquad \forall z \in \mathbb{R}, \tag{1.8}$$

then G has the same type as one of the following distributions :

$$\boxed{\textit{I}} : \qquad G_1(z) = \exp\Big\{-e^{-z}\Big\}, \qquad -\infty < z < \infty. \tag{1.9}$$

$$\boxed{\textit{II}} : \quad G_{2,\alpha}(z) = \begin{cases} 0, & z \leq 0; \\ \exp\Big\{-z^{-\alpha}\Big\}, & z > 0. \end{cases} \tag{1.10}$$

$$\boxed{\textit{III}} : \quad G_{3,\alpha}(z) = \begin{cases} \exp\Big\{-(-z)^{\alpha}\Big\}, & z < 0; \\ 1, & z \geq 0, \end{cases} \tag{1.11}$$

for some parameter $\alpha > 0$ in case II and III. □

These are termed the *standard extreme value distribution functions* and are differentiated by three types. Note that each real parameter α determines the type. **Type** $\boxed{\textbf{I}}$ is commonly known as the *Gumbel* family while the **type** $\boxed{\textbf{II}}$ and **type** $\boxed{\textbf{III}}$ are known as the *Fréchet* and the *Weibull* families respectively. From the fact that G is of the *same type* as one of the three distribution, we can rescale these distributions by some normalizing parameters $a > 0$ (scale) and b (location), that is $G_{i,\alpha,a,b}(z) = G_{i,\alpha}\Big(\frac{z-b}{a}\Big)$ for $i = 2, 3$, and similarly for G_1 to obtain the *full* EV distributions.

This theorem considers an iid random sample, but it holds true even if the original scheme is no longer independent. We will present the stationary case in Section **3.1**. Furthermore, we will see in Section **1.4** that F is in the *domain of attraction* of G.

1.2.1 Generalized Extreme Value Distribution

Von Mises [1936] showed that another representation is possible by taking the reparametrization $\xi = \alpha^{-1}$ of the extreme values dfs to obtain a continuous, unified model.

Hence, these three classes of extreme distributions can be expressed in the same functional form as

special cases of the single three-parameter *Generalized Extreme Value* (GEV) distribution

$$G(z) := G_{\xi,\mu,\sigma}(z) = \exp\left\{ -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}}\right\}, \qquad \xi \neq 0. \tag{1.12}$$

where $-\infty < \mu, \xi < \infty$ and $\sigma > 0$ with (μ, σ, ξ) being the three parameters of the model characterizing location, scale and shape respectively. We introduce the notation $y_+ = \max(y, 0)$ to denote that (1.12) is defined on $\{z : 1 + \xi\sigma^{-1}(z-\mu) > 0\}$. It ensures the term in the exponential function is negative, and the df converges to 1. It is important to note that this yields a vital condition for the GEV as it defines the endpoints from the three different characterizations of this distribution from the values of the shape parameter.

The GEV corresponds to the *Fréchet* family (1.10) whenever $\xi > 0$ and to the *Weibull* family (1.11) as $\xi < 0$. When $\xi = 0$, i.e. for the *Gumbel* family (1.9), the situation in (1.12) is not defined but is taken as the limit as $\xi \to 0$, leading to

$$G(z) := G_{\mu,\sigma}(z) = \exp\left\{ -\exp\left(\frac{z-\mu}{\sigma}\right)\right\}, \qquad \xi = 0. \tag{1.13}$$

The shape parameter $\xi \in \mathbb{R}$ is called the *extreme value index* (EVI) and is at the center of the analysis in EVT. It determines, in some degree of accuracy, the type of the underlying distribution.

Following Coles [2001], we introduce an important theorem in Extreme Value Theory and that has many implications. This theorem states the following :

Theorem 1.2. *For any df F,*

$$F \text{ is } \textit{max-stable} \iff F \text{ is GEV.} \tag{1.14}$$

□

Hence, any df that is *max-stables* (see Definition **1.2**) is also GEV (1.12)-(1.13), and vice-versa. To gain interesting insights of the implications of this theorem, we think it is useful to give an informal proof but only for the "⇐" as the converse requires a significant mathematical background. By using max-stability (Definition **1.2**) and Theorem **1.2**, this "proof" also gives intuition for Theorem **1.1**.

Outline Proof of **Extremal Types** Theorem **1.1** (and Theorem **1.2**) :

- If $a_n^{-1}(X_{(n)} - b_n)$ has the GEV as limit distribution for large n as defined in (1.8), then

$$\Pr\left\{a_n^{-1}(X_{(n)} - b_n) \leq z\right\} \approx G(z).$$

  Hence for any integer k, since nk is large, we have

$$\Pr\left\{a_{nk}^{-1}(X_{(n)k} - b_{nk}) \leq z\right\} \approx G(z). \tag{1.15}$$

- Since $X_{(n)k}$ is the maximum of k variables having identical distribution as $X_{(n)}$,

$$\Pr\left\{a_{nk}^{-1}(X_{(n)k} - b_{nk}) \leq z\right\} = \left[\Pr\{a_{nk}^{-1}(X_{(n)} - b_{nk}) \leq z\}\right]^k, \tag{1.16}$$

giving two expressions for the distribution of $X_{(n)}$, by (1.15) and (1.16) :

$$\Pr\{X_{(n)} \leq z\} \approx G\Big(a_n^{-1}(z-b_n)\Big) \quad \text{and} \quad \Pr\{X_{(n)} \leq z\} \approx G^{1/k}\Big(a_{nk}^{-1}(z-b_{nk})\Big).$$

- It follows that G and $G^{1/k}$ are identical apart from location and scale coefficients. Hence, G is *max-stable* and therefore GEV. This gives intuition of the **extremal types** Theorem **1.1**.

□

In words, it means that taking power of G results only in a change of location and scale, and hence by recalling the expression of the distribution of $X_{(n)}$(1.4), it is possible to find the non-degenerate GEV in the limit for $X_{(n)}$. More technical details can be found in Leadbetter et al. [1983].

Density

The density of the GEV distribution (1.12), $g(z) = \frac{dG(z)}{dz}$ (since we have absolute continuity) can be expressed in two forms, as depicted in Table 1.1.

Table 1.1: *Two cases for the* density distribution *of the GEV*

$\xi \neq 0$	$g(z) = \sigma^{-1}\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}-1} \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}}\right\};$
$\xi \to 0$	$g(z) = \sigma^{-1}\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}.$

We can now try to visually represent these three families. The following Figure 1.1 depicts the GEV, defined with respect to the value of the shape parameter ξ.

It is important to point out that the location parameter μ does not represent the mean as in the classic statistical view rather it represents the "center" of the distribution; and the scale parameter σ is not the standard deviation but does govern the "size" of the deviations around μ. This can be visualized in Figure C.1 in Appendix **C** where we show the variation of the GEV distribution when we vary these parameters[2]. We notice that the location parameter only implies a horizontal shift of the distribution, without changing its shape, and w see the influence of the scale parameter on the spread of the distribution around μ. For example if $\sigma \nearrow$, then the density will appear more flat.

In the following, we define the *left* and the *right endpoint* of a particular df F as respectively ${}_*x$ and x_*, by :

$${}_*x = \inf\{x \in \mathbb{R} : F(x) > 0\}, \qquad \text{and} \qquad x_* = \sup\{x \in \mathbb{R} : F(x) < 1\}. \tag{1.17}$$

Note that the Gumbel distribution is unbounded. The Fréchet distribution has a finite left endpoint in ${}_*x = \mu - \sigma \cdot \xi^{-1}$ (blue circle in Figure 1.1), and its upper endpoint is $+\infty$ while the Weibull distribution has a finite right endpoint in $x_* = \mu - \sigma \cdot \xi^{-1}$ (red circle in Figure 1.1) and is unbounded in the left. This has serious impact on modeling. Since these endpoints are functions of the parameter values, will

[2]Moreover, a Shiny application has been built through the R package to visualize in the best way the influence of the parameters on this distribution. See intro of Chapter **5** for an explanation on its use.

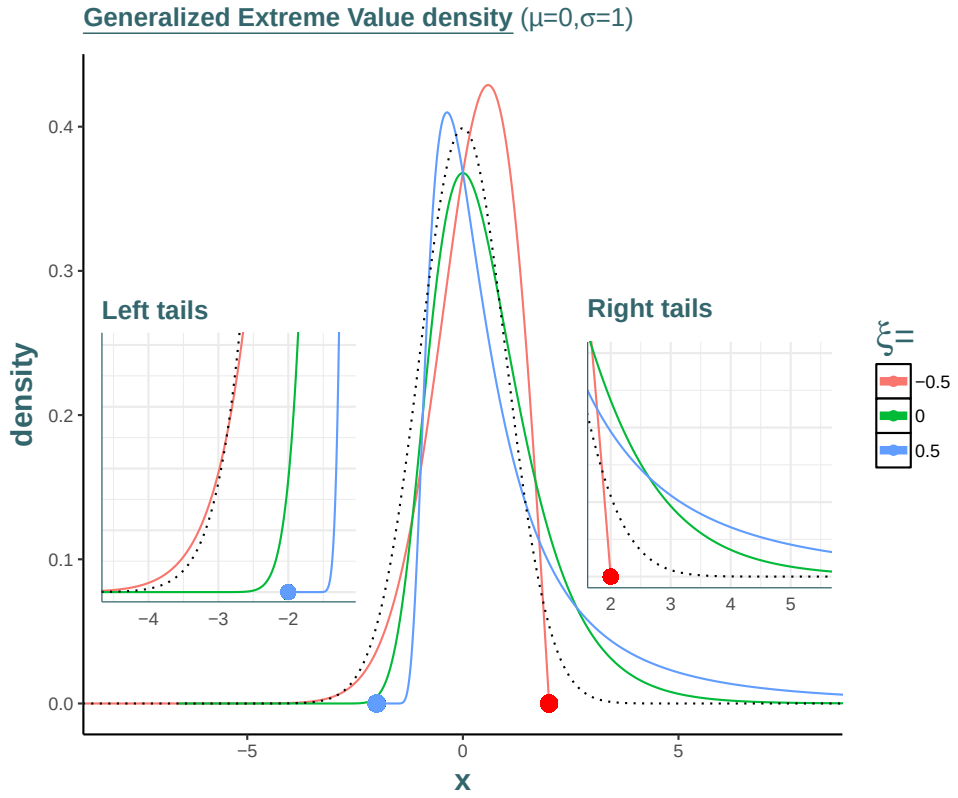

Figure 1.1: *GEV distribution with the normal as benchmark (dotted lines) and a zoom on the parts of interest to better visualize the behaviour in the tails. In green, we retrieve the Gumbel distribution ($\xi = 0$). In red, we retrieve the Weibull-type ($\xi < 0$) while in blue, we get the Fréchet-type ($\xi > 0$). The endpoints for the Weibull and the Fréchet are denoted by the red and blue filled circles respectively.*

see later in Section **1.6.1** that it can make the likelihood computation unstable. We will particularly face this in the Bayesian Chapter **4**.

It can be useful to think not only about the specific form of data or the distribution they will fit and its characteristics, but also about how to retrieve these specific distributions in practice. That is why we detail some examples of how to construct such EV distributions for the three types in concrete cases, playing with the appropriate choice of sequences a_n and b_n to retrieve the pertaining distribution family.

1.3 Applications : Examples of Convergence to GEV

In real applications it is not easy to find the appropriate sequences, but it is useful to understand the concept of convergence to GEV by looking at some theoretical examples.

Convergence to Gumbel distribution

The **Type** **I** or **Gumbel** distribution G_1 can be retrieved by considering, for example, an iid exponentially distributed sequence $\{X_j\}$ of n random variables, that is $X_j \overset{iid}{\sim} Exp(\lambda)$ and taking the largest

of these values, $X_{(n)}$, as defined earlier. By definition, if the X_j have df F, then $F(x) = 1 - e^{-\lambda x}$ for $x > 0$. Hence, our goal is to find non-random sequences $\{b_n\}$, $\{a_n > 0\}$ such that

$$\lim_{n\to\infty} \Pr\Big\{a_n^{-1}(X_{(n)} - b_n) \leq z\Big\} = G_1(z). \tag{1.18}$$

Hence, we can easily find that

$$\begin{aligned}\Pr\Big\{a_n^{-1}(X_{(n)} - b_n) \leq z\Big\} &= \Pr\{X_{(n)} \leq b_n + a_n z\} \\ &= \Big[\Pr\{X_1 \leq b_n + a_n z\}\Big]^n \\ &= \Big[1 - \exp\{-\lambda(b_n + a_n z)\}\Big]^n,\end{aligned}$$

from the iid assumption of the random variables and their exponential distribution. Hence, by choosing the sequences $a_n = \lambda^{-1}$ and $b_n = \lambda^{-1}\log n$ and reminding that

$$\Big[1 - \exp\{-\lambda(b_n + a_n z)\}\Big]^n = \Big[1 - \frac{1}{n}e^{-z}\Big]^n \overset{n\to\infty}{\longrightarrow} \exp(-e^{-z}) := G_1(z),$$

we find the so-called standard *Gumbel* distribution in the limit. Note that the same can be retrieved with $X_j \overset{iid}{\sim} N(0,1)$ and with sequences $a_n = -\Phi^{-1}(1/n)$ and $b_n = 1/a_n$.

Typically unbounded distributions, for example the Exponential and Normal, whose tails fall off exponentially or faster, will have the Gumbel limiting distribution for the maxima. They will have, in particular, medians (and other quantiles) that grow as $n \to \infty$ at the rate of 'some power of' $\log n$. This is a typical example of light-tailed distribution (i.e., whose tails decay exponentially, as defined in Appendix **A.1**).

Convergence to Fréchet distribution

The **Type** **II** or **Fréchet type** (or *Fréchet-Pareto*) distribution $G_2(x)$ has strong relations with the Pareto distribution and also the Generalized Pareto Distribution that will be presented in Chapter **2**. These are distributions which are typically heavy- or fat-tailed (see Appendix **A.1**).

Following Beirlant et al. [1996], when starting with a sequence $\{X_j\}$ of n iid random variables following a *basic* (or *generalized* with scale parameter set to 1) Pareto distribution with shape parameter $\alpha \in (0, \infty)$, $X_j \sim Pa(\alpha)$, we have that

$$F(x) = 1 - x^{-\alpha}, \qquad x \in [1, \infty). \tag{1.19}$$

Then, by setting appropriately $b_n = 0$, we can write

$$\begin{aligned}-n\bar{F}(a_n z + b_n) &= -n(a_n z + b_n)^{-\alpha} \\ &= \Big[F^{\leftarrow}(1 - \frac{1}{n})\Big]^{\alpha}(a_n)^{-\alpha}(-z^{-\alpha}),\end{aligned}$$

where we define the quantity $F^{\leftarrow}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}$ for $t < 0 < 1$ as the *generalized*

inverse[3] *of F*. Hence, it is easy to see that by setting $a_n = n^{1/\alpha}$ and keeping $b_n = 0$, we have that

$$\Pr\{a_n^{-1} X_{(n)} \leq z\} \to \exp(-z^{-\alpha}),$$

showing that for those particular values of the normalizing constants, we retrieve the Fréchet distribution in the limit of a basic Pareto distribution. The fact that b_n is set to zero can be understood intuitively since for heavy-tailed distribution (see) such as the Pareto distribution, a correction for location is not necessary to obtain a non-degenerate limiting distribution, see Beirlant et al. [1996, pp.51]. More generally, we can state the more general following theorem :

Theorem 1.3 (Pareto-type distributions)**.** *For the same choice of normalizing constants as above, i.e.* $a_n = F^{\leftarrow}(1 - n^{-1}))$ *and* $b_n = 0$ *and for any* $x \in \mathbb{R}$*, if*

$$n[\bar{F}(a_n x)] = \frac{\bar{F}(a_n x)}{\bar{F}(a_n)} \to x^{-\alpha}, \qquad n \to \infty, \tag{1.20}$$

then we say that "$\bar{F}$ is of Pareto-type*" or, more technically, "*$\bar{F}$ is regularly varying with index -α*".* □

This theorem is interesting to get an understanding of the shape of the tails of this kind of distributions. We define the concepts of *regularly varying functions*, together with *slowly varying functions* in Appendix **A.3**

Convergence to EV Weibull distribution

The **type III** or **Weibull** family (?) of distributions $G_3(x)$ arises, for example, in the limit of n iid uniform random variables $X_j \sim U[L, R]$ where L and $R > L$ are both in $\mathbb{R}$ and denote respectively the Left and the Right endpoint of the domain of definition. We have by definition

$$F(x) = \frac{x - L}{R - L}, \qquad x \in [L, R].$$

It is 0 for $x < L$ and 1 for $x > R$. We assume we are in the general case, i.e. $[L, R]$ can be $\neq [0, 1]$. When choosing $a_n = R$ and $b_n = (R - L)/n$, we find the unit Reversed Weibull distribution $We(1, 1)$ in the limit

$$\begin{aligned}
\Pr\{a_n^{-1}(X_{(n)} - b_n) \leq z\} &= \Pr\{X_{(n)} \leq b_n + a_n z\} \\
&= \left[1 - \frac{R - b_n - a_n z}{R - L}\right]^n, && L \leq b_n + a_n z \leq R \\
&= \left(1 + \frac{z}{n}\right)^n \to e^z, && z \leq 0 \quad \text{and} \quad n > |z|.
\end{aligned}$$

That is, the Weibull-type GEV with $\xi = -1$. This is a typical example of the maximal behavior for bounded random variables with continuous distributions.

[3] from which we can retrieve $x_t = F^{\leftarrow}(t)$, the t-quantile of F. Even if we deal in this text only with continuous and strictly increasing df, we prefer consider *generalized inverse*, for sake of generalization.

Conditions and Comments

Intuitively, it stands to reason that the df F needs certain conditions for the limit to exist in Theorem **1.1**. There exists a *continuity condition* at the right endpoint x_* of F which actually rules out many important distributions. For example, it ensures that if F has a jump at its finite x_* (e.g. discrete distributions), then F cannot have a non-degenerate limit distribution as in (1.8). Examples are well documented in Embrechts et al. [1997, section 3.1] for the Poisson, Geometric and Negative Binomial distributions. We cannot find a nondegenerate distribution in the limit for these distributions even after normalization, which limit the scope for applications of the Extremal Types Theorem **1.1**.

Let's finish by noting that it is actually not mandatory to find the normalizing sequences for inferential purposes. We can ignore the normalizing constants in practical applications and fit directly the GEV in our set of maxima. The estimated parameters μ and σ will implicitly take the normalization into account, while the shape parameter ξ is not affected. We will see more details about this at the beginning of Chapter **3** for the stationary case, while methods to estimate (μ, σ, ξ) will be presented in Section **1.6** . Now let's characterize more precisely the distributions pertaining to the GEV.

1.4 Maximum Domain of Attraction

The preceding results can be more easily summarized and obtained when considering *maximum domain of attraction* (MDA). The term "*maximum*" is typically used to make the difference with *sum-stable* distributions but as we only study maxima here, there is no possible confusion in our work. We will then only write "*domain of attraction*" (DA) in the following for convenience, considering these two names as synonyms.

Definition 1.3 (Domain of attraction)**.** *We say that a distribution F is in the* **domain of attraction** *of an extreme value family G in (1.9)-(1.11), denoted by $F \in D(G)$, if there exist $a_n > 0$ and $b_n \in \mathbb{R}$ such that the distribution of $a_n^{-1}(X_{(n)} - b_n)$ converges in distribution to G, where $X_{(n)}$ is the maximum of an iid sequence $\{X_i\}$ with distribution F.* △

Let ξ_k denote the EVI pertaining to some EV distribution G_k $(k = 1, 2, 3)$. From Theorem **1.1**, the domains of attraction are well defined in the sense that $F \in D(G_i)$ and $F \in D(G_j)$ implies $\xi_i = \xi_j$.

We have all the necessary tools to the pertaining DA now. But, before proceeding, we would like to point out that the fact that the characterization of the first DA (the Gumbel type) requires more technicalities going beyond the scope of this thesis. Although this class is important in theory, see e.g. Pinheiro and Ferrari [2015], it is less relevant for our purpose of modelling extremes in a practical case. It often requires other generalizations, for instance with additional parameters to surpass the issues of fitting empirical data. In the last subsection, we will present the unified framework, the domain of attraction pertaining to the GEV distributions, which is a kind of summary for the three first domains of attraction presented.

In each of the characterization of the DA, we will present some of their most useful, necessary (and sometimes sufficient) conditions. We will especially derive their *von Mises conditions*, coming from Von Mises [1936] but revisited in Falk and Marohn [1993]. These conditions are very important in practice and sometimes more intuitive because they make use of the *hazard function* of a df F, defined in the following, for sufficiently smooth distributions :

$$r(x) = \frac{f(x)}{\bar{F}(x)} = \frac{f(x)}{1 - F(x)}. \tag{1.21}$$

It involves the density function $f(x) = \frac{dF(x)}{d(x)}$ in the numerator and it can be thought as a measure of risk. It can be interpreted as the probability of "failure" in an infinitesimally small time period between x and $x + \delta x$ given that the subject has "survived" up till time x.

1.4.1 Domain of attraction for the 3 types of GEV

Domain of attraction for Gumbel distribution (G_1)

We derive here two ways of formulating necessary and sufficient condition for a df F to be in the Gumbel DA, namely $F \in D(G_1)$.

Theorem 1.4. *Following [Beirlant et al., 2006, pp.72], $F \in D(G_1)$ if and only if for some auxiliary function $b(\cdot)$, for every $v > 0$, the condition*

$$\frac{\bar{F}(x + b(x) \cdot v)}{\bar{F}(x)} \to e^{-v}, \tag{1.22}$$

as $x \to x_*$. Then,

$$\frac{b(x + v \cdot b(x))}{b(x)} \to 1.$$

□

A lot of more technical characterizations and conditions together with proofs can be found, for example in Haan and Ferreira [2006, pp.20-33] based on the pioneering thesis of Haan [1970].

Let's now present his ***von Mises criterion*** as in [Beirlant et al., 2006, pp.73]:

Theorem 1.5 (von Mises). *If $r(x)$ (1.21) is ultimately positive in the neighbourhood of x_*, is differentiable there and satisfies*

$$\lim_{x \uparrow x_*} \frac{dr(x)}{dx} = 0, \tag{1.23}$$

then $F \in D(G_1)$. □

In words, the slope of the hazard function with respect to x is zero at the limit when x approaches the (infinite) right-endpoint. This ensures a condition on the lightness of the tails of F.

Examples of distributions in $D(G_1)$: distributions having tails which are exponentially decaying (light-tailed i.e. the exponential, Gamma, Weibull, logistic, ...) but also distributions which are moderately heavy-tailed such as the lognormal. To see that, consider a Taylor expansion, we have that

$$\bar{G}_1(x) = 1 - \exp(-e^{-x}) \sim e^{-x}, \qquad x \to \infty,$$

where "$\sim$" refers to the asymptotic equivalence function. Hence, we directly see the exponential decay of the tails for the Gumbel distribution.

Domain of attraction for Fréchet distribution (G_2)

Let's define $\alpha := \xi^{-1} > 0$ as the *index* of the Fréchet distribution in (1.10).

Definition 1.4 (Power law). *If we look at the tail of the distribution G_2, a Taylor expansion tells us that*

$$\bar{G}_2(x) = 1 - \exp(-x^{-\alpha}) \sim x^{-\alpha}, \qquad x \to \infty, \tag{1.24}$$

which means that G_2 tends to decrease as a power law. △

Theorem 1.6. *We have $F \in G_2$ if and only if*

$$\bar{F}(x) = x^{-\alpha} L(x), \tag{1.25}$$

for some slowly varying *function L* (see Appendix **A.3** for the definition). □

In this case and with $b_n = 0$,

$$F^n(a_n x) \xrightarrow{d} G_2(x), \qquad x \in \mathbb{R},$$

with

$$a_n := F^{\leftarrow}\Big(1 - \frac{1}{n}\Big) = \Big(\frac{1}{1-F}\Big)^{\leftarrow}(n).$$

This previous theorem informs us that all dfs $F \in D(G_{2,\alpha})$ necessarily have an infinite right endpoint, that is $x_* = \sup\{x : F(x) < 1\} = \infty$. These distributions are all with regularly varying right-tail with index $-\alpha$ (see Appendix **A.3**), that is $F \in D(G_{2,\alpha}) \iff \bar{F} \in R_{-\alpha}$.

Finally, let's now present the (revisited) ***Von Mises condition*** for this DA which states the following in Falk and Marohn [1993]:

Theorem 1.7 (von Mises). *If F is absolutely continuous with density f and $x_* = \infty$ such that*

$$\lim_{x \uparrow \infty} x \cdot r(x) = \alpha > 0,$$

then $F \in D(G_{2,\alpha})$. □

We illustrate this with the standard Pareto distribution, that is

$$F(x) = \Big(1 - (\frac{x_m}{x})^{\alpha}\Big) 1_{x \geq x_m}, \qquad \alpha > 0 \text{ and } x_m > 0.$$

Clearly, we can see that by setting $K = x_m^{\alpha}$, we obtain $\bar{F}(x) = K x^{-\alpha}$. Therefore, we have that $a_n = (Kn)^{\alpha^{-1}}$ and $b_n = 0$.

Examples of distributions in $D(G_2)$: distributions that are typically (very) fat-tailed (or heavy-tailed, see Appendix **A.1**) distributions, such that $E(X_+)^{\delta} = \infty$ for $\delta > \alpha$. This class of distributions is thus appropriate for phenomena with extremely large maxima, think for example of the rainfall process in some tropical zones. Common distributions include Pareto, Cauchy, Burr,... An example to get an idea of this is by looking (1.24) showing that G_2 tends to decrease as a *power law*.

Domain of attraction for the EV Weibull distribution (G_3)

We start by recalling an important relation between the Fréchet and the EV Weibull distributions

$$G_3(-x^{-1}) = G_2, \qquad x > 0.$$

We pointed out the certain symmetry that occurs for these two types (e.g. recall figure 1.1). Hence, this will be useful to characterize $D(G_3)$ using what we know about the Fréchet case.

Theorem 1.8. *We say that $F \in G_3$ as in (1.11) with index $\alpha = \xi^{-1} > 0$ if and only if there exists a finite right endpoint $x_* < \infty$ such that*

$$\bar{F}(x_* - x^{-1}) = x^{-\alpha} L(x), \tag{1.26}$$

where $L(\,\cdot\,)$ is a slowly varying function. □

Hence for $F \in D(G_{3,\alpha})$, we have

$$a_n = x_* - F^{\leftarrow}(1 - n^{-1}), \qquad b_n = x_*,$$

and hence

$$a_n^{-1}\Big(X_{(n)} - b_n\Big) \xrightarrow{d} G_3.$$

Finally, we present the ***Von Mises condition*** related to the G_3 DA.

Theorem 1.9 (von Mises)**.** *For F having positive derivative on some $[x_0, x_*)$, with finite right endpoint $x_* < \infty$, then $F \in D(G_3)$ if*

$$\lim_{x \uparrow x_*} (x_* - x) \cdot r(x) = \alpha > 0. \tag{1.27}$$

□

Similarly to the Fréchet case, we remark that there is still a probability mass from the hazard rate when x approaches its finite right endpoint, characterized by a non-null constant α which defines the left heavy tail and the right endpoint.

Examples of distributions in $D(G_3)$: dfs that are bounded to the right ($x_* < \infty$). Whereas the Fréchet type is often more preferable in an extreme analysis context because it allows for arbitrarily large values, most phenomena are typically bounded, hence we will think at the EV Weibull for the most attractive and flexible class for modelling extremes. For example, in our case of modelling a process of maximum temperatures, it seems to be the perfect candidate.

1.4.2 Closeness under tail equivalence property

An interesting property of all the three types of DA $D(G_k)_{k=1,2,3}$ we have derived, is that those are *closed under tail-equivalence*. This is useful for characterizing tail's types of the distributions falling in the pertaining DA. In this sense,

1. For the **Gumbel** DA, let $F \in D(G_{1,\alpha})$. If H is another df such that, for some $b > 0$,

$$\lim_{x\uparrow x_*} \frac{\bar{F}(x)}{\bar{H}(x)} = b, \tag{1.28}$$

then $H \in D(G_{1,\alpha})$. This emphasizes exponential type of the tails for H in the Gumbel DA.

2. For the **Fréchet** DA, let $F \in D(G_{2,\alpha})$. If H is another df such that, for some $c > 0$,

$$\lim_{x\to\infty} \frac{\bar{F}(x)}{\bar{H}(x)} = c, \tag{1.29}$$

then $H \in D(G_{2,\alpha})$.

3. For the **Weibull** DA, let $F \in D(G_{3,\alpha})$. If H is another df such that, for some $c > 0$,

$$\lim_{x\uparrow x_*} \frac{\bar{F}(x)}{\bar{H}(x)} = c, \tag{1.30}$$

then $H \in D(G_{3,\alpha})$.

This emphasizes the polynomial decay for the tails of the distributions falling in the Fréchet or in the Weibull DA.

1.4.3 Domain of attraction of the GEV

The conditions that have been stated for the three preceding DA can be restated under an "unified" framework for the GEV distribution defined in (1.12). For a given df F that is sufficiently smooth, by letting the sequences b_n, a_n, and the shape parameter such that

$$b_n = F^{\leftarrow}(1 - n^{-1}), \qquad a_n = r(b_n) \qquad \text{and} \qquad \xi = \lim_{n\to\infty} r'(x),$$

then, $a_n^{-1}(X_{(n)} - b_n)$ has the GEV as nondegenerate limiting distribution which density is denoted in Table 1.1. This is a sufficient condition. Among many characterizations, we present the following.

Theorem 1.10. *Let F be the df of a sequence $\{X_i\}$ iid. For $u(\,\cdot\,) > 0$ measurable and $\xi \in \mathbb{R}$, $F \in D(GEV)$ if and only if:*

$$\lim_{v\uparrow x_*} \Pr\left\{ \frac{X - v}{u(v)} > x \mid X > v \right\} := \lim_{v\uparrow x_*} \frac{\bar{F}(v + x\cdot u(v))}{\bar{F}(v)} = \begin{cases} \left(1 + \xi x\right)_+^{-\xi^{-1}}, & \xi \neq 0; \\ e^{-x}, & \xi = 0. \end{cases} \tag{1.31}$$

□

We will see in Chapter 2 that it actually defines the "Peaks-Over-Threshold" model.

1.5 Return Levels and Return Periods

After having defined the theoretical properties of distributions pertaining to the GEV family precisely, we are now interested in finding a quantity that could significantly improve the interpretability of such

models. *Return levels* play a major role in environmental analysis. For such tasks, it is usually more convenient to interpret EV models in terms of insightful return levels rather than individual parameter estimates.

Assuming for this introductory example our time unit reference is in year -as usually assumed in meteorological analysis-, let us consider the *m-year return level* r_m which is defined as the high quantile for which the probability that the annual maximum exceeds this quantile is $(\lambda \cdot m)^{-1}$, where λ is the mean number of events that occur in a year. For yearly blocks, we have $\lambda = 1$ which will facilitate the interpretation. We call m the *return period* and define it to a reasonable degree of accuracy as the expected time between the occurrence of two so-defined high-quantiles. For example, under stationary assumption, if the 100-year return level is $37°c$ for the sequence of annual maximum temperatures, then $37°c$ is the temperature that is expected to be reached once in average within a period of 100 years. More precisely, you can see it such that r_m is exceeded by the annual maximum in any particular year with probability m^{-1}.

Let $\{X_{(n),y}\}$ denote the iid sequence of n random variables representing the annual maximum for a particular year y. From (1.12), we have

$$\begin{aligned} F(r_m) &:= \Pr\{X_{(n),y} \leq r_m\} = 1 - 1/m \\ \Leftrightarrow \quad & \left[1 + \xi\left(\frac{r_m - \mu}{\sigma}\right)\right]^{-\xi^{-1}} = \frac{1}{m}. \end{aligned}$$

Hence, by inverting this relation, and by letting $y_m = -\log(1 - m^{-1})$, we can retrieve the quantile of the GEV, namely the *return level* r_m

$$r_m = \begin{cases} \mu + \sigma\xi^{-1}(y_m^{\xi} - 1), & \xi \neq 0; \\ \mu + \sigma \log(y_m), & \xi = 0. \end{cases} \tag{1.32}$$

Henceforth, after having estimated the model (that will be the subject of Section **1.6**), we can replace the estimated parameters $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$ in (1.32) to obtain an estimate of the m-year return level.

However, we recall that the definition of return period is easily misinterpreted and the one given above is thus not universally accepted. To tackle this issue, it is important to distinguish stationary from nonstationary sequences. We investigate the return periods and return levels more precisely by relaxing the independence assumption (stationary) and then under a climate change environment (nonstationary) in Section **3.3**. Regarding the diagnostic issues of the model, we present the *return level plot* in Section **1.7.1**.

1.6 Inference

As already stated, a great advantage for the modeling of GEV is that we actually do not have to find the normalizing sequences to estimate the parameters of the model. Hence, we will present in this section the main (frequentists) methods of inference for the GEV. These are mostly based on the likelihood (Section **1.6.1**) but we will also present a few other methods that are widely used to estimate GEV parameters like the (probability weighted) moment estimator (Section **1.6.2**). Finally, note that there exist estimators for the EVI ξ only, but we leave that for Section **A.5**. After all, we will mostly rely on

the Bayesian inference in Chapter **4**.

1.6.1 Likelihood-based Methods

The most usual inference we will first consider is Maximum Likelihood (ML). It generally does a good job, it is intuitive to understand. Only its implementation could bring some problems.

Depicted by Smith [1985], the potential difficulty with the use of likelihood methods for the GEV concerns the regularity conditions that are required for the usual asymptotic properties associated with the Maximum Likelihood Estimator (MLE) to be valid. Such conditions are not satisfied by the GEV model because the endpoints of the GEV distribution are functions of the parameter value[4]. Depending on the value of the EVI ξ, the special cases are :

1. $\boxed{\boldsymbol{\xi < -1}}$: MLE's are unlikely to be obtainable.
2. $\boxed{\boldsymbol{\xi \in (-1, -0.5]}}$: MLE's are usually obtainable but standard asymptotic properties do not hold.
3. $\boxed{\boldsymbol{\xi > -0.5}}$: MLE's are regular, in the sense of having the usual asymptotic properties.

Fortunately in practice, problematic cases ($\xi \leq -0.5$) are rarely encountered in most environmental problems. This corresponds to distributions in the Weibull family with very short bounded upper tail, see for example the red density in Figure 1.1 or Figure C.1, or directly in the Shiny application where we better see what defines the borders of the problematic case. The 'bell' of the curve becomes very narrow. In the problematic cases, Bayesian inference which does not depend on these regularity conditions may be preferable. We will see in Part **II** that the distribution of the yearly maximum temperature is upper bounded which lead us to consider Bayesian inference in Chapter **4**.

Other forms of likelihood-based methods have also emerged to remedy this problem of instability for low values of ξ. Close to a Bayesian formulation, *penalized ML* method has been proposed by Coles and Dixon [1999] which adds a penalty term to the likelihood function to "force" the shape parameter to be > -1, values closer to -1 being more penalized. We will actually use this concept in Section **??** trying to circumvent issues of these likelihood computations in nonstationary sequences, and to add more flexibility through the neural architecture.

We are now considering a sequence $\{Z_i\}_{i=1}^n$ of independent random variables sharing each the same GEV distribution. Let $\boldsymbol{z} = (z_1, \ldots, z_n)$ denote the vector of observations. From the densities of the GEV distribution $g_\xi(z)$ defined in Table 1.1, we can derive the log-likelihood $\ell = \log\left[L(\mu, \sigma, \xi; \boldsymbol{z})\right]$, for the two different cases $\xi \neq 0$ or $\xi = 0$ respectively:

1.
$$\ell(\mu, \sigma, \xi \neq 0\,; \mathbf{z}) = -m\log\sigma - (1+\xi^{-1})\sum_{i=1}^{n}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]_+ - \sum_{i=1}^{n}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}}, \tag{1.33}$$

2.
$$\ell(\mu, \sigma, \xi = 0\,; \mathbf{z}) = -m\log\sigma - \sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\exp\left\{-\left(\frac{z_i-\mu}{\sigma}\right)\right\}. \tag{1.34}$$

[4] We already pointed this in the paragraph below (1.17).

Maximization of this pair of equations with respect to $\boldsymbol{\theta} = (\mu, \sigma, \xi)$ leads to the MLE with respect to the entire GEV family. Note that there is no analytical solution and hence, it must be numerically optimized.

From standard MLE theory, we know that the estimated parameter vector $\hat{\boldsymbol{\theta}}$ will be approximately multivariate normal. Inference such as confidence intervals can thus be applied, relying on this approximate normality of the MLE. Hence, problems of this method arise when the approximate normality cannot hold. The underlying inferences will not be sustainable. Whereas Zhou [2010] closed the discussion on the theoretical properties of the MLE, another method is usually more preferable for inference, the *profile likelihood*.

Profile Likelihood

In general, the normal approximation to the true sampling distribution of the respective estimator is rather poor. The *profile likelihood* is often more convenient when a single parameter is of interest. Let's denote it θ_j. Now let's consider the parameter vector $\boldsymbol{\theta} = (\theta_j, \boldsymbol{\theta_{-j}}) = (\mu, \sigma, \xi)$ typically in EVT in a stationary context, where $\boldsymbol{\theta_{-j}}$ corresponds to all components of $\boldsymbol{\theta}$ except θ_j. Hence, $\boldsymbol{\theta_{-j}}$ can be seen as a vector of nuisance parameters. The profile log-likelihood for θ_j is defined by

$$\ell_p(\theta_j) = \arg\max_{\boldsymbol{\theta_{-j}}} \ell(\theta_j, \boldsymbol{\theta_{-j}}). \tag{1.35}$$

Henceforth for each value of θ_j, the profile log-likelihood is the maximised log-likelihood with respect to $\boldsymbol{\theta_{-j}}$, i.e. with respect to all other components of $\boldsymbol{\theta}$ but not θ_j. Generalization where $\boldsymbol{\theta_j}$ is of dimension higher than one (e.g. in a nonstationary context) is possible.

Another interpretation is related to the χ^2 distribution and the equality with the hypothesis testing the Gumbel case. Details can be found in Beirlant et al. [2006, pp.138]. Applications of likelihood inferences will be provided in Section **6.1**.

1.6.2 Other Estimator : Probability-Weighted-Moments

Introduced by Greenwood et al. [1979], the *Probability-Weighted-Moments* (PWM) of a random variable X with df F, are the quantities

$$M_{p,r,s} = \mathbb{E}\Big\{X^p[F(X)]^r[1 - F(X)]^s\Big\}, \tag{1.36}$$

for real p, r and s. From (1.36), we can retrieve the PWM estimator from specific choices of p, r and s.

1.7 Model Diagnostics : Goodness-of-Fit

After having fitted a statistical model to data, it is important to assess its accuracy in order to infer reliable conclusions from this model. Ideally, we aim to check that our model fits well the whole population, e.g. the whole distribution of maximum temperatures, i.e. all the past and future temperature maxima... As this cannot be achieved in practice, it is common to assess a model with the data that were used to estimate this model. The aim here is to check that the fitted model is acceptable for the available data.

As these concepts are generally known for a statistician, we decide to let in Appendix **A.4** a reminder of *quantile* and *probability plots* applied in the world of extremes.

1.7.1 Return Level Plot

Section **1.5** introduced the concept of return levels and its intuitive interpretations. Now, we will use this quantity as a diagnostic tool for model checking. Approximate confidence intervals for the return levels can be obtained by the delta method which relies on the asymptotic normality of the MLE and hence produces symmetric confidence intervals.

Standard errors of the estimates

We naturally expect the standard errors to increase with the return period. Indeed, it is less accurate to estimate 100-year than a 2-year return level. As r_m is a function of the GEV parameters, we use the *delta method* to approximate the variance of $\hat{r}_m$. Specifically,

$$\text{Var}(\hat{r}_m) \approx \nabla r_m^{'} V \nabla r_m,$$

with V the variance-covariance matrix of the estimated parameters $(\hat{\mu}, \hat{\sigma}, \hat{\xi})'$ and

$$\begin{aligned} \nabla r_m^{'} &= \left[\frac{\partial r_m}{\partial \mu}, \frac{\partial r_m}{\partial \sigma}, \frac{\partial r_m}{\partial \xi} \right] \\ &= \left[1,\ \xi^{-1}(y_m^{-\xi} - 1),\ \sigma\xi^{-2}(1 - y_m^{-\xi}) - \sigma\xi^{-1} y_m^{-\xi} \log y_m \right], \end{aligned} \tag{1.37}$$

with $y_m = -\log(1 - m^{-1})$ and the gradient being evaluated at the estimates $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

A problem arise for the so-computed standard errors when considering long-range return levels. They can increase so drastically with the return period that the confidence intervals of the *return level plot* can become difficult to work with. To try to get rid of this issue and to allow for we will construct intervals on the basis of the *profile* log-likelihood. Finally, note that this inference relies on the model adequacy and hence, more uncertainty should be given if the model fit is not perfect.

Profiled likelihood Return levels

Usual likelihood methods are not the most accurate for inference in EVT. The problem was that confidence intervals computed in the usual method, with standard errors computed by the Delta method in (1.37), relying on the normal approximation, was not reliable for inference on return levels. This is due to severe asymmetries that are often observed in the likelihood surface for return levels, especially for large quantiles, see e.g. Bolívar et al. [2010].

Profile likelihood method is hence more accurate for confidence intervals as it better captures the skewness generally associated with return level estimates. We are now specifically interested in computing the profile log-likelihood for the estimation of the return level r_m. To do that, we present a method which consists of three main steps :

1. To include r_m as a parameter of the model, by (1.32) we can rewrite μ

$$\mu = r_m - \sigma\xi^{-1}\Big[\Big(-\log\{1-m^{-1}\}\Big)^{-\xi} - 1\Big].$$

as a function of ξ, σ and r_m. By plugging it in the log-likelihood in (1.33)-(1.34), we obtain the new GEV log-likelihood $\ell(\xi, \sigma, r_m)$ as a function of r_m.

2. We maximise this new likelihood $\ell(\xi, \sigma, r_m = r_m^-)$ at some fixed low value of $r_m = r_m^- \leq r_m^+$ with respect to the nuisance parameters (ξ, σ) to obtain the profiled log-likelihood

$$\ell_p(r_m = r_m^-) = \underset{(\xi,\sigma)}{\arg\max}\, \ell\Big(r_m = r_m^-\,,\ (\xi, \sigma)\Big).$$

We choose arbitrarily large value of the upper range r_m^+, and conversely for starting point of r_m^-.

3. Repeat the previous step for a range of values of r_m such that $r_m^- \leq r_m \leq r_m^+$ and then choose r_m which attain the maximum value of $\ell_p(r_m)$.

Doing this little algorithm gives the (profiled log-likelihood) *return level plot*.

Interpretation

Generally plotted against the return period on a logarithmic scale, the return levels has different shapes depending on the value of the shape parameter ξ, namely :

- If $\boldsymbol{\xi = 0}$, then return level plot will be **linear**.
- If $\boldsymbol{\xi < 0}$, then return level plot will be **concave**.
- If $\boldsymbol{\xi > 0}$, then return level plot will be **convex**.

This can be easily understood as we have seen that $\xi < 0$ implies an upper endpoint and heavy left tail while $\xi > 0$ implies the converse. Henceforth, the "increasing rate" of the return level will decrease as the return period increases for $\xi < 0$ as it cannot go too far away beyong the upper endpoint, and the converse holds for $\xi > 0$. You can already look at Figure 6.2 to visualize shape of this plot in our application ($\xi < 0$).

Chapter 2

Peaks-Over-Threshold Method

Contents

This new chapter will focus on another major approach of EV models by modeling only the excess over a certain threshold. This approach is very popular in practice as it can handle all the extremes with more flexibility and not only the maximum of one block. From this fact, numerous techniques have emerged and we will quickly navigate the main ideas.

In Section **2.1**, we will intuitively introduce the approach that will help us to formally characterize the resulting distribution of interest in Section **2.2**. Then, Section **2.3** will present the concept of return levels applied to this concept. Next, we will assess the maximum temperature threshold in a statistical sense in Section **2.5** by reviewing available methods, regardless common suggested thresholds of $25°c$ or $30°c$ from meteorologists.

Unfortunately, the resulting analysis of this chapter will not be present in Part **II** because it would have made the text become too voluminous. But empirical results (code and html reports) will remain available in the repository[1] which structure is explained in Appendix **D**. Moreover, *point process* will not be covered for the same reason but we remind that this is a powerful and flexible method that summarizes the two techniques considered in the two first chapters, and it should then not be missed.

This chapter is mainly based on Coles [2001, chap.4 and 7], Beirlant et al. [2006, chap.4], Reiss and Thomas [2007, chap.5] and Embrechts et al. [1997, chap.5], and other relevant articles.

[1] https://github.com/proto4426/PissoortThesis/

2.1 Preliminaries

Threshold models relying on the *Peaks-Over-Threshold* (POT) method propose an alternative to the blocking method seen in the previous Chapter. Focusing exclusively on observations greater than a pre-specified *threshold* provides a natural way to expose extreme values. POT solves the issue of choosing a single observation per data block, but makes the trade-off at the expense of threshold determination and independence issues since cold days are more likely to be followed by cold days, etc ; more details follow in Chapter **3**. The notion of "extremes" is therefore intrinsically different.

Let $\{X_j\}$ be a sequence of n iid random variables with marginal df F, shown in Figure **2.1** representing our application's data. Next we look at observations exceeding a threshold u (blue) that must be lower than the right endpoint (1.17) of F. The aim here is to find the child distribution that is depicted in red, say H, from the parent distribution F depicted in grey. It will allow us to model the exceedances $Y = X - u$ with H expressed as

$$H(y) = \Pr\{X - u \leq y \mid X > u\}.$$

Throughout this chapter we will aim to model this empirical distribution.

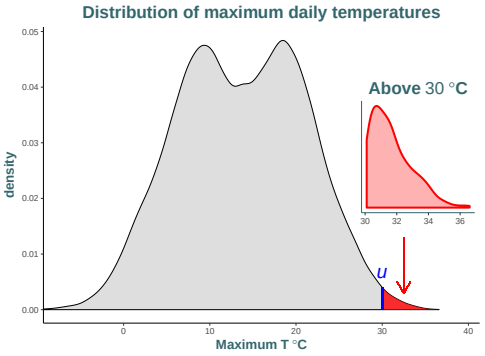

Figure 2.1: *Kernel density estimates for the whole series (grey) and for the series of excess over the threshold of $u = 30°c$ (red). More information on the data will be given in Part **II**.*

Threshold models can be seen as the conditional survival function of the exceedances Y, knowing that the threshold is exceeded, according to Beirlant et al. [2006, pp.147] :

$$\Pr\{Y > y \mid Y > 0\} = \Pr\{X - u > y \mid X > u\} = \frac{\bar{F}(u+y)}{\bar{F}(u)}. \tag{2.1}$$

Or similarly, in terms of the exceedance df $F^{[u]}(x) = \Pr\{X \leq u + x \mid X > u\}$, see Reiss and Thomas [2007, pp.25-29] or following Charras-Garrido and Lezaud [2013] and Rosso [2015], using the conditional probability law.

If the parent distribution F was known, the threshold exceedances distribution could be computed (2.1). But, as in the method of block maxima, F is unknown in practice. We will still rely on approximations[2]. Whereas Theorem **1.1** were used to find an approximate distribution for block maxima, we will attempt here to approximate the distribution of the exceedances $H(y)$.

2.2 Characterization of the Generalized Pareto Distribution

Similarly to the GEV (1.12) in the limit for the block maxima, we look for a limit distribution for exceeding a certain threshold. As for max-stability in Definition **1.2**, another formulation can be given for POT models and will help derive a theorem.

[2]We quote here the well-known "*All models are wrong, but some are useful*" from George Box & Draper (1987), *Empirical model-building and response surfaces*, Wiley, p.424

Definition 2.1 (POT-stability). *The dfs H are the only continuous one such that, for a certain choice of constants a_u and b_u,*

$$F^{[u]}(a_u x + b_u) = F(x).$$

△

We can now state the key theorem discovered by Balkema and Haan [1974] and Pickands [1975].

Theorem 2.1 (Pickands–Balkema–de Haan). *Let $\{X_j\}$ be the sequence of n iid random variables having marginal df F for which Theorem **1.1** holds. Then,*

$$Pr\{X - u \leq y \mid X > u\} \longrightarrow H_{\xi,\sigma_u}(y), \qquad u \to x_*. \tag{2.2}$$

It means that for large enough u, the df of $Y = X - u > 0$, conditional on $X > u$, is approximately $H_{\xi,\sigma_u}(y)$ where $H_{\xi,\sigma_u}(y)$ is the Generalized Pareto Distribution *(GPD) :*

$$H_{\xi,\sigma_u}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)_+^{-\xi^{-1}}, & \xi \neq 0; \\ 1 - \exp\left\{-\frac{y}{\sigma_u}\right\}, & \xi = 0. \end{cases} \tag{2.3}$$

□

The scale parameter is denoted σ_u to emphasize its dependency with the specified threshold u :

$$\sigma_u = \sigma + \xi(u - \mu), \tag{2.4}$$

where we notice the absence of location parameter μ in (2.3) as it appears in (2.4).

2.2.1 Outline proof of the GPD and justification from GEV

In the following sections we describe a meaningful yet not too technical way of retrieving the GPD.

Outline Proof of Theorem **2.1** :

- From Theorem **1.1**, we have for the distribution of the maximum, for large enough n,

$$F_{X_{(n)}}(z) = F^n(z) \approx \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\xi^{-1}}\right\}, \tag{2.5}$$

  with $\mu, \sigma > 0$ and ξ the GEV parameters. Hence, by taking logarithm on both sides, we get

$$n \ln F(z) \approx -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\xi^{-1}}. \tag{2.6}$$

- We also have that, from Taylor expansion, $\ln F(z) \approx -[1 - F(z)]$ as both sides go to zero as $z \to \infty$. Therefore, substituting into (2.6), we get the following for large u :

$$1 - F(u + \boldsymbol{y}) \approx n^{-1}\left[1 + \xi\left(\frac{u + \boldsymbol{y} - \mu}{\sigma}\right)\right]^{-\xi^{-1}}.$$

where we added the term $\boldsymbol{y} > 0$ to retrieve something in the form of (2.1).

- Finally, by mathematical manipulation, as $u \to x_*$:

$$\begin{aligned}\Pr\{X > u + y \mid X > u\} = \frac{\bar{F}(u+y)}{\bar{F}(u)} &\approx \frac{n^{-1}\left[1 + \xi\sigma^{-1}(u + y - \mu)\right]^{-\xi^{-1}}}{n^{-1}\left[1 + \xi\sigma^{-1}(u - \mu)\right]^{-\xi^{-1}}} \\ &= \left[1 + \frac{\xi\sigma^{-1}(u + y - \mu)}{1 + \xi\sigma^{-1}(u - \mu)}\right]^{-\xi^{-1}} \\ &= \left[1 + \frac{\xi y}{\sigma_u}\right]_+^{-\xi^{-1}},\end{aligned} \tag{2.7}$$

By simply taking the survivor of (2.7), we retrieve

$$\begin{aligned}\Pr\{X - u \leq y \mid X > u\} &= 1 - \Pr\{X > u + y \mid X > u\} \\ &= 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)_+^{-\xi^{-1}},\end{aligned} \tag{2.8}$$

which is $GPD(\xi, \sigma_u)$ as required.

□

More insights come from Reiss and Thomas [2007, pp.27-28] with analysis of rates of convergence. Examples of how to retrieve the GPD from specific distributions are available in Coles [2001, pp.77].

2.2.2 Dependence of the scale parameter

We chose to express the scale parameter as σ_u to emphasize its dependency with the threshold u. If we increase the threshold, say to $u' > u$, then the scale parameter will be adjusted :

$$\sigma_{u'} = \sigma_u + \xi(u' - u), \tag{2.9}$$

and in particular, this adjusted parameter $\sigma_{u'}$ will increase if $\xi > 0$ and decrease if $\xi < 0$. If $\xi = 0$, there would be no change in the scale parameter[3]. Also, we consider it noteworthy that as for the GEV, the scale parameter σ_u for GPD models differs from the standard deviation since it governs the "size" of the excess, as mentioned in **?**, pp.20. The issue of threshold selection will be discussed in Section **2.5**.

2.2.3 Three different types of GPD : Comparison with GEV

One should remark the similarity with the GEV distributions. Indeed, parameters of the GPD of the threshold excesses are uniquely determined by the corresponding GEV parameters of block maxima. Hence, the shape parameter ξ of the GPD is equal to that of the corresponding GEV and, most of all,

[3] Consistent with the well-known *memoryless property* of the exponential distribution H_{0,σ_u}

it is invariant. In block maxima, the GEV parameters would shift for a different block length, while in POT the GPD parameters are not affected by the choice of the threshold due to the self-compensation arising in (2.9).

Hence, as for the block-maxima approach, there are also three possible families of the GPD depending on the value of the shape parameter ξ which determines the qualitative behaviour of the corresponding GPD. Hosking and Wallis [1987], Singh and Guo [1995]

- **First** type $H_{0,\sigma_u}(y)$ comes by letting the shape parameter $\xi \to 0$ in (2.3), giving :

$$H_{0,\sigma_u}(y) = 1 - \exp\Big(-\frac{y}{\sigma_u}\Big), \qquad y > 0. \tag{2.10}$$

  One can easily notice that it corresponds to an **exponential** df and hence light-tailed with parameter $1/\sigma_u$, namely $Y \sim \exp(\sigma_u^{-1})$.

- **Second** and **third** types, i.e. when $\xi < 0$ and $\xi > 0$ (resp.), differ only by their support :

$$H_{\xi,\sigma_u}(y) = 1 - \Big(1 + \frac{\xi y}{\sigma_u}\Big)^{-\xi^{-1}}, \qquad \text{for} \quad \begin{cases} y > 0, & \xi > 0; \\ 0 < y < \sigma_u \cdot |\xi|^{-1}, & \xi < 0. \end{cases} \tag{2.11}$$

  Therefore, if $\xi > 0$ the corresponding GPD is of **Pareto**-type, hence is heavy-tailed, and has no upper limit while if $\xi < 0$, the associated GPD has an upper bound $y_* = u + \sigma_u/|\xi|$ and is then of **Beta**-type. Special case arise when $\xi = -1$ where the pertaining distribution becomes Uniform$(0, \sigma_u)$, result coming from Grimshaw [1993, pp.186].

The **density** of the GPD is written here as

$$h_{\xi,\sigma_u}(y) = \begin{cases} \frac{1}{\sigma_u}\Big(1 + \xi\frac{y}{\sigma_u}\Big)^{-\xi^{-1}-1}, & \xi \neq 0; \\ \sigma_u^{-1} \cdot e^{-y}, & \xi = 0. \end{cases} \tag{2.12}$$

In the example of Figure 2.1, we can see that the red density seems to have an upper bound around $36.5°c$. thence, we can conclude that this distribution should be of Beta-type. Part **II** will confirm that $\xi < 0$.

After looking at the behaviour of the density of these functions, we will procure a more comprehensive view by defining some examples of how to retrieve these different types of Generalized Pareto Distributions.

2.3 Return Levels

Presented in Section **1.6** for block maxima, return levels are also useful in POT to bring valuable information. However, unlike for block maxima, the quantiles of the GPD cannot be as readily interpreted as return levels because the observations no longer derive from predetermined *blocks* of equal length. Instead, it is now required to estimate the *probability of exceeding the threshold* $\zeta_u = \Pr\{X > u\}$, from which a natural estimator is $\hat{\zeta}_u = k/n$ with k the number of points exceeding u. We can now retrieve the return level r_m, i.e. the *value that is exceeded on average once every* m *observations*. This is given by

$$r_m = \begin{cases} u + \sigma_u \, \xi^{-1} \Big[(m\zeta_u)^{\xi} - 1 \Big], & \xi \neq 0; \\ u + \sigma_u \log(m\zeta_u), & \xi = 0. \end{cases} \tag{2.13}$$

provided m is sufficiently large. Computations are very similar as for the GEV in Section **1.7.1**.

Interpretation

Whereas the interpretation of the plot in function of the shape parameter value is the same as for the block-maxima method (see Section **1.8**), it is more convenient to replace the value of m by $N \cdot n_y$ in (2.13), where n_y is the number of observations per year, to give return levels on an annual scale. This method allows us to obtain the *N-year return level* which is now commonly defined as the level expected to be exceeded once every N years.

2.4 Inference : Parameter Estimation

We will not develop likelihood techniques here as is ressembles that of GEV in Section **1.6**, and requires numerical techniques as well.

The two approaches we have encountered so far, i.e. block-maxima and POT share the same parameter ξ. Therefore, it is not necessary to differentiate between these methods for the sole estimate of the shape parameter. We give in Appendix **A.5** some of the most used methods to estimate the EVI.

Finally, methods dedicated to POT exist to estimate the parameters of the GPD. For example, they can include the probability-weighted-moment (PWM) which is formulated differently as for GEV in Section **1.6.2**, see e.g. Ribereau et al. [2016]. The L-moment estimator is also important, especially for rainfall application, where Hosking and Wallis [1997] emphasized that L-moment method came historically as a modification of the PWM estimator.

2.5 Inference : Threshold Selection

Threshold selection is crucial in a POT context. It involves a *bias-variance trade-off*, that is :

- *Lower threshold* will induce *higher bias* due to model mispecification. In other words, the threshold must be sufficiently high to ensure that the asymptotics underlying the GPD approximation are reliable.
- *Higher threshold* will imply higher estimation uncertainty, i.e. *higher variance* of the parameter estimate as the sample size is reduced for high threshold.

2.5.1 Standard Threshold Selection Methods

Applications of all the subsequent methods can be viewed in Section 3.1 of the **Summary1_intro.Rmd** in the **/vignettes** folder of the repository. Vignette can be downloaded in html from the compressed file in the same folder. Recall that all empirical results of this chapter will not be included in this thesis.

Based on Mean Residual Life

The *mean residual life* function or *mean excess* function is defined as

$$mrl(u_0) := E(X - u_0 \mid X > u_0) = \frac{\int_{u_0}^{x_*} \bar{F}(u)du}{\bar{F}(u_0)}, \tag{2.14}$$

for X having survival function $\bar{F}(u_0)$ computed at u_0. It denotes in an actuarial context the expected remaining quantity or amount to be paid out when a level u_0 has been chosen. However, there are also interesting and reliable applications in an environmental context. Moreover, this function yields valuable properties about the tail of the underlying distribution of X. In fact, we expect that :

- If $mrl(u_0)$ is constant, then X is exponentially distributed.
- If $mrl(u_0)$ ultimately increases, then X has a heavier tail than the exponential distribution.
- If $mrl(u_0)$ ultimately decreases, then X has a lighter tail than the exponential distribution.

This can be particularly interesting for our purpose when considering threshold models. For this case, we have the excesses $\{Y_i\}$ that follow a GPD (2.3) and which are generated by the sequence $\{X_i\}$. From the theoretical mean of this distribution, we retrieve (provided $\xi < 1$)

$$\begin{aligned} mrl(u) := E(X - u \mid X > u) &= \frac{\sigma_u}{1-\xi} \\ &= \frac{\sigma_{u_0} + \xi u}{1-\xi}, \end{aligned} \tag{2.15}$$

from dependence of the scale parameter σ with the threshold u, see (2.9). Hence, we remark that $mrl(u)$ is linearly increasing in u, with gradient $\xi \cdot (1-\xi)^{-1}$ and intercept $\sigma_{u_0} \cdot (1-\xi)^{-1}$. Furthermore, we can estimate empirically this function by

$$\widehat{mrl}(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{[i]} - u), \tag{2.16}$$

where the $x_{[i]}$ denote the i-th observations out of the n_u that exceed u.

Mean residual life plot The *mean residual life plot* results from combining the linearity detected between $mrl(u)$ and u in (2.15) with (2.16). Therefore, worthwhile information can be retrieved from the locus of the points :

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{[i]} - u) \right) : u < x_{max} \right\}. \tag{2.17}$$

Even if its interpretation is not straightforward, this graphical procedure will give insights for the choice of a suitable threshold u_0 to model extremes via a GPD, that is the threshold u_0 above which we can detect linearity in the plot. Relying on this well-chosen threshold u_0, the GPD approximation should be correct, even though its interpretation is subjective. Furthermore, information in the far right-hand-side of this plot is unreliable. Variability is high due to the limited amount of data above high thresholds.

Based on the stability of the parameter's estimates

Due to its simplicity, *stability plots of the parameter's estimates* is one of the preferred tools for practitioners. The aim is to plot MLE's of the parameters against different values for the threshold. In theory, MLE's are independent of the threshold choice, and hence the threshold is chosen at the lowest value for which the MLE's remain near-constant.

But this method is also criticized, especially owing to its lack of interpretability, and the pointwise confidence interval strongly dependent across the range of thresholds. Other techniques have thus been proposed, see e.g. Wadsworth [2016] which suggests complementary plots with greater interpretability, with a likelihood-based procedure allowing for automated and a more formal threshold selection. In short, this new method relies on the independent-increments structure of MLE and makes use of likelihood ratio tests to identify the threshold that significantly provides the best fit to the data. Thresholds are then compared iteratively, and significance is assessed by simulations.

Based on the Dispersion Index Plot

As mentioned, methods considered above involve substantial amount of subjectivity. Following Ribatet [2006], the *Dispersion Index* (DI) plot is particularly useful for time series. Point process that we have not developed here, can be used to characterize the excess over a threshold as a Poisson process. Hence, $\mathbb{E}[X] = \text{Var}[X]$. The DI statistic introduced by Cunnane [1979] is defined by $DI = s^2 \cdot \lambda - 1$, where s^2 is the intensity of the Poisson process and λ is the mean number of events in a block.

Based on *L*-Moments plot

L-Moments are linear combinations of the ordered data values. From the GPD, we have that

$$\tau_4 = \tau_3 \cdot \frac{1 + 5\tau_3}{5 + \tau_3}, \tag{2.18}$$

where τ_4 is the *L-Kurtosis* and τ_3 is the *L-Skewness*. See e.g. Hosking and Wallis [1997] for more details on L-moments or Peel et al. [2001] for a known application of this method in hydrology.

We can then construct the *L-Moment plot* which consist of the points :

$$\left\{(\hat{\tau}_{3,u}, \hat{\tau}_{4,u}) : u \leq x_{\max}\right\} \tag{2.19}$$

where $\hat{\tau}_{3,u}$ and $\hat{\tau}_{4,u}$ are estimations of L-kurtosis and L-skewness based on u and $x_{\max}$ is the maximum observation. Note that interpretation of this plot is often tedious.

2.5.2 Varying Threshold : Mixture Models

The so-called "fixed threshold" approaches (as renamed in Scarrott and MacDonald [2012], among others) considered so far have been criticized since it leads to a fixed and subjective threshold, sometimes also seen as an arbitrary choice where the uncertainty cannot be taken into account.

Hence, recent models have emerged that allow a dynamic view of the threshold. In *mixture models*, the threshold is either implicitly or explicitly defined as a parameter to be estimated, and in most cases

the uncertainty associated with the threshold choice can be naturally accounted for in the inferences. The model can be presented in a general way :

$$f(x) = (1 - \zeta_u) \cdot b_t(x) + \zeta_u \cdot g(x), \tag{2.20}$$

where $\zeta_u = \Pr\{X > u\}$ is now called the *tail fraction* and is a new parameter of the model, $b_t(x)$ is the density of the *bulk model* (i.e., the data that does not exceed the threshold) and $g(x)$ is the *tail model*, e.g. the GPD density. We ignored parameter dependence for clarity. There is abundant literature on the subject and numerous models have emerged, not necessarily parametric, see e.g. Dey and Yan [2016, chap.3] for the univariate case. The guiding principle in choosing EV mixture models is to combine a flexible bulk model with a reliable tail fit which is robust to the bulk data. These two components are not independent as they share common information about the threshold.

Serious issues arise in these models, for example regarding the discontinuity that often occurs in the density function at the junction between the bulk and the tail model. Alternative models have emerged to force continuity on the density but still, mixture models are often regarded as over-complex with respect to the benefits these models can offer in practice. Research is in great progress but we think improvements have still to be made before yielding straightforward modeling and valuable results.

Chapter 3

Relaxing The Independence Assumption

Contents

In most environmental applications, the independence assumption made in the first chapters is questionable and never completely fulfilled. From hydrological process as stated in Milly et al. [2008] to temperature data where climate warming is a famous issue, such theoretical assumptions that have been previously made are not sustainable in practice. First, Section **3.1** will provide tools that enable EV models to hold in presence of a limited long-range dependence. Section **3.2** will allow EV modeling under nonstationary processes allowing us for instance to study the possible increasing behavior of the maximum temperatures. Whereas Section **3.3** redefined return levels under those less restricted cases, Section **3.4** will introduce a new flexible way to model nonstationary extremes under a redefined Multi Layer Perceptron framework providing inference techniques that minimize chance of overfitting.

This chapter is mostly based on Coles [2001, chap.5-6], Beirlant et al. [2006, chap.10] and Reiss and Thomas [2007, chap.7], and other relevant articles.

3.1 Stationary Extremes

So far, we considered the maximum $X_{(n)} = \max_{1\leq i\leq n} X_i$ being composed of independent random variables only. Now, we are interested by modeling $X^*_{(n)} = \max_{1\leq i\leq n} X^*_i$ where $\{X^*_i\}$ will now denote a *stationary* sequence of n random variables sharing the same marginal df F as the sequence $\{X_i\}$ of independent random variables.

Definition 3.1 (Stationary process). *We say that the sequence $\{X_i\}$ of n random variables is (strongly)* **stationary** *if, for $h \geq 0$ and $n \geq 1$, the distribution of the lagged random vector $(X_{1+h}, \ldots, X_{n+h})$ does not depend on h.* △

It corresponds to physical processes whose stochastic properties are homogeneous but which may be dependent. There exist other formulations of *stationarity* but we stick to this general definition.

This dependence can take many forms and hence we need to relax the independence condition. Let $F_{i_1,\ldots,i_p}(u_1,\ldots,u_p) := \Pr\{X_{i_1} \leq u_1, \ldots, X_{i_p} \leq u_p\}$ denote the joint df of $X_{i_1}, \ldots, X_{i_p}$ for any arbitrary positive integers $(i_1, \ldots, i_p)$.

Definition 3.2 ($D(u_n)$ dependence condition from Leadbetter [1974]). *Let $\{u_n\}$ be a sequence of real numbers. We say that the* $\boldsymbol{D(u_n)}$ **condition** *holds if for any set of integers $i_1 < \cdots < i_p$ and $j_1 < \cdots < j_q$ such that $j_1 - i_p > \ell$, we have that*

$$|F_{i_1,\ldots,i_p,j_1,\ldots,j_q}(u_n,\ldots,u_n;u_n,\ldots,u_n) - F_{i_1,\ldots,i_p}(u_n,\ldots,u_n)\cdot F_{j_1,\ldots,j_q}(u_n,\ldots,u_n)| \leq \beta_{n,\ell}\ , \quad (3.1)$$

where $\beta_{n,\ell}$ is nondecreasing and $\lim_{n\to\infty}\beta_{n,\ell_n} = 0$ for some sequence $\ell_n = o(n)$. △

This condition ensures that, when the sets of variables are separated by a relatively short distance, typically $s_n = o(n)$, the long-range dependence between such events is limited in a sense that is sufficiently close to zero to have no effect on the limit extremal laws. This result is remarkable in the sense that, provided a series has limited long-range dependence at extreme levels (i.e., where $D(u_n)$ condition holds), maxima of stationary series follow the same distributional limit laws as those of independent series.

Theorem 3.1 (Limit distribution of maxima under $D(u_n)$, Leadbetter [1974]). *Let $\{X^*_i\}$ be a stationary sequence of n iid random variables. If there exists sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that $D(u_n)$ condition holds with $u_n = a_n x + b_n$ for every real x, and*

$$Pr\{X^*_{(n)} \leq u_n\} \longrightarrow G^*(x), \qquad n \to \infty, \quad (3.2)$$

*where G^* is a non-degenerate df, then G^* is a member of the GEV family as presented in Theorem **1.1**.*

□

Theorem 3.2 (Leadbetter et al. [1983]). *Let $\{X^*_i\}$ be a stationary sequence and let $\{X_i\}$ be a iid sequence of n random variables. We have under regularity conditions,*

$$\Pr\{a_n^{-1}(X_{(n)} - b_n) \leq x\} \longrightarrow G(x), \qquad n \to \infty,$$

for normalizing sequences $\{a_n > 0\}$ *and* $\{b_n\}$*, where* G *is non-degenrate, if and only if*

$$\Pr\{a_n^{-1}(X^*_{(n)} - b_n) \leq x\} \longrightarrow G^*(x), \qquad n \to \infty,$$

where G^* *is the limiting df coming from a stationary process, defined by*

$$G^*(x) = G^\theta(x), \tag{3.3}$$

for some constant $\theta \in (0, 1]$ *called the* **extremal index**.

□

It is evident from (3.3) that the maximum of a stationary series will have a tendency to decrease.

3.1.1 The extremal index

The *extremal index* is an important indicator quantifying the extent of extremal dependence, that is the degree at which the assumption of independence is violated. From (3.3), it is clear that $\theta = 1$ lead to an independent process, but the converse does not hold. The case $\theta = 0$ will not be considered as it is too "far" from independence and brings problems. Moreover, results of Theorem **3.2** would not hold.

Formally, it can be defined as

$$\theta = \lim_{n\to\infty} \Pr\Big\{ \max(X_2, \ldots, X_{p_n}) \leq u_n \mid X_1 \geq u_n \Big\}, \tag{3.4}$$

in the POT approach, where $p_n = o(n)$ and the sequence u_n is such that $\Pr\{X_{(n)} \leq u_n\}$ converges. Hence, θ can be thought for example as the probability that an exceedance over a high threshold is the final element in a *cluster of exceedances*.

Clusters of exceedances

From (3.4) and in a POT context, extremes have the tendency to occur in clusters whose *mean cluster size* is θ^{-1} at the limit. Equivalently, θ^{-1} can be viewed as the factor with which the mean distance between cluster is increased. This problem of temporal dependence make inference based on the likelihood invalid. We name two methods that can be used to circumvent this issue :

- **Filtering out** an (approximate) independent sequence of threshold exceedances.
- **Declustering** : compute the maximum value in each cluster and then we model these clusters maximums as independent GP random variable. In this approach, we remove temporal dependence but we do not estimate it. However, information is discarded and this could be a substantial loss in meteorological applications, for instance to determine heat or cold waves.

New parameters

When $0 < \theta \leq 1$, we have from Theorem **3.2** that G^* is an EV distribution but with different scale and location parameters than G. If we note by (μ^*, σ^*, ξ^*) the parameters pertaining to G^* and those from

G kept in the usual way, we have the following relationships, when $\xi \neq 0$

$$\mu^* = \mu - \sigma\xi^{-1}(1-\theta^{\xi}), \quad \text{and} \quad \sigma^* = \sigma\theta^{\xi}. \tag{3.5}$$

In the Gumbel case ($\xi = 0$), we simply have $\sigma^* = \sigma$ and $\mu^* = \mu + \log\theta$. The fact that $\xi^* = \xi$ induce that the two distributions G^* and G will have the same form, following Theorem **3.2**.

Return levels

Because of clustering, notion of return levels is more complex and the dependence appear in the definition of return levels for excess models :

$$r_m = u + \sigma\xi^{-1}\Big[(m\zeta_u\theta)^{\xi} - 1\Big]. \tag{3.6}$$

Hence, we see that ignoring dependence will lead to an overestimation of return levels. For example, we have that :

- If $\theta = 1$, then the *100-year-event* has probability 0.368 of not appearing in the next 100 years.
- If $\theta = 0.1$, the event has probability of 0.904 of not appearing in the next 100 years.

Return levels will be redefined in Section **3.3** for the stationary case.

3.1.2 Modelling in Block Maxima

With dependent series, modeling by means of GEV as in Chapter **1** can be used in a similar way since the shape parameter ξ^* will remain invariant. The difference is that the effective number of block maxima $n^* = n\theta$ will be reduced and hence convergence in extremal Theorem **1.1** will be slower. Indeed, approximation is expected to be poorer and this will be exacerbated with increased levels of dependence in the series. Efforts must be made to either try to increase n for example by reducing the blog length, or make sure the model fit is convincing with diagnostic tools presented in Section **1.7**.

3.2 Non-Stationary Extremes

Whereas previous Section relaxed the first "i" of the "iid" assumption made during Chapter **1** and **2** by allowing temporal dependence under stationary process, this section will now tackle the "**id**" part, i.e. assumption that the observations are **i**dentically **d**istributed. The stationarity assumption is not likely to hold for climatological data such as temperatures. For instance, the most obvious departure from stationarity is the presence seasonal patterns as seen in Figure C.2 with higher spread in spring or in autumn for example. Seasonal concerns should disappear for very high thresholds in excess models but this is not a valid argument since the number of data would become very small. For a sufficiently large block size in block maxima, meteorological seasons should not be an issue.

Furthermore, the aim of this thesis will focus on the modeling of the possible trend in order to assess the climate warming. Our modeling will hence more focus on the analysis of different parametrizations for the mean by allowing the location parameter μ to vary with time. Even if it seems less interesting, we

will also allow the scale parameter to vary in order to check if the annual maxima have varying spread over time. We will avoid to vary ξ with time in order to stay in the same EV family of distributions.

There are no new general theory a for nonstationary processes in previous section and we will then use a pragmatic approach of combining standard EV models with statistical modeling. We will only discuss the approach in block maxima but the extrapolation to POT is straightforward for inference, model comparison and diagnostics.

3.2.1 Block-Maxima

As we continue to consider yearly blocks, we do only face nonstationary concerns for the trend which could probably be imputed to the Global Warming. The evidence of seasonality arising when we decrease the block lengths is an issue in block maxima. For example if one consider daily maximum temperatures, seasonality will be present. However, we loose information by taking yearly blocks only as we do not use all the information since at least one half of the daily temperatures from October to March will not be used. To overcome this issue, dataset could be divided and different models can be applied on sub-blocks of the data, for example on the month July and August and fitting a GEV model with block length of 62 days and do similar analysis with other months. This will provide us a set of GEV models that will describe different aspects of the process. In fact, this method would lead to some kind of a "manual" nonstationary GEV modeling.

Inference for GEV

In a nonstationary context, ML is preferred for its adaptability to changes in model structure. In a general setting, we let a nonstationary GEV model describe the distribution Z_t for $t = 1, \dots, m$:

$$Z_t \sim \text{GEV}\big(\mu(t), \sigma(t), \xi(t)\big), \tag{3.7}$$

where each of $\mu(t), \sigma(t), \xi(t)$ are expressed as

$$\theta(t) = b\big(X^{'}\beta\big), \tag{3.8}$$

for a specified inverse link function $b(\,\cdot\,)$ where θ denotes either μ, σ or ξ and where β denote the complete vector of parameters. In our example, Z_t will describe the annual maximum temperature of year t for $m = 116$ years. As already stated, it will not be recommended to allow ξ to vary with time. Examples of parametric expressions from (3.8) will be given in Section **6.2.1**.

If $g\big(z_t; \mu(t), \sigma(t), \xi(t)\big)$ denotes the GEV density (Table 1.1) with parameters $\mu(t), \sigma(t), \xi(t)$ evaluated at z_t, the log-likelihood of the model (3.7) is, provided $\xi(t) \neq 0\ \forall t$,

$$\begin{aligned}\ell(\beta) &= \sum_{t}^{m} \log g\big(z_t; \mu(t), \sigma(t), \xi(t)\big) \\ &= -\sum_{t}^{m} \Bigg\{ \log\sigma(t) + \big[1 + \xi^{-1}(t)\big] \log \Big[1 + \xi(t)\cdot\sigma^{-1}(t)\cdot\Big(z_t - \mu(t)\Big)\Big]_{+} \\ &\qquad + \Big[1 + \xi(t)\cdot\sigma^{-1}(t)\cdot\Big(z_t - \mu(t)\Big)\Big]_{+}^{-\xi^{-1}(t)} \Bigg\},\end{aligned} \tag{3.9}$$

where the notation $y_+(t) = \max\{y(t), 0\}$ holds for all t. The parameters $\mu(t), \sigma(t), \xi(t)$ are replaced by their respective expressions from (3.8). If $\xi(t) = 0$ for any t, we replace the likelihood by using the limit $\xi(t) \to 0$ in (3.9) as in Table 1.1. Numerical techniques are then used to maximize (3.9) in order to yield the MLE of β and evaluate standard errors.

Model Comparisons

In order to compare our models, that is for example to check whether a trend is statistically significant, or if the nonstationnary models provide an improvement over the simpler (stationary) model, we will use two techniques :

1. The *deviance statistic* which is defined as

$$D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\}, \tag{3.10}$$

   for two nested models $\mathcal{M}_0 \subset \mathcal{M}_1$, where $\ell_1(\mathcal{M}_1)$ and $\ell_0(\mathcal{M}_0)$ are the maximized log-likelihoods (3.9) under models $\mathcal{M}_1$ and $\mathcal{M}_0$ respectively. Asymptotically, the distribution of D is χ^2_k with k degrees of freedom representing the difference of parameters between model $\mathcal{M}_1$ and $\mathcal{M}_0$. Comparisons of D with the χ^2_k critical value will guide our decision.

2. It is sometimes preferable to rely on other criterion, for example when the number of models to be compared is large or their construction is not straightforward. We will make use of the *Bayesian Information Criterion* (BIC) and the *Akaike Information Criterion (corrected)* (AIC_c). For n observations and p parameters, we define

$$\text{BIC} = -2\ell + p\log(n), \qquad \text{AIC}_\text{c} = -2\ell + 2p + \frac{2p(p+1)}{n-p-1}. \tag{3.11}$$

   Used by Cannon [2010], these two criterion both have a likelihood term which represent the quality of fit of the model and a term which penalizes the complexity of the model represented by its number of parameters to be estimated. These two criterion are advised for small samples and to prevent overfitting (i.e., fitting to noise instead of the true underlying process). BIC will penalize more heavily models that are more complex.

The basic principle of parsimony holds for both methods. In the first method, it is incorporated in the statistical test since the critical value χ^2_k will increase with k, the difference of complexity of two models. In the second method, it is directly included in the criterion since the partial derivative with p is positive for both criterion(3.11) and these are to be minimized.

We will use the first technique in Section **6.2.1** to make successive comparisons of parametric models that we propose for the location parameter. We will then summarize all the results in Table 6.5 in the following Section by means of BIC and AIC_c, taking all relevant models into account.

Model Diagnostics

When the "best" model is selected following the method, it is still necessary to assess that the model fits well enough the data at hand that we can infer conclusions about some aspect of the population. We will

use tools seen in the independent (or stationary) case, the quantile and the probability plots presented in Appendix **A.4**. From the inhomogeneous distribution across years, it is needed to standardize the data. For example, when model (3.7) is estimated, the *standardized variables* $\tilde{Z}_t$ are

$$\tilde{Z}_t = \hat{\xi}^{-1}(t) \cdot \log\left\{1 + \hat{\xi}(t) \cdot \hat{\sigma}^{-1}(t) \cdot \Big(Z_t - \hat{\mu}(t)\Big)\right\}, \tag{3.12}$$

each having standard Gumbel distribution (1.9). This yield the *residual probability plot* :

$$\left\{\Big(i/(m+1),\ \exp(-e^{-\tilde{z}_{(i)}})\Big)\ :\ i = 1, \dots, m\right\}, \tag{3.13}$$

with the Gumbel as reference, and the *residual quantile plot* :

$$\left\{\Big(\tilde{z}_{(i)},\ -\log(-\log(i/(m+1)))\Big)\ :\ i = 1, \dots, m\right\}. \tag{3.14}$$

The choice of the Gumbel as reference distribution can be discussed but this is a reasonable choice regarding its place in EVT. Figure C.10 in Appendix **C** will present such plots.

3.3 Return Levels : New Definitions

Whereas we already defined return levels in Section **1.5** for independent sequences, we will now give a more general definition for return levels.

Stationarity

Under assumption of a stationary sequence, the return level is the same for all years. The m-year return level r_m is associated with a return period of m years. Let $X_{(n),y}$ denote the annual maximum for a particular year y. Assuming $\{X_{(n),y}\} \overset{iid}{\sim} F$, there are two main interpretations for return periods in this context, following Amir AghaKouchak [2013, chap.4] :

1. **Expected waiting time until an exceedance occurs :** let T be the year of the first exceedance. Recalling $F(r_m) = \Pr\{X_{(n),y} \leq r_m\} = 1 - 1/m$, we write

$$\begin{aligned}
\Pr\{T = t\} &= \Pr\{X_{(n),1} \leq r_m, \dots, X_{(n),t-1} \leq r_m, X_{(n),t} > r_m\} \\
&= \Pr\{X_{(n),1} \leq r_m\} \dots \Pr\{X_{(n),t-1} \leq r_m\} \Pr\{X_{(n),t} > r_m\} && \boxed{\text{iid assumption}} \\
&= \Pr\{X_{(n),1} \leq r_m\}^{t-1} \Pr\{X_{(n),1} > r_m\} && \boxed{\text{stationarity}} \\
&= F^{t-1}(r_m)(1 - F(r_m)) \\
&= (1 - 1/m)^{t-1}(1/m).
\end{aligned}$$

   We easily recognize that T has a geometric distribution with parameter m^{-1}. Hence, its expected value is $1/m^{-1} = m$, showing that the expected waiting time for an m-year event is m years.

2. **Expected number of events in a period of m years is exactly 1 :** to see that, we define

$$N = \sum_{y=1}^{m} \mathbb{I}(X_{(n),y} > r_m)$$

as the random variable representing the number of exceedances in m years, with $\mathbb{I}$ the indicator function. Hence, each year is a "trial", and from the fact that $\{X_{(n),y}\}$ are iid, we can compute the probability that the number of exceedances in m-years is k by

$$\Pr\{N = k\} = \binom{m}{k}(1/m)^k(1 - 1/m)^{m-k},$$

where we retrieve $N \sim \text{Bin}(m, 1/m)$ and hence N has an expected value of $m \cdot m^{-1} = 1$.

Non-stationarity

As demonstrated in Amir AghaKouchak [2013, Section 4.2], we can retrieve the same two interpretations of return period as for a stationary process. However, mathematical derivations go beyond the scope of this thesis. Moreover, from definition of non-stationary process, as parameter(s) will be function of time, return levels will also change over time. This will have big impacts on modeling, since an inappropriate model will lead to inappropriate return levels, as we will see to a certain extent in Figure 6.3. We will now try more complex models in order to improve this fit.

3.4 Neural Networks for Nonstationary Series : GEV-CDN

In the era of Artificial Intelligence and Machine Learning or even the trendy term "Deep Learning", it is interesting to see how artificial Neural Networks (NN) can effectively deal with nonstationarity in EVT for the block-maxima approach. In practice, assumptions made by models (3.8) may not be accurate enough, in the sense that it could not take the complex temporal relationship with GEV parameters. One could for example expect to have particular relationships between the covariates[1] and the GEV parameters. Only considering parametric models in location or scale could thus be seen as too restrictive. For example, Kharin and Zwiers [2005] allowed for nonlinear trends in temperature extremes by making simulations over a 110-year transient global climate to estimate linear trends in the three GEV parameters based on a series of overlapping 51-year time windows.

Here, we follow an automated approach allowing to take into account all possible relationships through a flexible modeling approach. The well-known result from Hornik et al. [1989] says that provided enough data, hidden units and an appropriate optimization, NN's can capture any smooth dependencies of the parameters given the input, and hence, it can theoretically capture any conditional continuous density, be it asymmetric, multimodal, or heavy-tailed. This can be particularly interesting as we do not have particular prior knowledge on the form of the underlying process of annual maximum temperatures. NN's have this facility of being capable of automatically modeling any nonstationary relationships without explicitly specify it a priori, including interactions between covariates. As demonstrated by the reference article of Cannon [2010], physical process such as rainfall or other meteorological data have a tendency to show nonlinearities and so NN's become interesting. However from its flexibility, attention must be given to the danger of overfitting the data. Another pitfall is its

[1] Here we will still consider the time itself only but we could have other time-varying covariates.

lack of interpretation of the relationships retrieved by the model between inputs and outputs but it bears noting that sensitivity analysis methods as in Cannon and McKendry [2002] could be used to identify the form of nonlinear relationships between covariates and GEV distribution parameters or quantiles.

We will use a *conditional density estimation network* (CDN), which is a probabilistic variant of the *multilayer perceptron* (MLP). An extensive review of the MLP can be found for example in Hsieh and Tang [1998] in the context of meteorological or climatological predictions. Parameters will be estimated via generalized maximum likelihood.

Parametric or nonparametric ? It is always a difficult task to state whether Neural Networks (NN) are parametric models or not. NN are somewhere in the gray area between a *parametric* and a *non-parametric* model, in the sense that it assumes a GEV distribution from the output layer defined by the three parameters of interest, while it also allows for a fabulous flexibility coming from the hidden layers which lead to think that these aree rather nonparametric. Note that all transformations applied inside the network are in general parametric and nonlinear. However, this terminological question is not very relevant and this will not impact the particular modeling.

3.4.1 Generalized Maximum Likelihood

In Section **1.6.1** we introduced the concept of penalized likelihood. The idea is that the MLE's may diverge for some values of ξ, especially when sample size is small. To resolve this problem, Martins and Stedinger [2000] suggest the use of a prior distribution for the shape parameter ξ of the GEV model such that the most probable values of the parameter are included. This method extend the usual ML and is called the *Generalized Maximum Likelihood* (GML). In this method, the penalty is in the form of a prior distribution on ξ :

$$\pi(\xi) \sim \text{Beta}(\xi + 0.5;\ c_1, c_2), \tag{3.15}$$

in which ξ is limited to the range $-0.5 \leq \xi \leq 0.5$ to limit the search space of ξ during optimization to the support of the shifted beta prior. It is recommended by the authors to set c_1 and c_2 to 6 and 9 respectively, resulting in a Beta density function with a mode at -0.1 and $\approx 90\%$ of the probability concentrated between -0.3 and 0.1. However, these two values can be tuned depending on the application or relying on results of preceding inferential methods, based on the characteristics of the Beta distribution. For a sequence $\boldsymbol{x} = x_1, \ldots, x_n$ of observations, the GML estimator corresponds to the mode of the empirical posterior distribution, i.e. the *generalized-likelihood*

$$\text{GL}(\mu, \sigma, \xi | \boldsymbol{x}) = L(\mu, \sigma, \xi | \boldsymbol{x}) \cdot \pi(\xi), \tag{3.16}$$

where $L(\cdot)$ can be the log-likelihood $\ell(\cdot)$ (3.9). When dealing with nonstationnary processes El Adlouni et al. [2007] have proven that GML estimators are likely to outperform the usual ML's. This method will be used in the GEV-CDN framework also because it is more flexible and will avoid issues of usual likelihood computations.

3.4.2 Architecture of the GEV-CDN Network

The MLP architecture of the general GEV-CDN framework is pictured in Figure 3.1. Given a set of covariates $\{x_i(t),\ i = 1, \ldots, I\}$ at time t, outputs are evaluated following these steps :

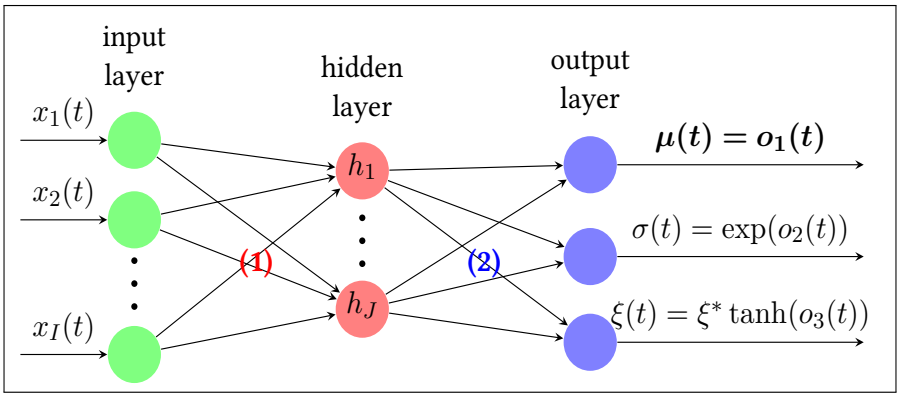

Figure 3.1: *General framework of the fully-connected nonstationary GEV-CDN based on Cannon [2010]. The input layer will still the time itself in our application, i.e. $x_i(t) = t,\ \forall i = 1, \ldots, I$ but it can be other covariates. The hidden layer represent additional complexity incorporated in the model and the output layer represent the three GEV parameters. **(1)** and **(2)** represent the functional relationships (3.17)-(3.18) between layers.*

- The j-th hidden layer node h_j is given by

$$h_j(t) = m\Big(\sum_{i=1}^{I} x_i(t) \cdot w_{ji}^{(1)} + b_j^{(1)}\Big), \tag{3.17}$$

  with $m(\cdot)$ the hidden layer activation function, $w_{ji}^{(1)}$ and $b_j^{(1)}$ are the input-hidden layer weight and bias. The function $m(\cdot)$ is often sigmoidal to allow the GEV-CDN mapping to be nonlinear but it can be the identity function for a strictly linear mapping.

- The value of the k-th output is given by

$$o_k(t) = \sum_{j}^{J} h_j(t) \cdot w_{kj}^{(2)} + b_k^{(2)}, \qquad k = 1, 2, 3, \tag{3.18}$$

  where $w_{kj}^{(2)}$ is the hidden-ouput layer weight and $b_k^{(2)}$ is the hidden-output layer bias.

- The GEV parameters are obtained by applying the output-layer activation functions $g_k(\cdot)$ denoted in Figure 3.1. As usual, the function $g_2(\cdot)$ is to force σ to take positive values and $g_3(\cdot)$ is to constraint ξ to lie within $[-\xi^*,\ \xi^*]$. Again, we notice that it is not recommended to allow ξ to vary with time.

A hierarchy of models can be defined by varying the structure of the CDN (number of hidden layers, which activation function $m(\cdot)$ and weights connections) and compared by the selection criterion such as the BIC or AIC_c discussed in (3.11).

3.4.3 Prevent Overfitting : Bagging

*B*ootstrap *agg*regat*ing* or *bagging* is an ensemble method[2] used in many state-of-the-art machine learning algorithms such as Random Forests discovered by Breiman [2001]. This technique is praised in Machine Learning for its performance as it decreases variance of predictions (or estimates) and hence reduce the risk of overfitting. This method works by generating additional data with repetitions from the original dataset to produce multisets of the same size. The individual multisets' outputs having equal weights are then combined by averaging the ensemble members. This process is also known as *model averaging*. Indeed, by increasing the size of original data you cannot improve the model predictive force, but just decrease its variance.

Carney et al. [2005] have successfully applied this averaging method in the context of CDN and Cannon [2010] says it is worth exploring for GEV-CDN models. He implemented it soon after in its R package GEVcdn. *Early stopping* stopping can be added as a computationally intensive means of controlling overfitting as it stops training prior to convergence of the optimization algorithm and hence allow to reduce the complexity of the model.

Other techniques are available to prevent overfitting. First used by MacKay [1992], *weight penalty regularization* is popular and available in the GEV-CDN framework as a means of limiting the effective number of parameters. The amount of weight penalty is controlled via a Gaussian prior on the magnitude of the input-hidden layer weights. Optimal value for the variance of the Gaussian prior must be set relying on some form of split-sample or cross validation scheme.

3.4.4 Confidence Intervals : Bootstrapping Methods

Like the estimated parameters themselves, the standard errors may not be reliable for small samples. One way to improve the accuracy of the standard errors is to use *bootstrap*. Discovered by Efron [1979], this can be used for EV dfs including nonstationarity. The bootstrap samples are manufactured through Monte Carlo resampling of residuals to attend to the underlying assumption that original sample is iid. There exist a large panel of different bootstrap procedures to construct confidence intervals. For example, Khaliq et al. [2006] follows these steps for the residual bootstrap :

1. Fit a nonstationary GEV model to the data.
2. Transform residuals from the fitted model to be identically distributed :

$$\varepsilon(t) = \Big[1 + \xi(t) \cdot \sigma^{-1}(t) \cdot \Big(x(t) - \mu(t)\Big)\Big]^{-\xi^{-1}(t)},$$

   where $\varepsilon(t)$ is the t-th trandformed residual.
3. Resample $\varepsilon(t)$ with replacement to form the bootstrapped set $\Big\{\varepsilon^{(b)}(t),\ t = 1, \ldots, n\Big\}$.
4. Rescale the bootstrapped residuals by inverting the transformation :

$$x^{(b)}(t) = \mu(t) - \sigma(t) \cdot \xi^{-1}(t) \cdot \Big(\varepsilon^{(b)}(t) - 1\Big). \tag{3.19}$$

[2] For climatologists, *ensemble models* are different but of major utility, especially to make weather forecasting, see for example Suh et al. [2012] among others.

5. Fit a new nonstationary GEV model to the boostrapped samples (3.19) and estimate the parameters and quantiles from the fitted model.

6. Repeat steps 1 to 5 a large number of times B.

Cannon [2010] found that *parametric* bootstrap outperformed the *residual* bootstrap in the GEV-CDN framework, but he did not consider alternative bootstrap approaches such as the *bias-adjusted percentile* which might yield better calibrated confidence intervals. Empirical Monte-Carlo comparisons of coverage from all available methods considered so far would be interesting. Furthermore, the incoming chapter will introduce other methods to construct intervals in a strictly Bayesian approach relying also on Monte-Carlo sampling.

Chapter 4

Bayesian Extreme Value Theory

Contents

see evdbayes pdf package r

Attention : π ou f ??????

We let useful relevant tools regarding bayesian inference in appendix **B**

This chapter is mostly based on Coles [2001, sec. 9.1], Beirlant et al. [2006, chap.11] and Reiss and Thomas [2007, chap.1-4], and other relevant articles.

Definition 4.1 (Posterior distribution). *Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ denote the observed data of a random variable X distributed according to a distribution with density*

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta) \cdot L(\theta|\boldsymbol{x})}{\int_\Theta \pi(\theta) \cdot L(\theta|\boldsymbol{x}) \cdot d\theta} \propto \pi(\theta) \cdot L(\theta|\boldsymbol{x}) \tag{4.1}$$

where $L(\cdot)$ denotes the likelihood function, as in ?? but there it was the log-likelhiood !!! and θ usually denotes the multidimensional set of parameters in EVT, $\theta = (\mu, \sigma, \xi)$, at least in a univariate stationary context. △

Unlike other chapters, we will directly write θ as a 3D vector, unless stated differently. This allows not to overload the equations and to facilitate readability.

1. Whenever it is possible, it allows to introduce to introduce other source of knowledge coming from the domain at-hand, by the elicitation of a prior. The counter-argument of this advantage is that it also introduces (improper ?) subjectiveness.

2. "account- ing for parameter and threshold uncertainty is perhaps handled most easily in the Bayesian paradigm" [Dey and Yan, 2016, pp.106]

   As such, It permits an elegant way of making future predictions which is one of the most(?) important issue in EVT.

3. Bayesian framework can overcome the regularity conditions of the likelihood inference (see section **4.1**). Thus it usually provides a viable alternative in cases when MLE (for example) breaks down. And actually, we are not so fo far from the problematic situations depicted in section **3.1**. Moreover, the Highest Posterior Probability (HPD) region is constructed so that... and there is no more need to fall to asymptotic theory as in conventional methods.

4. For an asymmetric distribution, the HPD interval can be a more reasonable summary than the central probability interval (see illustration ...). For symmetric densities, HPD and central interals are the same while HPD is shorter for asymmetric densities. See Liu et al. [2015]....

As the dependence becomes stronger, the run length n must be larger in order to achieve the same precision. Dependence exists both within the output for a single parameter (autocorrelations) and across parameters (cross-correlations), we discuss this issue in section text.

4.1 Prior Elicitation

Sometimes viewed as advantage from the amount of information that can be retrieved, and sometimes viewed as an drawback due to the (rather unquantifiable) subjectivity that introduced, the construction of the prior is a key step in Bayesian analysis.

Priors are necessary in the Bayesian paradigm to be able to compute the posterior in (4.1).But, priors require the legitimate statement of domain's expert, to make this viewed the less subjective as possible

Prior may not be of great importance if the size (m) of the dataset is large. It can be seen from (4.1) where the amount of information contained in the data through $L(\theta|\boldsymbol{x})$ will be prominent compared to this contained in the prior through $\pi(\theta)$. Prior will have limited influence.

One is aware that this is not often the case in EVT cases. By design, we are dealing with rather small so-constructed datasets. And mostly for this reason, it could be important to incorporate additional information in this limited dataset through the prior distribution.

4.1.1 Non-informative Priors

Receive a correct and accepted advice from an expert is often difficult So, in many cases, we cannot inject information through the prior. We must then construct a prior which represent this lack of knowledge so that they do not influence posterior inferences.

There exists a vast amount of uninformative priors in the literature (see e.g. Yang and Berger [1996], Ni and Sun [2003]) This family of priors can be *improper*, i.e. priors for which the integral of $\pi(\theta)$ over the parameter space is not finite. It is valid to use improper priors only if the posterior target is proper.

Adjustments of these priors must always be thought in practical applications

Jeffrey's prior

is specified as

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \qquad \text{where} \qquad I_{ij}(\theta) = \mathbb{E}_\theta \left[- \frac{\partial^2 \log f(\boldsymbol{x}|\theta)}{\partial\theta_i \partial\theta_j} \right], \quad i,j = 1,\dots,d. \tag{4.2}$$

where $f(\boldsymbol{x}|\theta)$ is the density function of X. This prior is invariant to reparametrization, but has complex form for EV models, and it exists only when $\xi > -0.5$ in GEV models, where it is function of ξ and σ only.

MDI prior

Maximal Data Information priors

However, it has been showed by Northrop and Attalides [2016] that both Jeffrey and MDI priors give improper posterior when there are no truncation of the shape parameter, i.e. we must restrict the fact that $\pi(\theta) \to \infty$ as $\xi \to (-)\infty$ for Jeffreys (MDI), in order to obtain a proper posterior.

Vague priors

The last and often preferred alternative to construct uninformative priors is to use proper priors which are near flat, e.g. which are uniform or with exhibits large variance for the normal distribution.

In GEV we will take independent normal-distributed priors each with a large (tuned) variance. When these variances increase, we get at the limit

$$\pi(\theta) = \pi(\mu,\nu,\xi) \overset{(\perp\!\!\!\perp)}{=} \pi(\mu)\cdot\pi(\nu)\cdot\pi(\xi) \ \propto 1, \tag{4.3}$$

where $\nu = \log\sigma$.

Taking multivariate normal distribution as prior has also been proposed (see) is often difficult as it involves 9 (hyper)parameters in total and this can be difficult

4.1.2 Informative Priors

STAN : "It can also be a huge help with computation to have less diffuse priors, even if they're not informative enough to have a noticeable impact on the posterior. "

Gamma Distributions for Quantile Differences

Beta Distributions for Probability Ratios

The Bayes Factor

4.2 Bayesian Computation : Markov Chains

Methods have been developed for sampling from arbitrary posterior distributions $\pi(\theta|\boldsymbol{x})$. Simulations of N values $\theta_1, \theta_2, \ldots, \theta_N$ that are iid from $\pi(\theta|\boldsymbol{x})$ can be used to estimate features of interest.

But simulating from $\pi(\theta|\boldsymbol{x})$ is usually not achievable and this is why we need **Markov Chain Monte Carlo** (MCMC) techniques. We use it to simulate a markov chain $\theta_1, \theta_2, \ldots, \theta_N$ that conerge to the target distribution $\pi(\theta|\boldsymbol{x})$. This means that, after some *burn-in period* B, $\theta_{B+1}, \ldots, \theta_N$ can be treated as random sample from $\pi(\theta|\boldsymbol{x})$.

Let's now (a bit weakly) define one of the most important results in Markov Chain theory.

Definition 4.2 (*First-order discrete-time* Markov Property)**.** *Let $k_0, k_1, \ldots$ be the states associated to a sequence of time-homogeneous random variables, say $\{\theta_t : t \in \mathbb{N}\}$. The Markov property states that the distribution of the future state θ_{t+1} depends only on the distribution of the current state θ_t. In other words, given θ_t, we have that θ_{t+1} is independent of all the states prior to t. We can write this as*

$$\Pr\{\theta_{t+1} = k_{t+1} \mid \theta_t = k_t,\ \theta_{t-1} = k_{t-1}, \ldots\} = \Pr\{\theta_{t+1} = k_{t+1} \mid \theta_t = k_t\}. \tag{4.4}$$

△

or see Angelino et al. [2016, section 2.2.3] for more in-depth results.

The samples are not independent, and the dependence influences the accuracy of the posterior estimates. As dependence becomes stronger, we must increase the run-length N to achieve the same accuracy.

4.2.1 Algorithms

We are looking for a so-generated chain that has a stationary distribution $\pi(\theta|\boldsymbol{x})$. This is the case if the chain is

1. *aperiodic*

2. *irreducible* or *ergodic*, that is if any state for θ can be reached with probability > 0 in a finite number of steps from any other state for θ.

"The Markov chains Stan and other MCMC samplers generate are *ergodic* in the sense required by the Markov chain central limit theorem, meaning roughly that there is a reasonable chance of reaching one value of theta from another." **?**

With MH or Gibbs sampler, we need to tune individually the proposal standard deviations to reach a correct acceptance, and this is often done with trial-and-error methodology.

The performance of the standard Markov chain Monte Carlo estimators depends on how effectively the Markov transition guides the Markov chain along the neighborhoods of high probability. If the exploration is slow then the estimators will become computationally inefficient, and if the exploration is incomplete then the estimators will become biased Betancourt [2016]. It is then necessary to consider other form of sampling...

4.2.2 Hamiltonian Monte Carlo

Package Rstan

Neal and others [2011] and Betancourt and Girolami [2015] are really

HMC permit to better exploit the properties of the target distribution to make informed jumps through neighborhoods of high probability while avoiding neighborhoods of low probability entirely.

4.2.3 Computational efficiency comparison

In modern statistical area, computing methods have been widely ... And this need for compiutations will rise in the future.

We will then compare our 3 methods too see if effectively

4.3 Convergence Diagnostics

When applying MCMC algorithms to estimate posterior distributions, it is vital to assess convergence of the algorithm to try to ensure that we reached the stationary target distribution. Let's now enumerate some of the key steps we must keep in mind when thinking about convergence, an hence reliable results.

1. A sufficient *burn-in period* $B < N$ must be chosen to ensure that the convergence to the posterior distribution $\pi(\theta|\boldsymbol{x})$ has occurred.

2. For the same reason, a sufficient number of simulations N to eliminate the influence of initial conditions and ensure accuracy in the estimations ((and then make sure than we are sampling from the target stationary (posterior) distribution)).

3. Several dispersed starting values must have been simulated to ensure we explored all the regions of high probability. This is particularly important when the target distribution is complex.

4. The chains must have good mixing properties, in the sense that the whole parameter space (...) A common technique that we will apply is to run different chains several times and then combine

a proportion of each chain (typically 50%) to get the final chain. This procedure wants to ensure a proper mixing behaviour. The potential scale reduction factor (Gelman diagnostic) is also a popular tool, see .

We must keep in mind that no convergence diagnostics can prove that convergence really happened and validate the "model". However, a combined use of several relevant diagnostics will be required to increase our confidence that convergence actually happened.

4.3.1 Proposal Distribution

The main ideas are :

- If the variance of the proposal distribution is too large, most proposals will be rejected :

  ie the jumps through the chain are too large,

- If the variance of the proposal distribution is too low, then most proposals will be accepted

Both are harmful for the objective of an efficient "visit" of the whole parameter space.

Widely speaking, we consider 2 different types of algorithms in which it is preferable to target a certain acceptance rate. It is distinguished by the updating manner of the components of θ through the algorithm, i.e. the 3 univariate parameters of interest.

- When all components of θ are updated simultaneously, it is recommended to target an acceptance rate of around 0.20. Roberts et al. [1997] have shown that, under quite general conditions, the asymptotically optimal acceptance rate is 0.234. (for target density that has a symmetric product form) This quantity has been verified by Sherlock et al. [2009]. It holds for the *Metropolis-Hastings* algorithm.

- When the components are updated one at a time, an acceptance rate of around 0.40 is recommended. It holds for the *Gibbs sampler* algorithm.

Let's (see **?** for example for the first case)

Gelman-Rubin diagnostic : the $\hat{R}$ statistic

As discussed in item **4** above

Geweke diagnostic

Thinning

iteration k is stored only if $k \mod$ thin is zero (and if k greater than or equal to the burn-in B).

This typically reduces the precision of posterior estimates, but it may represent a necessary computational saving.

4.3.2 The problem of auto and cross-correlations in the chains

There exists exists 2 problems of correlations in the output delivered by a MC.

- **Autocorrelation** is the
- **Cross-correlation**

4.4 Posterior Predictive

notation for posterior ? pi or f

As discussed in item 2 above, prediction is of important interest in EVT, and this is "facilitated" in the Bayesian paradigm. This also permits a more straightforward quantification of the inferential uncertainty associated.

Definition 4.3 (Posterior Predictive density)**.** *Let X_{m+1} denotes a (one-step-ahead) future observation with density $f(x_{m+1}|\boldsymbol{\theta})$. Then we define the Posterior Predictive density of a future observation X_{m+1} given $\boldsymbol{x}$ as*

$$\begin{aligned} f(x_{m+1}|\boldsymbol{x}) &= \int_{\Theta} f(x_{m+1}, \theta|\boldsymbol{x}) \cdot d\theta = \int_{\Theta} f(x_{m+1}|\theta) \cdot \pi(\theta|\boldsymbol{x}) \cdot d\theta \\ &:= \mathbb{E}_{\theta|\boldsymbol{x}}\big[f(x_{m+1}|\theta)\big] \end{aligned} \tag{4.5}$$

where the last line emphasizes that we can evaluate $f(x_{m+1}|\boldsymbol{x})$ by averaging over the different possible parameter values.

△

The uncertainty in the model is reflected here through $\pi(\theta|\boldsymbol{x})$ while the uncertainty due to variability in future observations is also reflected through $f(x_{m+1}|\theta)$.

Definition 4.4 (Posterior Predictive probability)**.** *The posterior predictive probability of X_{m+1} exceeding some threshold x is accordingly given by*

$$\begin{aligned} \Pr\{X_{m+1} > x \mid \boldsymbol{x}\} &= \int_{\Theta} \Pr\{X_{m+1} > x \mid \theta\} \cdot \pi(\theta|\boldsymbol{x}) \cdot d\theta \\ &= \mathbb{E}_{\theta|\boldsymbol{x}}\big[\Pr(X_{m+1} > x \mid \theta)\big] \end{aligned} \tag{4.6}$$

△

This quantity is often of interest in EVT as we are rather concerned with the probability of future unknown observable exceeding some threshold.

However, this quantity is difficult to obtain analytically. Hence, we will more rely on simulated approximations. Given a sample $\theta_1, \ldots, \theta_r$ from the posterior $\pi(\theta|\boldsymbol{x})$, we use

$$\Pr\{X_{m+1} > x \mid \boldsymbol{x}\} \approx r^{-1} \sum_{i=1}^{r} \Pr\{X_{m+1} > x \mid \theta_i\}, \tag{4.7}$$

where $\Pr\{X_{m+1} > x \mid \theta_i\}$ follows directly from $f(x|\theta)$.

We will now analyse more in-depth the numerical computations in the Bayesian paradigm or how we can get numerically a sample of the posterior distribution.

4.5 Bayesian Predictive Accuracy for Model Validation

4.5.1 Cross-validation for predictive accuracy

When having large amount of data, we can use a well-known and widely used technique coming from Machine Learning. That is, dividing the dataset between a training (typically 75% of the whole set) and a test set containing the remaining observations. For example, having N draws $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ coming from the posterior $\pi(\theta|x_{train})$, we can score each value using (?)

$$\log\left[N^{-1}\sum_{t(i)?=1}^{N} f(x^*|\theta^{(t)})\right]. \tag{4.8}$$

However, we often do not have large amounts of data. Henceforth, we can use the *cross-validation* technique which is more relevant in smaller dataset, but which is computationally more demanding. There exists several variants of them.

Leave-one-out cross-validation

The *Leave-One-Out* (LOO) cross-validation is the

K-fold cross-validation

Vehtari et al. [2016]

Or we can use other criteria which avoid the computations. The basic approach is to use $\log f(x|\bar{\theta}) - p^*$ for N draws $\theta^{(1)}, \ldots, \theta^{(N)}$ from $\pi(\theta|x)$. where p^* represents the effective number of parameters and $\bar{\theta}$ the posterior mean. Several methods exists using this idea. We will see the two most important.

Deviance Information Criterion

The *Deviance Information Criterion* (DIC) was first used by Spiegelhalter et al. [2002] and use the following estimate for the effective number of parameters

$$p^* = 2 \cdot \left(\log f(x|\bar{\theta}) - N^{-1}\sum_{t=1}^{N} \log f(x|\theta^{(t)})\right) \tag{4.9}$$

It is defined on the deviance scale and smaller DIC values indicate better models.

$$\text{DIC} = 2\log f(x|\bar{\theta}) - \frac{4}{N}\sum_{t=1}^{N} \log f(x|\theta(t)) \tag{4.10}$$

Widely Applicable Information Criterion

The *Widely Applicable Information Criterion* (WAIC) is a more recent approach proposed by Watanabe [2010] and is given by

$$\text{WAIC} = 2\sum_{i=1}^{n}\left[\log\left\{\mathbb{E}_{\theta|x} f(x_i|\theta)\right\}\right] - \mathbb{E}_{\theta|x}\log f(x_i|\theta) \tag{4.11}$$

or

$$\text{WAIC} = \sum_{i=1}^{n}\left[2\log\left(N^{-1}\sum_{t=1}^{N} f(x_i|\theta^{(t)})\right) - \frac{4}{N}\sum_{t=1}^{N}\log f(x_i|\theta^{(t)})\right] \tag{4.12}$$

There exists for sure other several methods, as proposed by Gelman et al. [2014].

'LOO and WAIC have various advantages over simpler estimates of predictive error such asAIC and DIC but are less used in practice because they involve additional computational steps'

For each generated chains with dispersed starting values, we evaluate separately the information criteria. The discrepancies between the chains are small (?), which is a good sign.

4.6 Bayesien Inference ?

4.6.1 Bayesian Credible Intervals

The Bayesian *credible intervals* are inherently different from the frequentist's confidence intervals. In the Bayesian intervals, the bounds are treated as fixed and the estimated parameter as a random variable, while in the frequentist's setting, bound are random variables and the parameter is a fixed value.

There exist mainly two kinds of credible interval in the Bayesian sphere :

- The *Highest Probability Interval* (HPD) which is defined as the shortest interval containing $x\%$ of the posterior probability, e.g. if we want a 95% HPD interval (ξ_0, ξ_1) for ξ :

$$\int_{\xi_0}^{\xi_1} \pi(\xi|\boldsymbol{x})d\xi = 0.95 \qquad \text{with} \qquad \pi(\xi_0|\boldsymbol{x}) = \pi(\xi_1|\boldsymbol{x}). \tag{4.13}$$

  It is often the preferred interval as it gives the parameter's values having the highest posterior probability.

- The Quantile-based credible intervals or *equal-tailed interval* picks an interval which ensures a probability of being below this interval as likely as of being above it. For some posterior distribution which are not symmetric, this could be misleading, thus it is not the most recommended interval. (see ..) However, these are often easily obtained when we have a random smaple of the posterior...(?)

xxxxx

4.6.2 Distribution of Quantiles : Return Levels

The Markov chains generated can be transformed to estimate quantities of interest such as quantiles and hence return levels.

The values can be retrieved in the same manner as we have done in the GEV frequentist setting in (1.32). If the df F associated is GEV then $y_m = -\log(1 - m^{-1})$, and if F is GPD then $y_m = m^{-1}$.

r_m is the quantile corresponding to the upper tail probability $p = m^{-1}$.

We can use the values of the samples generated by the posterior to estimate features of this distribution. (... see edbayes)

4.7 Bayesian Model Averaging

4.8 Bayesian Neural Networks

Part II

Experimental Framework : Extreme Value Analysis of Maximum Temperatures

Chapter 5

Introduction to the Analysis

Contents

Since we have theoretically defined useful concepts in EVT, we will now introduce the practical analysis of this thesis that consist of analyzing daily maximum temperatures in Uccle from 1901 to 2016. We already showed the distribution of these data in Figure 2.1 to introduce the POT approach. This chapter will present an introductory analysis of the annual maxima of these data. We chose annual maxima in order to later provide a convenient GEV modeling. We will then first use several techniques to analyze these data before going further into the EV models in the following chapters.

After a brief presentation of the repository which contains the R package created for this thesis and the structure for the scripts that contains the code that created all the analysis, we will briefly present the shiny applications created to enhance visualizations. Then, Section **5.1** will present data and their source(s). Section **5.2** will first describe data and compare models to represent the data before going further on the trend analysis with splines derivatives in a Generalized Additive Model to assess significance of the trend with correction for simultaneous tests and provide first aspects in the issue of Climate Warming without using extreme models.

This chapter will then not explicitly use techniques presented in the previous chapters. It will be mostly based on Ruppert et al. [2003] for the model presented in Section **5.2.3**.

Repository for the code : R Package

The created **R package** can be easily downloaded from its **repository** :

https://github.com/proto4426/PissoortThesis

by following instructions in the README. It follows the standard structure of an usual R package (see e.g. Leisch [2008]) and makes use of the `roxygen2` package for the documentation.

For your convenience, an external folder has been created in this repository containing all the scripts created during this thesis. It allows to reproduce all the results often with more details, tables or plots. It is located in the **/Scripts-R/** folder of the repository. We will notice in the beginning of each of the following chapter to which script(s) it corresponds. We give more details on the structure of the repository in Appendix **D** to give you a better idea about each folders and files.

Visualization Tool : Shiny Application

Shiny applications that have been developed can be run directly through R environment after having the package loaded in your environment, by executing

```
runExample()   # in the R console.
```

Then, choose between the propositions displayed and write it inside (' '). So far, we phave

- ('GEV_distributions') : application based on Figure 1.1 where you can smoothly visualize the GEV and the influence of its parameters.
- ('trend_models') : application dedicated to annual maxima that can you can smoothly visualize with some preliminary methods (see Section **5.2.2**).
- ('splines_draws') : application to simulate splines of GAM model that allow to visualize coverages of the pointwise and simultaneous confidence intervals. (see Section **5.2.3**).

Moreover, a dynamic overview of the applications is available in the repository's `readme`.

All-in-one : Dashboard For the reader's convenience, we decided to smoothly gather all applications made into a *dahsboard*. Moreover, we put this on a server so that, if you do not care about installing the package or looking at the functions, you just have to go to the following URL[1]

The rest of the analysis in this chapter will rely on `1intro_stationary.R` and `1intro_trends(splines).R` codes from the **/Scripts-R/** folder of the repository.

[1] https://proto4426.shinyapps.io/All_dashboard/

5.1 Presentation of the Analysis : Temperatures from Uccle

Data used during this thesis come from the "Institut Royal de Météorologie" (IRM) located in Uccle. We were provided official data used by meteorologists in Belgium. We wanted to have data that are reliable enough in order to be able to produce relevant results.

We were provided a set of databases from the IRM. We decided to focus on temperature analysis in order to assess climate warming. Rainfall data analysis would have also been interesting in EVT.

The fact that the dataset starts only at year 1901 is for reasons of homogeneity since the evolution of the shelters that occurred since previous period. Technology that were used to measure temperatures also lead to several types of measurements errors.

For meteorological considerations and again for reasons of homogeneity, it is better to analyze the temperatures in *closed shelters*, following for example Lindsey and Newman [1956]. Indeed, thanks to grateful advices given by C.Tricot, climatologist at the IRM, temperatures in closed shelters have the advantage of not being too influenced by the solar radiation which artificially increase temperatures, especially in periods of *extreme* heats and thus high solar activity. This is actually what we want to study. For example, the maximum temperature of $36.6°c$ that occurred the 27th June 1947 in closed shelters was actually measured at $38.8°c$ in open shelters the same day.

Comparisons with freely available data

A similar dataset for Uccle is publicly available on the Internet[2]. It was a project initially performed by the KMNI and which was used by Beirlant et al. [2006]. However, we did not want to simply analyze these data as we did not know if we could really trust them, for reasons mentioned above.

Afterwards, we compared these two datasets for years 1901 to 2001. We remark effectively that there are large differences in these two datasets. For example, 54% of measurements are equal with those of the dataset in open shelters against 14.4% in the closed shelters. For this reason, we have confidence that the public dataset have temperatures measured in open shelters which is not recommended. Hence, we conclude that large errors of measures can easily occur in unofficial data which confirms the fact that it is important to get reliable data if one wants to provide reliable analysis.

5.2 First Analysis : Annual Maxima

5.2.1 Descriptive Analysis

There are plenty of very interesting descriptive analysis we could make, but most of them are not very useful. The idea here is just to have a global overview of the data. Code provide most of the descriptive analysis. To summarize most of the information, we plotted a violin-plot and a density plot of the **daily** maximum temperatures in Figure C.2 in Appendix **C** where we divided data in each meteorological seasons to emphasize this difference. We see for example that the the spread is higher for autumn and spring.

Now, we will rather stick on the block-maxima approach by taking yearly blocks leaving us with $n = 116$ data which seem justifiable for further GEV analysis and the conditions to hold. We will

[2] http://lstat.kuleuven.be/Wiley/Data/ecad00045TX.txt

discuss the choice of the block length further in the next Chapter.

5.2.2 First visualization with simple models

We present the series of yearly maxima in Figure 5.1 where we introduce 3 **models for the trend** :

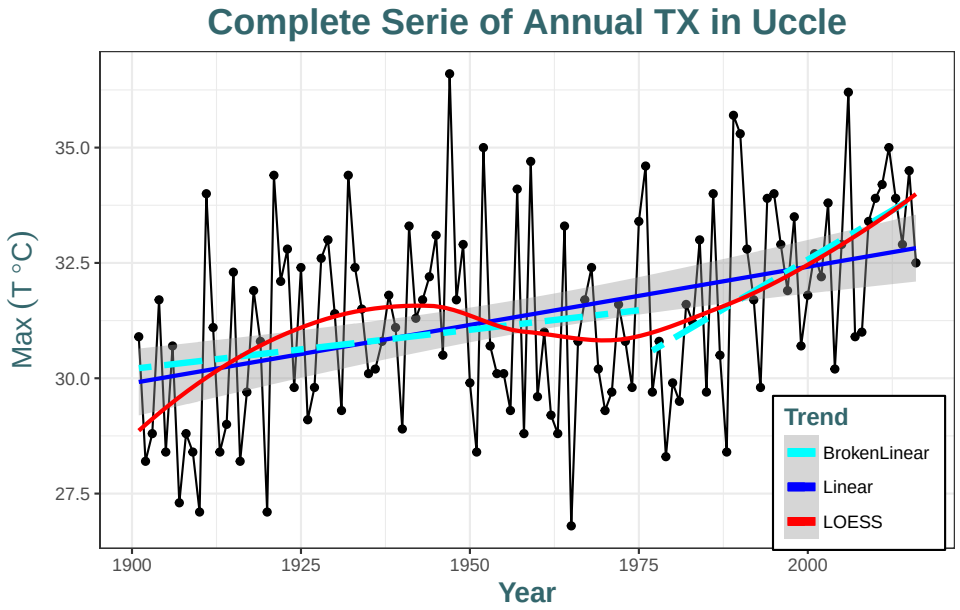

Figure 5.1: *Yearly maxima together with three first models that represent the trend. Note that shaded grey area around the linear trend represent its* 95% *pointwise confidence interval on the fitted values.*

- *Linear regression* is a **parametric** fit. We remark that it is slightly but significantly increasing over time (p-value $\approx 10^{-5}$). From this model, one rough interpretation we can have is that each year we expect the annual maximum temperature to increase by $\hat{b} = 0.025°c$ ($\hat{\sigma}_{\hat{b}} = 0.005$).

- *Local Polynomial regression* or LOESS is a **nonparametric** fit. It tells us that the yearly maxima process is rather "smooth". The drop in the series visible around years 1950 to 1975 is probably due to noise rather than a real decrease or freezing of the maximum temperatures at this time. Moreover, it disappears if we change the parameter controlling the degree of smoothing. We will assess that more formally in the next Section with another nonparametric model.

- *Broken-linear regression* is a **parametric** fit. We wanted to emphasize visually the difference in trend between the period [1901-1975] and [1976-2016]. These two periods have been chosen arbitrarily and we remark that period [1901-1952] has also a very high positive slope. We will study that in more details in next Section.

These different models give us insights on the process we will study throughout the rest of this thesis. We will not develop further the results of each model as this is not the subject of this thesis and we are more interested to assess this trend and its statistical significance overt time with a more interesting method.

5.2.3 Deeper Trend Analysis : Splines derivatives in GAM

Apart from the significant pointwise increasing linear trend, we did not find concrete results. Moreover, we see in the series in Figure 5.1 in that a linear trend to the entire series could be seen as too restrictive. Regarding the broken-linear trend for example, we would like to assess if there is indeed a difference in the slope of the trend over some time periods.

We assume the reader knows about *Generalized Additive Models* (GAM) developed by Hastie and Tibshirani [1986]. We also assume the reader knows about *penalized splines* and *splines smoothing*. Theoretical explanations of these concepts can be found in Ruppert et al. [2003, chapter 3, 6 and 11] which will be the reference book of this section. Replacing covariates with smooth functions such as smoothing splines will result in a more flexible nonparametric model. We will also follow the simulation-based Bayesian approach of Marra and Wood [2012] that compare coverage properties of intervals.

Pointwise or Simultaneous confidence intervals ?

For this analysis, we thought important to differentiate between the two types of intervals in order to better understand what is truly meant by "confidence interval". A special example is the grey area around the linear fit in Figure 5.1 which represent a 95% interval for the regression line (and not for the data), taking into account variability in the data. If we make repeated sampling over and over and take the predicted values, then the new linear fit will be in the grey zone approximately 95% of the time.

Let $\mathcal{X}$ denote the set of x values of interest, i.e. $\mathcal{X} = [1901, 2016]$ in our case and $f(\cdot)$ is the model of interest, e.g. the linear regression, or the GAM model that we will fit. Hence, we define

- A **pointwise** $100(1-\alpha)\%$ confidence interval $\Big\{[L(x), U(x)] \ : \ x \in \mathcal{X}\Big\}$ approximately satisfies

$$\Pr\Big\{L(x) \leq f(x) \leq U(x)\Big\} \geq 1-\alpha, \qquad \forall x \in \mathcal{X}. \tag{5.1}$$

- A **Simultaneous** $100(1-\alpha)\%$ confidence interval must satisfy

$$\Pr\Big\{L(x) \leq f(x) \leq U(x), \ \forall x \in \mathcal{X}\Big\} \geq 1-\alpha. \tag{5.2}$$

From the pointwise confidence intervals we can say that (example) $f(1980)$ has 95% chance to lie within (-1,0) (say) and $f(2000)$ has also 95% to lie within (0.2,1.2) BUT it is a fallacy to say simultaneously that both are contained in these intervals at the same time with 95% confidence.

For the interested reader, theoretical details on how to mathematically derive simultaneous intervals are available in Ruppert et al. [2003, pp.142-144].

Methodology

- We fitted a simple GAM model on the annual maxima relying on the `mgcv` package from Maindonald and Braun [2006].
- Figure C.3 let in Appendix **C** shows the correlation structure of the serial normalized residuals. As selection of a model for the residuals is not trivial from this figure, we fitted several time

series models for the residuals. It is not necessary to consider too complex models so we stopped at 2 additional degrees of freedom. Results are let in Table C.1 in Appendix **C** where we see that BIC will prefer the independent model for the residuals but the AIC will not. Parsimony being important, we chose the independent model to draw plots with simultaneous intervals but we will also keep MA(1) for comparisons. Likelihood ratio tests confirmed our choice. Diagnostics of the model presented in Figure C.4 in Appendix **C** are fine.

- As we took a Gaussian (identity) link, our model can hence be written as

$$Y_{\text{GAM}}(\text{year}) = \alpha + f_{(\text{k})}(\text{year}) + \epsilon, \qquad \epsilon \sim \text{WN}. \tag{5.3}$$

where f is modelled by smoothing splines, ϵ can be MA(1) and Y represent annual TX. We set the dimension of the basis k for the spline to 20 to ensure a reasonable degree of smoothness as there is modest amount of non-linearity in the series and we did not perform cross-validation as it will not influence the results that much.

- To account for uncertainty, we simulated $M = 10^4$ draws (quite fast) from the posterior of the GAM model in (5.3). Hence, the confidence intervals (or bands) that we compute will be of the form of Bayesian credible intervals, discussed in Section **4.6.1** and which is highlighted by Marra and Wood [2012]. 50 posterior draws are displayed in Figure 5.2 displaying pointwise (in yellow) and simultaneous (in red) intervals. We point out that pointwise intervals are too narrow.

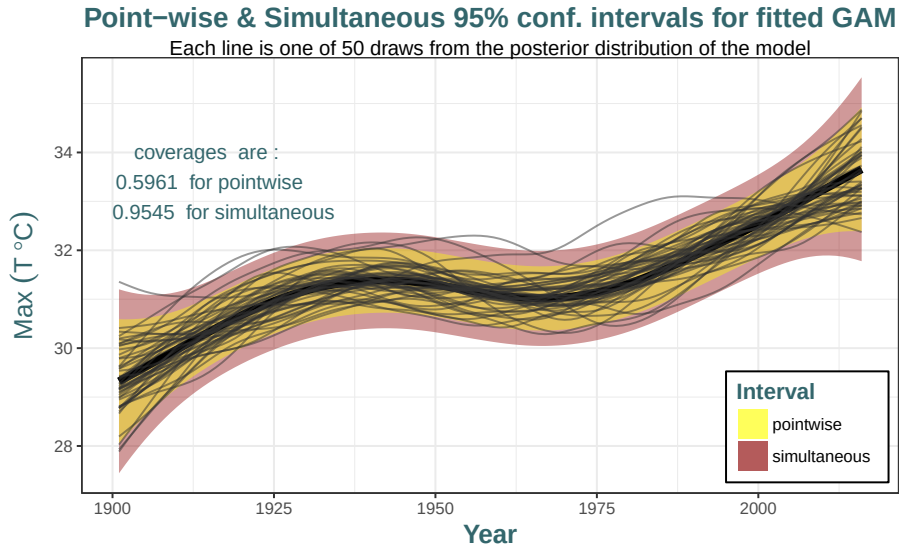

Figure 5.2: *displays draws from the posterior distribution of the model. Notice the large uncertainty associated to the posterior draws (grey lines). The coverages are calculated for $M = 10^4$ simulations.*

Coverage Analysis

We would like to highlight the inadequacy of the pointwise intervals. Hence, we computed posterior draws of the fitted GAM and we counted how many draws lied within each intervals.

Table 5.1: *Proportion of the M posterior simulations which are covered by the confidence intervals*

Coverage at 95%	$M = 20$	$M = 100$	$M = 10^3$	$M = 10^5$
Pointwise	40%	63%	61.1%	59.463%
Simultaneous	80%	91%	94.9%	95.019%

We clearly see that the coverages indeed converge to the true value 95% of the confidence level for the simultaneous intervals when M becomes very large, while it converges rather to $\approx 60\%$ for the pointwise intervals.

Shiny application has been build from this graph to better visualize the impact of the number of simulations on the confidence intervals and how their coverage vary, both visually and quantitatively.

Note that the results are similar if we do the experiment on the first derivatives $f^{'}(\cdot)$ of the splines rather than on the splines itself.

Final Results

First, we can see the two significant increases in trend from Figure C.5 let in Appendix **C** when we do not correct for simultaneous intervals. The decrease we have pointed out in the preceding section between 1945 and 1970 with LOESS is hence not significant here and is hence due to randomness.

Some remarks from the two plots in Figure 5.3, considering now splines derivatives :

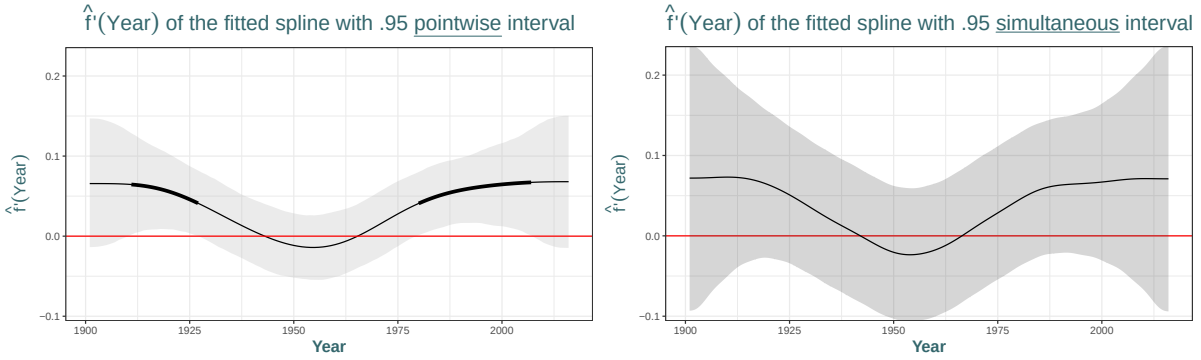

Figure 5.3: *Plots of the first derivative $f^{'}_{(20)}(year)$ of the estimated splines on the retained GAM model. Grey area represents 95% confidence bands. Sections of the spline where the confidence interval does not include zero are indicated by thicker lines.*

- Together with the series in Figure 5.1, we can make the link between the trend behavior of the series and the splines first derivative which accurately models the slope behavior of the trend.
- We notice the increasing trend with decreasing slope, i.e. $f^{''}_{(20)}(\text{year}) < 0$, until year ≈ 1945 where it becomes negative. Then the slope increase and the trend starts to re-increase around 1962. This upward trend that the series of annual maxima is facing for this last period brings light to the climate warming we are all talking about. However, we see that we are now at a "critical point" for several years, i.e. $f^{''}_{(20)}(\text{year}) \approx 0$, bringing light to the likely linear trend.

- Whereas the pointwise confidence interval include significant regions (i.e. 0 is not included in the interval), the simultaneous interval which is accounting for the increase in uncertainty, have no significant regions. Then, we cannot conclude that there has been any significant increasing in the annual maximum temperatures in any periods.

5.3 Comments and Structure of the Analysis

We have found that there is indeed an upward (linear) trend for the series of yearly maxima but this is not significant when doing the correction for simultaneous intervals. Hence, a nonstationary analysis is worth to be continued. After having made this introductory analysis, we will go through with the specific subject of this thesis, the extreme value analysis. Hence, next chapter will present first a stationary GEV analysis and then allow nonstationarity in various ways.

POT As you have seen, the analysis in POT or in GEV involves different methods and different subsets of data. In this text, we will not display the results of the POT analysis to keep the text not too enormous. Henceforth, we will be able to focus on the GEV analysis only. But **note that** the analysis in POT have been done is are available on the same repository presented above. Appendix **D** summarize its contents.

Moreover, we have also conducted some analysis by dividing the dataset by seasons, by hot or cold months (July-August or January-February), ... We also analyzed annual minimum temperatures and we found a trend that is less pronounced for minimum temperatures. Comparisons are interesting, also available on the repository, but we will only focus on a GEV analysis on yearly maxima.

Chapter 6

Analysis in Block Maxima

Contents

In this first analysis, we rely on

This analysis relies on `1intro_stationary.R` code from the **/Scripts-R/** folder of the github repository.

This analysis relies on `2Nonstationary.R` and `2NeuralsNets.R` codes from the **/Scripts-R/** folder of the github repository. A few results of nonstationary analysis in a POT approach are presented in Section 4.1 of the **Summary1_intro** files in the **/vignettes** folder of the repository.

Block-length

The block-length selection an important issue of the analysis. It is important to choose a block-length which is large enough for the limiting arguments supporting the GEV approximation (1.8) to be valid, a large bias in the estimates could occur. But a large block-length implies less data to work with, and thus a large variance of the estimates. A compromise must be found as pointed out in Chapter **1**. We chose yearly blocks that seem justifiable not only for this reason but especially for their interpretability and ease of use. This will induce wastage of data since we will not use 6 month (from October to March) as they will never have an annual maximum. An idea to prevent this wastage is to fit several GEV models on some subsets of the dataset divided by seasonal criterion.

R packages in EVT

A bunch of packages exist for modeling extreme values in R. We have explored and used most of them some to do the following analysis. We made some comparisons and the results are obviously the same for similar methods. Regarding classical EVT analysis, we must name the following :

- ▷ `ismev`, `evd`, `extRemes` (good for a wide nonstationary analysis with POT and nice tutorials, see e.g. Gilleland and Katz [2016]), `POT` (see Ribatet [2006]), `evir`, `fExtremes`, ...

Whereas lots of the package are doing the same analysis but with different tools, we decided to rely mostly on `ismev` as it is the package used in the book of Coles [2001].

6.1 First Inferences of the Model

Whereas the whole content of Chapter **1** is important to understand the concepts used in this section, we will now be mostly based on inferential methods discussed in Section **1.6**, return levels in Section **1.5** and diagnostics in Section **1.7**.

Maximum Likelihood

Relying on functions from the R packages cited above, but also by checking it manually, that is by numerically solving the optimization problem which is a minimization of the negative log-likelihood. This can be done with the `nlm` routine using a Newton-Raphson algorithm (Dennis and Schnabel [1987]). This algorithm is based on an approximation of the log-likelihood by a quadratic function, the second order Taylor series approximation of the log-likelihood for a given point.

Estimation results for the annual maxima series are shown in Table 6.1 :

Table 6.1: *Maximum likelihood estimation of the three GEV parameters*

	Location μ	Scale σ	Shape ξ
Estimates (s.e.)	30.587 (0.216)	2.081 (0.155)	$\mathbf{-0.254}$ (0.067)

The important thing to note from Table 6.1 is the value of the **shape** parameter which is negative. It means that we are under a Weibull-type of the GEV family. From Figure 1.1 the corresponding density

has the form of the red line, i.e. having a finite right endpoint. Moreover, we confirmed that by doing a likelihood ratio test comparing this with a Gumbel distribution. Results are let in Table 6.4.

The Weibull-type implies that the distribution have an estimated right endpoint given by $\hat{x}_* = \hat{\mu} - \hat{\sigma} \cdot \hat{\xi}^{-1} = 38.77°c$. Comparing this value with the maximum value of the series ($= 36.6°c$) tells us that properties of this model take into account the uncertainty from the fact that there are only 116 years of data. Hence, it allows to go beyond this maximum value, with very small probability. This will be highlighted in Figure 6.2 (right plot) where we remark that there are still probabiliy mass beyond the minimum and the maximum values of the series.

Profile log-likelihood intervals As discussed in Section **1.6**, *profile likelihood intervals* are often preferred for individual parameters to handle the poor normal approximation of the MLE. Results provided by the `ismev` package are in Figure C.6 let in Appendix **C**. These intervals are constructed in the following way : search for the horizontal line and then subtract the maximum log-likelihood by half the corresponding upper quantile of the χ^2_{df} for df $= 1$ parameter of interest. We notice that :

- Even at 99%, the interval for $\hat{\xi}$ does not contain 0 supporting our statement that the distribution is left heavy-tailed and right bounded.
- 95% intervals for the location and scale parameters are $[30,\ 31]$ and $[1.8,\ 2.4]$ respectively.
- The intervals do not present much asymetries. In fact, this will be more relevant for return levels as we will see in the next Section.

Probability-Weighted-Moments

It is always a good practice to check whether different methods lead to significantly different results. A second estimator we have seen is the *probability-weighted-moments*. Results are shown in Table 6.2 :

Table 6.2: *GEV parameters estimated by PWM*

	Location μ	Scale σ	Shape ξ
Estimates	30.552	2.115	**−0.232**

We directly see that these results are very close to the MLE's of Table 6.1, in particular for the EVI. This is encouraging for further inference, and we have hence confidence to be under a Weibull-type GEV model. For convenience, we will only keep the maximum likelihood estimates to work with in the following.

6.1.1 Return Levels

First presented in Section **1.5**, return levels are very appreciated by the practitioners for inference in EVT in an environmental context. Usual likelihood intervals relying on the normal approximation are not reliable for return levels. Hence, we decided to compute the profile likelihood intervals and compare both values.

The interpretation we can make from Table 6.3 is that, for example, the estimated 100-year return level is $36.23°c$ and that is the temperature which will be exceeded on average once every 100 years.

Table 6.3: *m-year return level estimates and 95% intervals. Last line computes the difference between the length of the normal interval with the length of the profile likelihood interval*

	2-year	10-year	100-year	1000-year
Estimates	31.315	34.153	36.229	37.982
Normal interval	(30.88, 31.75)	(33.63, 34.67)	(35.21, 37.25)	(35.67, 39.04)
Profile likelihood interval	(31.16, 31.68)	(33.95, 34.74)	(35.54, 37.84)	(36.58, 40.25)
Difference of lengths	0.348	0.247	−0.260	−0.294

Note that very long term extrapolation should be tempered by caution since we only have 116 years of data and predicting far beyond this value will be unreliable. We clearly see the shift of the profile likelihood confidence intervals compared with the normal intervals. This can also be seen on the return level plot in Figure 6.2 where we clearly see blue points going higher than red lines for higher values of the return period. Moreover, we see that the profile likelihood intervals are more precise[1] for "small" return periods, approximatively until half the total number of annual data, and then profile likelihood intervals become wider than normal intervals. This shows how profile likelihood intervals take into account the uncertainty of long-term predictions. Hence, in addition to arguments already provided, uncertainty but also climate warming lead us to have a preference for profile likelihood intervals and, quite surprisingly, this method is not used by default in EV packages such as `ismev`.

6.1.2 Diagnostics

Section **1.7** provided us tools to check the accuracy of the fitted model. The goal is to check that the model $\hat{F}$ fitted by MLE is accurate enough for the true distribution F which is estimated by the empirical df (A.11). First, we present the quantile and the probability plots in Figure 6.1.

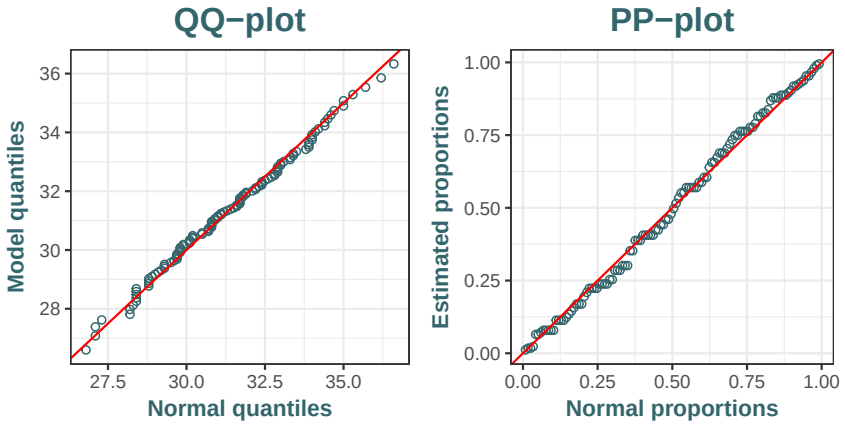

Figure 6.1: *Quantile (left) and probability (right) plots for the stationary GEV model fitted by MLE.*

Both plots show points lying very close to the unit diagonal, demonstrating the empirical df is very close to the fitted model and hence putting confidence that our model fits our data accurately. Right plot

[1] In the sense of a narrower confidence interval. A Monte-Carlo study of accuracy could be made using observed data.

of Figure 6.2 has a similar interpretation and leads to the same conclusion, although the fitted model has a higher peak.

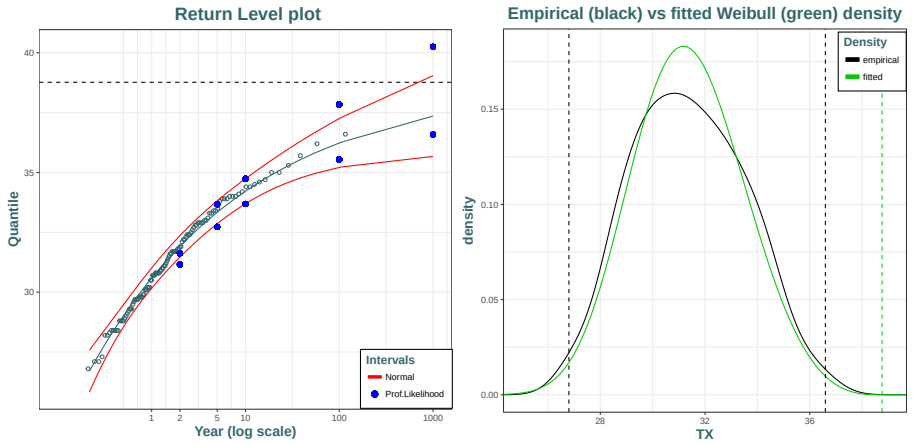

Figure 6.2: *(Left) Return level plot with red lines representing normal confidence intervals, blue points are individual profile likelihood intervals for return levels and horizontal dotted line just represent the right endpoint of the fitted model. (Right) kernel density in black compared with the density of the fitted model in green, with dotted lines representing the endpoints of the distributions, and hence empirical density should not continue outside these lines.*

Return Level Plot

Another tool is available in EVT to check the fit of a model is the return level plot. It allows to compare observations with return levels coming from the EV Weibull model fitted by MLE. Left plot of Figure 6.2 shows us a concave shape of the return levels with the return period wich asymptotes to the right endpoint x_* of the fitted EV Weibull model. This comes from the fact that $\xi < 0$. We remark that all points are very close the estimated return level and hence we put confidence that our model is suitable. Moreover, all these points are inside the normal confidence intervals (and recall that profile likelihood intervals are not suitable for very small return periods).

Whereas the estimated return level cannot go beyond x_*, we see that for very high return periods, the upper bound of the confidence intervals goes beyond this right endpoint. Again, this is justified since for such far periods, these intervals allow to go beyond the domain of the fitted distribution.

Profile Likelihood Intervals for Return Levels

We let in Appendix **C** the Figure C.7 representing the profile log-likelihoods for three return periods and their corresponding intervals, that is at the intersection between the blue line and the curve. We clearly visualize the asymmetries on these graphs and the positive skew which is increasing for higher values of the return period. This was expected since the data at hand provide increasingly weaker information about high levels of the process. We also displayed on the different plots the return levels from Table 6.3, represented by the green lines. These were computed relying on another method from package `extRemes`. Results are slightly different for the 2-year return level.

6.1.3 Stationary Analysis

In Section **3.1** we have proven that when a sequence is not independent we can still have the GEV distribution in the limit of the normalized sequence. This will only induce different location and scale parameters compared to an independent sequence. We can visualize the dependence in the series for example by hand of (partial) autocorrelation functions. Corresponding plots are shown in Figure C.8 let in Appendix **C** where we see that temporal dependence is light but well present. Actually, the estimates shown in Table 6.1 already take this dependence into account.

POT

Dependence is hence not really a concern for GEV while it is more problematic for POT. Indeed, by analyzing only data that are above a threshold (say $30°c$), serial dependence in the data can be strong and points will have the tendency to occur in clusters. We wanted to illustrate this with Figure C.9 in Appendix **C** which highlights this dependence with red lines corresponding to most heavy heat waves in the history of Uccle in summers 1911 and 1976. Indeed, observations lying on the red lines have a dependence since they occurred during a same period of extreme heat. "*Hot days are more likely to be followed by hot days*".

Moreover, we estimated the extremal index θ by the method of Ferro and Segers [2003] to have an idea on the extent of this extremal dependence. We obtained $\hat{\theta} \approx 0.42$ and hence, one interpretation is that the extremes are expected to cluster by groups of mean size $0.42^{-1} \approx 2.4$. We can visualize from Figure C.9 that the points have indeed some tendency to form groups of size 2.

6.2 Parametric Nonstationary Analysis

As depicted in Figure 5.1, even the assumption of a stationarity is likely to be poor for the sequence of annual maxima. Whereas the oscillatory behavior caught by the LOESS model is probably due to noise rather than a true characteristic of the process, the increasing trend is more alarming. Indeed, the trend analysis made in Chapter **5** confirmed that the trend is not significant when we control for simultaneous tests

Our particularly flexible modeling of the trend confirmed that the trend is statistically significant when doing pointwise comparisons but this method is inadequate since we proved that the coverage did not match with the assumed confidence level. If we control for simultaneous tests, significance of the trend completely disappear. However, one could argue that these intervals are very large (look back at Figures 5.2 or 5.3) and thus they are not very precise. The next step (not displayed here) would be to decluster this series.

6.2.1 Comparing Different Models

The first approach we will conduct is by hand of the deviance statistic comparing sequentially nested models. The number of degrees of freedom (df) represent the number of parameters of the model (i.e., its complexity). The parametric models we will first consider are :

1. *Gumbel* : most simple EV-model with only 2 parameters as $\xi = 0$.

2. *stationary* : EV-Weibull model fitted in Section by MLE.
3. *linear in μ* : the location parameter follow $\mu(t) = \beta_0 + \beta_1 \cdot t$.
4. *quadratic in μ* : the location parameter follow $\mu(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2$.
5. *linear in μ and σ* : Same as model 3 for the location but with $\sigma(t) = \exp(\beta_0 + \beta_1 \cdot t)$. The use of the inverse link $b(\cdot) = \exp(\cdot)$ is to ensure positivity of σ $\forall t$.
6. *cubic in μ* : the location parameter follow $\mu(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3$.

In the next Section we will allow for more flexibility in the parameters. Sequential pairwise comparisons will always be made with the best retained model

Table 6.4: *Comparisons of proposed (nested) models for the trend. Significant p-values at 5% are in bold.*

Model	ℓ	df	p-value
Gumbel	-256.84	2	
stationary	-251.75	3	$\mathbf{0.14\%}$
linear in μ	$\mathbf{-241.81}$	**4**	$\mathbf{10^{-3}\%}$
quadratic in μ	-241.48	5	42%
linear in μ and in σ	-241.69	5	63%
cubic in μ	-241.37	6	65%

Note that the system is computationally singular[2] for cubic and more complex models. The model that is chosen by this procedure is model 3 which allows a linear model for the location. The choice is clearly supported by likelihood ratios and hence we can put great confidence that this selection is straightforward.

6.2.2 Diagnostics and Inference

As explained in Section **3.2.1**, diagnostic tools such as quantile and probability plots can still be used in the context of nonstationarity with some transformations.

Results are shown in Figure C.10 let in Appendix **C** for the selected model. We can appreciate that the fit seems accurate. Problems for large quantiles in the QQ-plot is not problematic.

Return Levels

We will still use the return levels as our tool to make inference in EVT. We will then plot these return levels against years by taking

We clearly see the linear pattern...

Quite surprinsignly(?), we see that the fitted return level after n years where n is the number of data is $36.4°c$ which is actually very close to the maximum of the series $36.6°c$.

[2]

Caution should be exercised in practice concerning whether or not it is believable for the upward linear trend in maximum temperatures to continue to be valid.

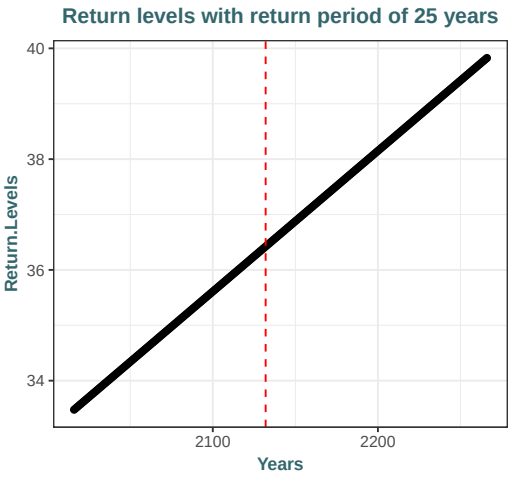

Figure 6.3

6.3 Improvements with Neural Networks

Model parameters are estimated via GML using a quasi-Newton BFGS optimization algorithm, and an appropriate GEV-CDN architecture which is the one that minimizes the appropriate cost-complexity model selection criteria (AIC_c or BIC). Different structures are tested with combinational cases of stationary and nonstationary parameters of the GEV distribution, linear and nonlinear architecture of the CDN and combinations of the input covariates

Final Results

Note that we allowed the shape parameter ξ to vary with time to consider most possible models while this was not advised for EV models. Anyway, we showed all the results and that does not change the final result.

As we mentioned, the NN is meant to approximate any functions with good accuracy. It comprise thus all the models considered so far.

weight penalty regularization (controlled via the gaussian prior controlled via sd.norm in gev.fit or also gev.bag must be controlled by cross validation. That means that we have to find the value of the variance for this prior that provide the best fit (aic or bic ?)

Based on Figure 3.1 which represent the fully connected architecture, the hierarchy of models we will consider is, by ascending complexity

•

We choosed the value of 6 and 9 (?) for c_1 and c_2

Cannon [2010] recommended to use between 1 and 3 (4) hidden layers due to the relatively small sample of annual extremes (here 117).

This Table actually confirms the finding of Table

(mis aussi ds chap3) "Another pitfall is its lack of interpretation of the relationships retrieved by the model between inputs and outputs but it bears noting that sensitivity analysis methods as in Cannon and McKendry [2002] could be used to identify the form of nonlinear relationships between covariates and GEV distribution parameters or quantiles."

Table 6.5: *Put the function nonlinear sigmoid or identity ?)*

model	AIC_c	BIC	hidden	df
stationary	-19.6	-11.5	0	3
$\boldsymbol{\mu_t}$	**-37.4**	**-26.7**	**0**	**4**
μ_t, σ_t	-35.4	-22.2	0	5
μ_t, σ_t, ξ_t	-34.2	-18.4	0	6
μ_t	-35.4	-19.6	1	6
μ_t, σ_t	-36.2	-17.9	1	7
μ_t, σ_t, ξ_t	-34	-13.3	1	8
μ_t	-37.4	-14.3	2	9
μ_t, σ_t	-32.5	-4.7	2	11
μ_t, σ_t, ξ_t	-38.4	3.9	2	13

Inference : Confidence intervals by Bootstrap

Empirical Coverage analysis of the bootstrap procedures ? (MOnte Carlo)

6.4 Comments and Comparisons with POT

In practical analysis of extreme values such data, there are plenty of ways to analyze and we considered some of them.

First of all, the other approach we have seen is in Chapitre 2 is the POT

As we did not have any precise goal to achieve, we decided to only consider the annual analysis in block maxima.

In excess over a threshold models, the Point Process approach

Chapter 7

Bayesian Analysis in Block Maxima

Contents

PUT the examples right in the place where it is mentioned in the theory.! "As we have seen in section 2.1.1.... and in section 2.2.2....." This analysis relies on all the codes which filenames start by "`Bayes`" from the **/Scripts-R/** folder of the github repository. Al the functions created that are used and are also made available through the package are in **/R/`BayesFunc.R`**.

"It is often the case that more than one model provides an adequate fit to the data. Sensitivity analysis determines by what extent posterior inferences change when alternative models are used" book risk analysis other section pp.2.

"The basic method of sensitivity analysis is to fit several models to the same problem. Posterior inferences from each model can then be compared."

7.1 From `evdbayes` R package : MH algorithm

The `evdbayes` is the only package available on CRAN for (see Ribatet [2006]) is a very old package which has not been updated since a while. We had problems to understand both its structure and the famous "black-box"

7.2 From Our Functions (R package)

7.3 From HMC algorithm using STAN language

The problem is maybe from 1.6.1. The parameter ξ is relatively near the region that could be problematic, causing convergence issues.

7.4 Ratio of Uniform : `revdbayes` package

https://cran.r-project.org/web/packages/revdbayes/vignettes/revdbayes-vigne
html

From the `revdbayes`

Helped with `rust`

7.5 Comparisons

Hartmann and Ehlers [2016] We can calculate the effective sample size (ESS) using the posterior samples for each parameter :

$$\text{ESS} = N \cdot (1 + 2\sum_{k} \gamma(k))^{-1} \tag{7.1}$$

where N is still the number of posterior samples and $\gamma(k)$ are the monotone lag k sample autocorrelations. ?. We can thus interpret this as the number of effectively independent samples.

7.5.1 STAN

Benefits :

- Allows more flexibility (?) through the mathematical formulation of the formula
- It is really smoother and clearer (straightforward) for this kind of problems

Drawbacks :

- New language with all the problems/errors arising when learning it.

7.6 Comparison with frequentists results

In this first analysis, we rely on

Conclusion

During this thesis, we have statistically assessed the presence of a trend in the extreme temperatures in Uccle. We first detected that the trend is significative by the method of linear regression. We also discovered that the best fitted GEV model is the one with a linear trend in the location parameter.

"A key issue in applications is that inferences may be required well beyond the observed tail of the data, and so an assumption of stability is required:" Davison et al. [2012]

"Another approach would be to use something other than time as the covariate in the model. For instance, one could imagine linking temperature data directly to CO2 level rather than time. However, linking to a climatological covariate makes extrapolation into the future more difficult, as one would need to extrapolate the covariate as well. No obvious climatological covariate comes to mind for the Red River application. "

Timescale-uncertainty effects on extreme value analyses seem not to have been studied yet. For stationary models (Sect. 6.2), we anticipate sizable effects on block extremes–GEV estimates only when the uncertainties distort strongly the blocking procedure. For nonstationary models (Sect. 6.3), one may augment confidence band construction by inserting a timescale simulation step (after Step 4 in Algorithm 6.1) Mudelsee [2014, pp.262]

!!!!! not put too much references in the text !!!!!

A project to give analysis in time live could be implemented, with values being stocked each days/year !

be prudent that all "methods" are listed (numbered) in (each) ToC

add square brackets [] for cite ?

Careful for boldsymbols, especiallyin bayesian

PissoortThesis:: Behind our functions

[MOTS DEXPLICATIONS SUR TTES LES FORMULES (domain atraction condition, etc...]

Finish (Bayesian) documentation in the package.

See new graphs for stats desc. (in appendix), from workshop exam.

Create Gif animations for the github readme

Appendix

Appendix A

Statistical tools for Extreme Value Theory

A.1 Tails of the distributions

Heavy-tailed

Definition A.1 (Heavy-tails). *The distribution of a random variable X with distribution function F is said to have a* heavy *right* tail *if*

$$\lim_{n\to\infty} e^{\lambda x} \, Pr\{X > x\} = \lim_{n\to\infty} e^{\lambda x} \bar{F}(x) = \infty, \quad \forall \lambda > 0. \tag{A.1}$$

△

More generally, we can say that a random variable X has heavy tails if $\Pr\{|X| > x\} \to 0$ at a polynomial rate. In this case, note that some of the moments will be undefined.

Definition A.2 (Fat-tails). *The distribution of a random variable X is said to have a* fat tail *if*

$$\lim_{x\to\infty} Pr\{X > x\} = x^{-\alpha}. \tag{A.2}$$

△

Definition A.3 (Long-tails). *The distribution of a random variable X with distribution function F is said to have a* long *right* tail *if* $\forall t > 0$,

$$\lim_{x\to\infty} Pr\{X > x + t | X > x\} = 1 \;\Leftrightarrow\; \bar{F}(x + t) \sim \bar{F}(x) \quad as\, x \to \infty. \tag{A.3}$$

△

Definition A.4 (Light-tails). *Conversely, we say that X has* light tails *or* exponential tails *if its tails decay at an exponential rate, i.e.*

$$\lim_{x\to\infty} Pr\{|X| > x\} = e^{-x} \tag{A.4}$$

△

An intuitive example of a distribution with exponential tails such as the exponential distribution.

A.2 Convergence concepts

Convergence in distribution

Definition A.5 (Convergence in distribution). *We say that a sequence X_n with df F_n converges in distribution to X with df F, if*

$$F_n(x) := Pr\{X_n \leq x\} \longrightarrow Pr\{X \leq x\} := F(x), \tag{A.5}$$

at all continuity points of F. △

It means that, for large n, $\Pr\{X_n \leq t\} \approx \Pr\{X \leq t\}$. We denote this by $X_n \xrightarrow{d} X$.

Convergence in probability

Definition A.6 (Convergence in probability). *We say that a sequence X_n* converges to X in probability *if, $\forall \epsilon > 0$,*

$$Pr\Big\{|X_n - X| > \epsilon\Big\} \to 0, \qquad n \to \infty. \tag{A.6}$$

△

Hence, it means that the probability of the differnece between X_n and X goes to 0 as n is large. We denote this by $X_n \xrightarrow{P} X$.

An example of application of this convergence is the *Weak Law of Large Numbers.*

Theorem A.1. *[Weak Law of Large Numbers] Let a sequence of R.V. $\{X_i\}_{iid}$ be defined of the same probability space with mean μ and variance $\sigma^2 < \infty$. Then, we know that the difference between $\bar{X}_n$ and μ will go to 0 in probability, i.e. $\bar{X}_n \xrightarrow{p} \mu$.* □

But this law actually makes a stronger convergence, following Kolmogorov et al. [1956], that is an *almost sure convergence*

Almost Sure Convergence

This is the type of stochastic convergence that is most similar to pointwise convergence known from elementary real analysis.

Definition A.7 (Almost Sure convergence). *We say that a sequence of random variables X_n* converges almost surely *(or with probability one) to X if*

$$Pr\Big\{X_n = X\Big\} = 1, \qquad n \to \infty. \tag{A.7}$$

△

We can denote this by $X_n \xrightarrow{\text{a.s.}} X$. This means that the values of X_n approach the value of X, in the sense that events for which X_n does not converge to X have probability 0.

Well other forms of convergence do exist, but these ones are the most important in regard to EVT. However, the reader may refer e.g. to **?** for more in-depth results.

A.3 Varying functions

Definition A.8 (Regularly varying function)**.** *Let's consider the survival $\bar{F}$. We say that this survival function $\bar{F}$ is **regularly varying** with index $-\alpha$ if*

$$\lim_{x\to\infty} \frac{\bar{F}(xt)}{\bar{F}(x)} = t^{-\alpha}, \qquad t > 0. \tag{A.8}$$

We write it $\bar{F} \in R_{-\alpha}$. △

Definition A.9 (Slowly varying function)**.** *We say that a function f is **slowly varying** with index $-\alpha$ if*

$$\lim_{x\to\infty} \frac{f(tx)}{f(x)} = 1, \qquad t > 0. \tag{A.9}$$

△

We remark that a slowly varying function is a regularly varying function with index 0.

A.4 Diagnostic Plots : Quantile and Probability Plots

From Beirlant et al. [1996, pp.18-36], together with the nice view of Coles [2001, pp.36-37], we present two major diagnostic tools which aim at assessing the fit of a particular model (or distribution) against the real distribution coming from the data used to construct the model. These are called the *quantile-quantile* plot (or *qq*-plot) and the *probability* plot (or *pp*-plot).

These diagnostics are popular by their easy interpretation and by the fact that they can both have graphical (i.e. subjective, qualitative, quick) view but also a more precise (i.e. objective, quantitative, rigorous) analysis can be derived, for example from the theory of linear regression.
For these two diagnostic tools, we use the order statistics as seen (1.1) but now we rather consider an **ordered sample** of independent **observations** :

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \tag{A.10}$$

coming from a population from which we fit the estimated model (distribution) $\hat{F}$ and where $x_{(1)}$ (resp. $x_{(n)}$) is thus the minimum (resp. maximum) observation in the sample. We also define the **empirical distribution function**

$$\tilde{F}(x) = \frac{i}{n+1}, \qquad x_{(i)} \leq x \leq x_{(i+1)}. \tag{A.11}$$

$\tilde{F}$ is an estimate of the true distribution F and hence, by comparing $\hat{F}$ and $\tilde{F}$, it will help us to know if the fitted model $\hat{F}$ is reasonable for the data.

Quantile plot

Given a ordered sample as in (A.10), a *qq-plot* consists of the locus of points

$$\left\{\left(\hat{F}^{\leftarrow}\left(\frac{i}{n+1}\right),\ x_{(i)}\right)\ :\ i=1,\ldots,n\right\}. \tag{A.12}$$

This graph compares the ordered quantiles $\hat{F}^{\leftarrow}\left(\frac{i}{n+1}\right)$ of the fitted model $\hat{F}$ against the ordered observed quantiles, i.e. the ordered sample from (A.10). We used the continuity correction $\frac{i}{n+1}$ to prevent problems at the borders. Note that a disadvantage of Q–Q plots is that the shape of the selected parametric distribution is no longer visible Beirlant et al. [2006, pp.63]

Probability plot

Given the same sample in (A.10), a *probability plot* consists of the locus of points

$$\left\{\left(\hat{F}(x_{(i)}),\ \frac{i}{n+1}\right)\ :\ i=1,\ldots,n\right\}. \tag{A.13}$$

This graph compares the estimated probability of the ordered values $x_{(i)}$, thus from the fitted model $\hat{F}$, against the probability coming from the empirical distribution as in (A.11).

From these two graphical diagnostic tools, the interpretation is the same and we will consider that $\hat{F}$ fits well the data if the plot looks linear, i.e. the points of the plots lie close to the unit diagonal.

Besides the fact that the probability and the quantile plots contain the same information, they are expressed in a different scale. That is, after changing the scale to probabilities or quantiles (with probability or quantile transforms), one can gain a better perception and both visualizations can sometimes lead contradictory conclusions, especially in the graphical inspection. Using both is thus preferable to make our model's diagnostic more robust.

A.5 Estimators Based on Extreme Order Statistics for EVI

The following estimators allow to estimate the EVI ξ.

Pickands estimator

First introduced by Pickands [1975], this method can be applied $\forall \xi \in \mathbb{R}$ to give

$$\hat{\xi}_k^P = \frac{1}{\ln 2}\ln\left(\frac{X_{n-\lceil k/4\rceil+1,n} - X_{n-\lceil k/2\rceil+1,n}}{X_{n-\lceil k/2\rceil+1,n} - X_{n-k+1,n}}\right), \tag{A.14}$$

where we recall that $\lceil x\rceil$ denotes the integer (ceil) part of x.

A condition for the consistency of this estimator is that k must be chosen such that $k/n \to 0$ as $n \to \infty$. This condition will hold for the following estimators based on order statistics.

A problem with this intuitive estimator is that its asymptotic variance is very large (see e.g. Dekkers and Haan [1989]) and depends highly on the value of k. To improve this, we can quote the estimator of Segers [2001] which is globally more efficient.

The following estimators are only valid for $\xi > 0$. In general in EVT, this condition should hold

(rainfall data, finance, risk analysis,...) but in our application we know that ξ is likely to be negative and hence, the following estimators cannot be used.

Hill estimator

This is probably the most simple EVI estimator thanks to the intuition behind its construction. There exists plenty of interpretations to construct it (see e.g. Beirlant et al. [2006, pp.101-104]). It is defined as

$$\xi_k^H = k^{-1} \sum_{i=1}^{k} \ln X_{n-i+1,n} - \ln X_{n-k,n}, \qquad k \in \{1, \dots, n-1\}. \tag{A.15}$$

Following e.g. de Haan and Resnick [1998], this estimator is consistent under certain conditions. Besides that, this estimator has several problems :

- instability with respect to the choise of k.
- Severe bias due to the heavy-tails of the distribution and thus the slowly varying component which influences negatively.
- Inadequacy with shifted data.

Hence, this estimator should be carefully used.

Moment estimator

Introduced by Dekkers et al. [1989], this estimator is a direct generalization of the Hill estimator presented above. It is defined as

$$\hat{\xi}_k^M = \hat{\xi}_k^H + 1 - \frac{1}{2}\left(1 - \frac{(\hat{\xi}_k^H)^2}{\hat{\xi}_k^{H(2)}}\right)^{-1}, \tag{A.16}$$

where

$$\hat{\xi}_k^{H(2)} = k^{-1} \sum_{i=1}^{k} \left(\ln X_{n-i+1,n} - \ln X_{n-k,n}\right)^2.$$

This estimator is also consistent.

Appendix B

Bayesian Methods

B.1 Algorithms

B.1.1 Metropolis–Hastings Algorithm

The *Metropolis–Hastings* algorithm is one of the first and of the pioneering algorithm discovered by Hastings [1970] to compute MCMC for Bayesian analysis.

Algorithm 1: The Metropolis–Hastings Algorithm

1. Pick a starting point θ_0 and fix some number N of simulations.
2. **For** $t = 1, \ldots, N$ **do**
   (a) Sample proposal θ_* from a proposal density $p_t(\theta_*|\theta_{t-1})$,
   (b) Compute the ratio
   $$r = \frac{\pi(\theta_*|\boldsymbol{x}) \cdot p_t(\theta_{t-1}|\theta_*)}{\pi(\theta_{t-1}|\boldsymbol{x}) \cdot p_t(\theta_*|\theta_{t-1})} = \frac{\pi(\theta_*) \cdot \pi(\boldsymbol{x}|\theta_*) \cdot p_t(\theta_{t-1}|\theta_*)}{\pi(\theta_{t-1}) \cdot \pi(\boldsymbol{x}|\theta_{t-1}) \cdot p_t(\theta_*|\theta_{t-1})}.$$
   (c) Set
   $$\theta_t = \begin{cases} \theta_* & \text{with probability } \alpha = \min(r, 1); \\ \theta_{t-1} & \text{otherwise.} \end{cases}$$

This algorithm remains valid when π is only proportional to a target density function and thus it can be used to approximate 4.1 .

Note that the proposal density is often chosen to be symmetric so that we will just sample under a "simple Metropolis" algorithm where r is thus simplified to be only the ratio of the posterior densities, $r = \frac{\pi(\theta_*|\boldsymbol{x}) \cdot}{\pi(\theta_{t-1}|\boldsymbol{x})}$.

We can shortly summarize the *pros* and *cons* this algorithm :

- *PROS* : Very easy to program and works even for relatively complex densities.
- *CONS* : Can be very inefficient, in the sense that it will require lots of iterations before the stationary target distribution will be reached. This requires some tuning to the algorithm through

B.1.2 Gibbs Sampler

The *Gibbs Sampler* can be seen as a special case of the Metropolis-Hastings algorithm. Suppose our parameter vector θ can be divided into d subvectors $(\theta_1, \ldots, \theta_d)$, and let's say in our case that each of these "subvectors" represent a single parameter, thus typically one of the three (μ, σ, ξ), for the simplest case so that $d = 3$ in this model. At each $t = 1, \ldots, N$, the Gibbs sampler samples the subvectors $\theta_t^{(j)}$ conditional on both the data $\boldsymbol{x}$ and the remaining subvectors $\theta_{t-1}^{(-j)}$ at their current values. Therefore, we have $\theta_{t-1}^{(-j)} = \left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)}\right)$ and each $\theta_t^{(j)}$ is sampled from $\pi\left(\theta^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right)$.

Algorithm 2: PSEUDOCODE of the Gibbs Sampler

1. Pick a starting point θ_0 and fix some number N of simulations.
2. **For** $t = 1, \ldots, N$ **do**
   For $j = 1, \ldots, d$ **do**
   (a) Sample proposal θ_* from a proposal density $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)})$,
   (b) Compute the ratio

$$\begin{aligned} r =& \frac{\pi\left(\theta_*^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_{t-1}^{(j)}|\theta_*^{(j)}\right)}{\pi\left(\theta_{t-1}^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_*^{(j)}|\theta_{t-1}\right)} \\ =& \frac{\pi\left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_*^{(j)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)} \mid \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_{t-1}^{(j)}|\theta_*^{(j)}\right)}{\pi\left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_{t-1}^{(j)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)} \mid \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_*^{(j)}|\theta_{t-1}\right)}, \end{aligned}$$

   (c) Set

$$\theta_t^{(j)} = \begin{cases} \theta_*^{(j)} & \text{with probability } \alpha = \min(r, 1); \\ \theta_{t-1}^{(j)} & \text{otherwise.} \end{cases}$$

(better signs for conditional bar "|" !!!!!!!

This algorithm depends on being able to simulate from $\pi\left(\theta^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right)$ which is often impossible. However, one can use Metropolis-Hastings to $\pi\left(\theta^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right)$, giving the above.

A special case arise if one can simulate directly so that $r = 1$, i.e. we take $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)}) = \pi(\theta_*^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x})$. (The proposal $p_{t,j}(\,\cdot\,)$ is also often symmetric, i.e. $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)}) = p_{t,j}(\theta_{t-1}^{(j)}|\theta_*^{(j)})$. But it cannot be simplified in the equation.)

It is important for our tasks to tune the average probability of acceptance to be roughly between 0.4 and 0.5 (see e.g. Gelman et al. [2013, chapter 11]) so that the so-generated markov-chain has desirable properties. This is doneby setting the standard deviation $\sigma(j)$ of the univariate normal distribution taken for $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)})$. Whereas our $\theta(j)$ are (often?) univariate, it is difficult to set each $\sigma^{(j)}$ to achieve average acceptance probabilities for all parameters. We will then use a trial-and-error approach.

Note also the increase of complexity with this sampler compared to the Metropolis-Hastings, where the nested loop implies that there are d iterations with each simulation.

pros and *cons* :

- *PROS* : Easy to program and, for some problems, it can also be very efficient. It is a pleasant way to split multidimensional problems into simpler (typically univariate) densities.
- *CONS* : Sometimes hard to compute analytically the conditional distributions. Not all densities can be split into pleasant conditionals equations.

Metropolis-within-Gibbs?

B.1.3 Hamiltonian Monte Carlo

http://deeplearning.net/tutorial/hmc.html#hmc

A difficulty we have faced and that we would like to point out for GEV models is (see p.316-317 STAN manual) t

"Most of the computation [in Stan] is done using Hamiltonian Monte Carlo. HMC requires some tuning, so Matt Hoffman up and wrote a new algorithm, Nuts (the "No-U-Turn Sampler") which optimizes HMC adaptively. In many settings, Nuts is actually more computationally efficient than the optimal static HMC! "

The *Hamiltonian Monte Carlo* (HMC)

"The Hamiltonian Monte Carlo algorithm starts at a specified initial set of parameters theta; in Stan, this value is either user-specified or generated randomly. Then, for a given number of iterations, a new momentum vector is sampled and the current value of the parameter θ is updated using the leapfrog integrator with discretization time and number of steps L according to the Hamiltonian dynamics. Then a Metropolis acceptance step is applied, and a decision is made whether to update to the new state (theta*; rho*) or keep the existing state" **?**

? have demonstrated in similar application that HMC (and Riemann manifold HMC) are much more computationally efficient than traditional MCMC algorithms such as MH.

Definition B.1 (Total energy of a closed system : Hamiltonian function)**.** *For a certain particle; Let $\pi(\theta)$ be the posterior distribution and let $\boldsymbol{p} \in \mathbb{R}^d$ denote a vector of auxiliary parameters independent of θ and distributed as $\boldsymbol{p} \sim N(\mathbf{0}, \boldsymbol{M})$. We can interpet θ as the position of the particle and $-\log \pi(\theta|\boldsymbol{x})$ describes its potential energy while $\boldsymbol{p}$ is the momentum with kinetic energy $\boldsymbol{p}'\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{p} \cdot 2^{-1}$. Then the total energy of a closed system is the Hamiltonian function*

$$\mathcal{H}(\theta, \boldsymbol{p}) = -\mathcal{L}(\theta) + \boldsymbol{p}'\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{p} \cdot 2^{-1}, \qquad \text{where} \qquad \mathcal{L}(\theta) = \log \pi(\theta). \tag{B.1}$$

△

We define $\mathcal{X} = (\theta, \boldsymbol{p})$ as *the combined state* of the particle.

The unormalized joint density of $(\theta, \boldsymbol{p})$ is

$$f(\theta, \boldsymbol{p}) \propto \pi(\theta) \cdot \exp\{-\boldsymbol{p}'\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{p} \cdot 2^{-1}\} \propto \exp\{-\mathcal{H}(\theta, \boldsymbol{p})\}. \tag{B.2}$$

Following Hartmann and Ehlers [2016], the idea is to use the Hamiltonian dynamics equations (not shown here for..) to model the evolution of a particle that keep the total energy constant. Introducing the auxiliary variables $\boldsymbol{p}$ and using the gradients (..) will lead to a more efficient exploration of the parameter space

These differential equations cannot be solved so numerical integrators are required, for instance the "Störmer-Verlet" from **?** which will introduce discretization. A MH step is then required to correct the error and ensure convergence. The new proposal $\mathcal{X}_* = (\theta_*, \boldsymbol{p}_*)$ will be accepted with probability

$$\alpha(\mathcal{X}, \mathcal{X}_*) = \min\left[\frac{f(\theta_*, \boldsymbol{p}_*)}{f(\theta, \boldsymbol{p})}, 1\right] = \min\Big[\exp\left\{\mathcal{H}(\theta, \boldsymbol{p}) - \mathcal{H}(\theta_*, \boldsymbol{p}_*)\right\},\ 1\Big]. \tag{B.3}$$

As $\boldsymbol{M}$ is symmetric positie definite, $\boldsymbol{M} = m\boldsymbol{I}_d$. Then we can summarize the HMC algorithm in the following, in its 'simplest' form :

marie est la best : $Moyenne = \frac{1}{n}\sum_{i=1}^{n} X_i$

Algorithm 3: The Hamiltonian Monte Carlo algorithm

1. Pick a starting point θ_0 and set $i = 1$.
2. **Until** convergence has been reached **do**
   (a) Sample $\boldsymbol{p}_* \sim N_d(\mathbf{0}, \boldsymbol{I}_d)$ and $u \sim U(0, 1)$,
   (b) Set $(\theta_I, \boldsymbol{p}_I) = (\theta_{i-1}, \boldsymbol{p}_*)$ and $\mathcal{H}_0 = \mathcal{H}(\theta_I, \boldsymbol{p}_I)$,
   (c) **repeat** L times
   ▷ $\boldsymbol{p}_* = \boldsymbol{p}_* + \frac{\epsilon}{2}\nabla_\theta \mathcal{L}(\theta_{i-1})$
   ▷ $\theta_{i-1} = \theta_{i-1} + \epsilon \cdot \boldsymbol{p}_*$
   ▷ $\boldsymbol{p}_* = \boldsymbol{p}_* + \frac{\epsilon}{2}\nabla_\theta \mathcal{L}(\theta_{i-1})$,
   (d) Set $(\theta_L, \boldsymbol{p}_L) = (\theta_{i-1}, \boldsymbol{p}_*)$ and $\mathcal{H}^{(1)} = \mathcal{H}(\theta_L, \boldsymbol{p}_L)$,
   (e) Compute $\alpha\Big[(\theta_I, \boldsymbol{p}_I, (\theta_L, \boldsymbol{p}_L)\Big] = \min\Big[\exp\left\{H^{(0)} - H^{(1)}\right\},\ 1\Big]$,
   (f) **If** $\alpha\Big[(\theta_I, \boldsymbol{p}_I), (\theta_L, \boldsymbol{p}_L)\Big] > u$ **then** set $\theta_i = \theta_L$
   else set $\theta_i = \theta_I$,
   (g) Increment $i = i + 1$ and return to step (a).

As you can see, it is not trivial. The basic idea to keep in mind is that jumping rules are much more efficient than for traditional algorithms because they learn from the gradient of the log posterior density, so they know better where to jump to. As a result, it can be MUCH more efficient.

Chains are expected to reach stationarity faster as it proposes moves to regions of higher probabilities.

pros and *cons* :

- *PROS* : Easy to program as we just have to write down the model. Very efficient in general, and works for all types of problems.

- *CONS* : Need to learn how to use STAN, less control over the sampler bu maybe it is for the best?

Appendix C

Other Figures and Tables

C.1 GEV : Influence of the Parameters on the Shape of the Distribution

Regarding our future application, that is maximum temperatures, it is relevant to consider values of the location parameter μ around 30 degrees.

C.2 Introduction of the Practical Analysis (section 6)

Table C.1: *compares models for the residuals of the GAM model based on AIC and BIC criterion. These criterion take into account the quality of fit (based on likelihood) but also a penalty term to penalize more complex models.*

	df	AIC	BIC
Uncorrelated	4	494.635	505.650
AR(1)	5	494.356	508.124
MA(1)	5	493.706	507.474
ARMA(1,1)	6	492.511	509.033
AR(2)	6	495.133	511.654
MA(2)	6	494.698	511.219

C.3 Analysis by GEV

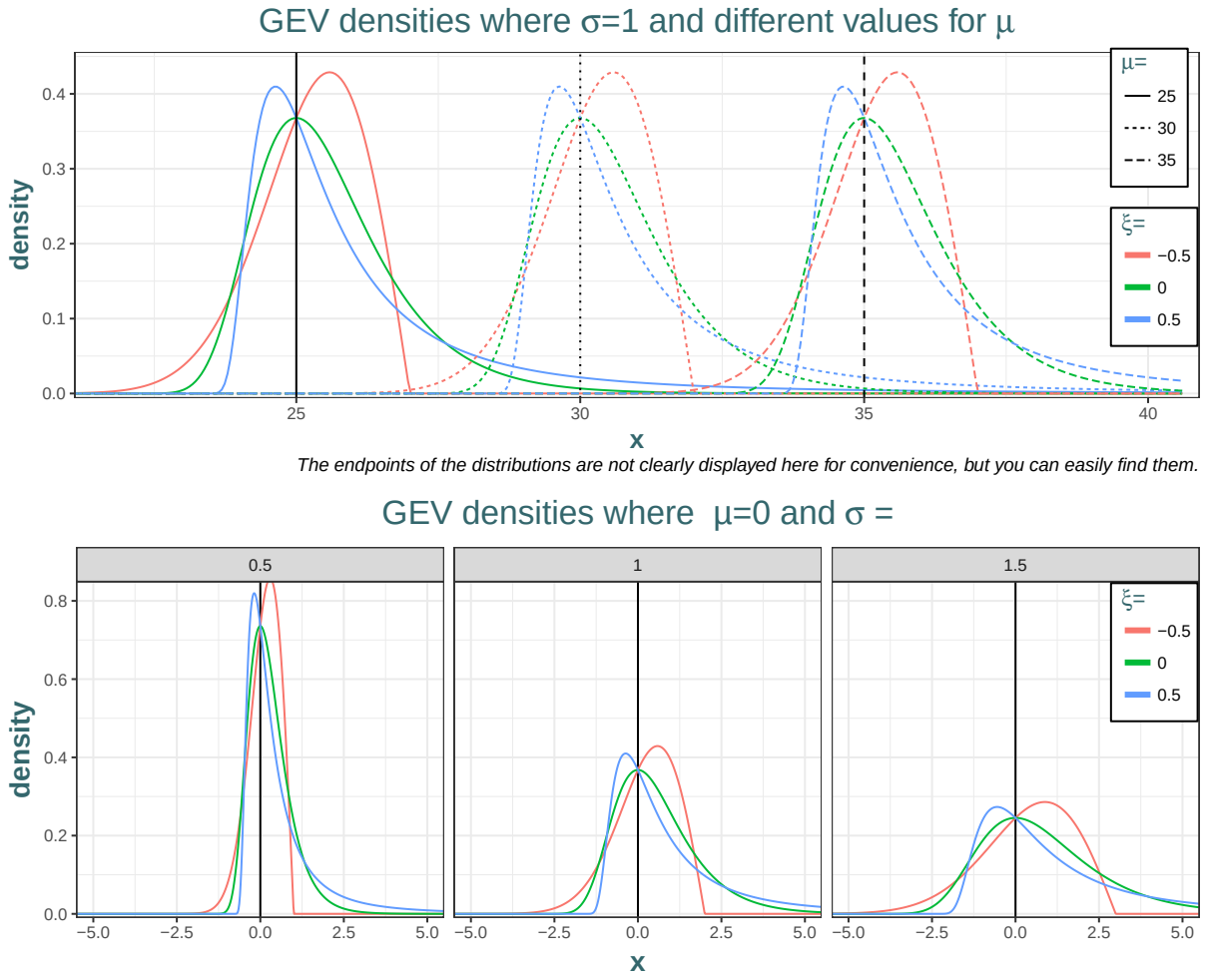

Figure C.1: *GEV distribution for different values of the three parameters*

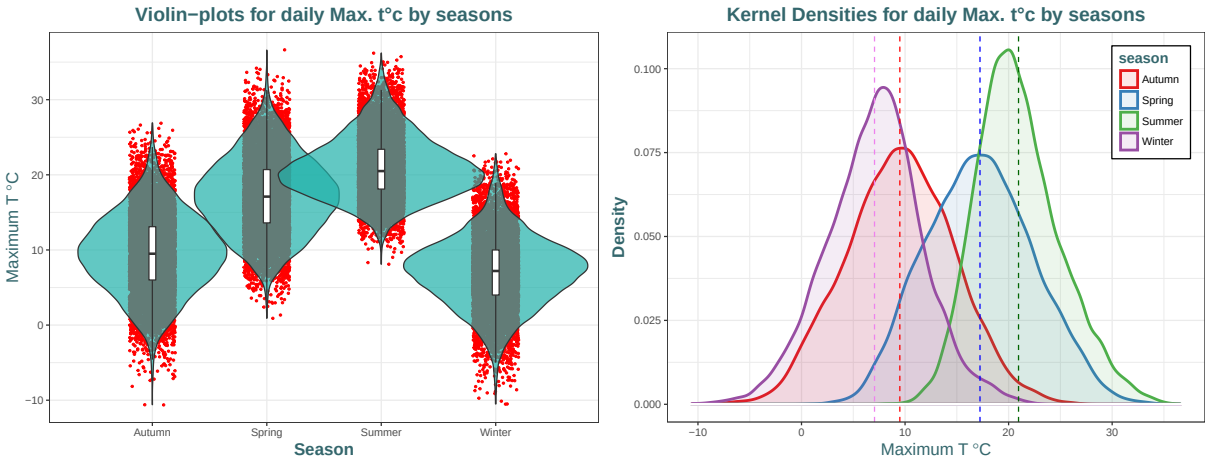

Figure C.2: *Violin-plot and density plot for each seasons. (Right) vertical dotted lines represent the mean of each distribution.*

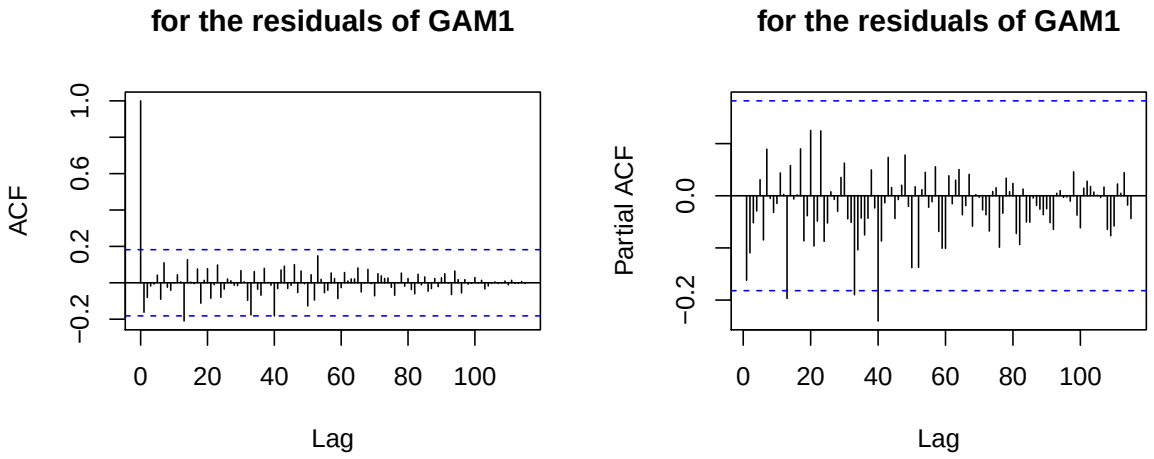

Figure C.3: *ACF and PACF for the residuals of the fitted GAM model with assumed independent errors*

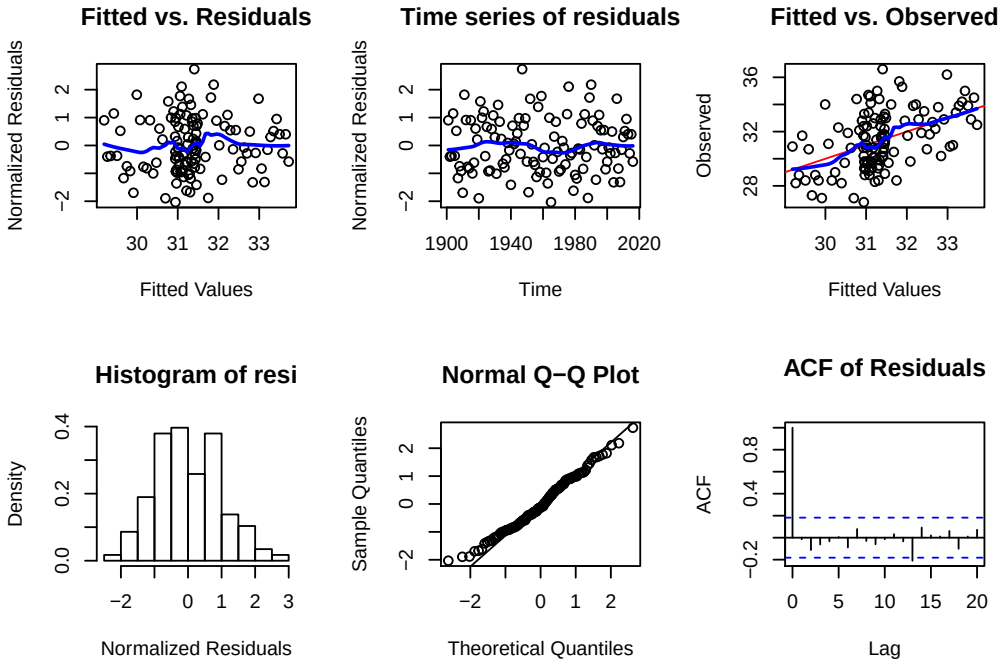

Figure C.4: *Diagnostics of the chosen GAM model with Whinte Noise process on the errors, based on the residuals.*

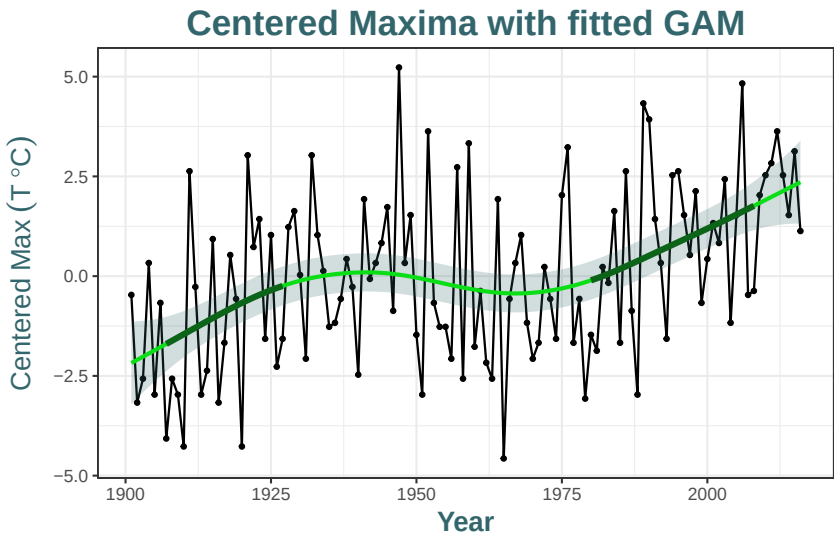

Figure C.5: *Series of annual maxima together with the fitted GAM model (in green)* ***with MA(1) model on the residuals****. Thicker lines indicate that the increase is significant for pointwise confidence interval. Shaded area represent a "95%" interval for the predicted values which looks quite narrow.*

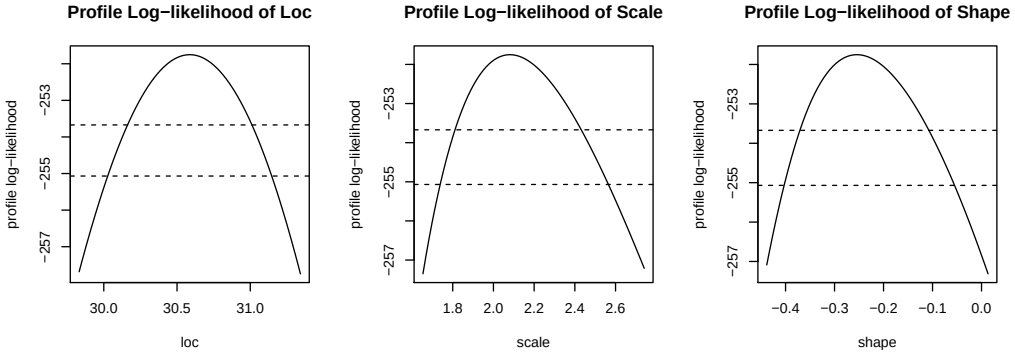

Figure C.6: *Profile likelihood intervals for the three GEV parameters. The two horizontal dotted lines represent the 95% (above) and 99% (below) confidence intervals when we take the intersection on the horizontal axis. Output makes use of the* `ismev` *package.*

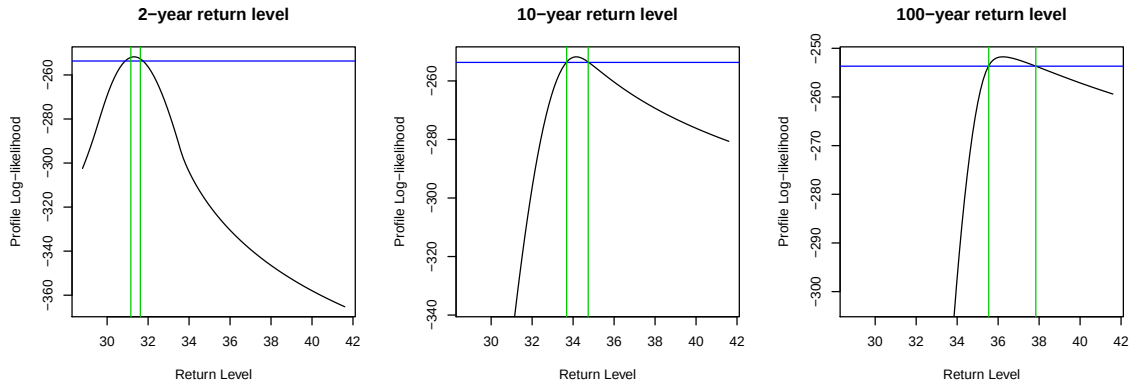

Figure C.7: *95% Profile likelihood intervals for three return levels. We kept the same x-scale for the three plots but not the y-scales. We used the* `ismev` *package but we modified the function to allow for more flexibility because the default y-scale in produced ugly visualizations for high return levels. Green lines represent the intervals from Table* **??** *computed with another package from E.Gilleland,* `extRemes`*.*

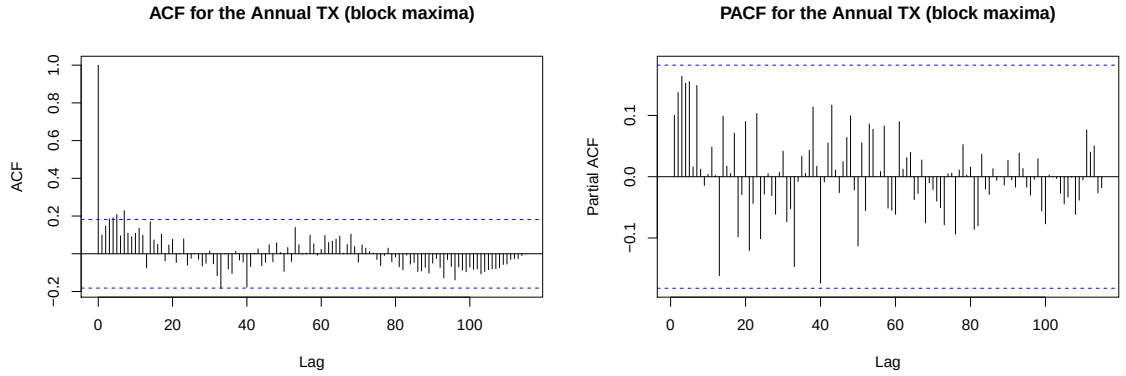

Figure C.8: *ACF and PACF for the residuals of the fitted GAM model with assumed independent errors*

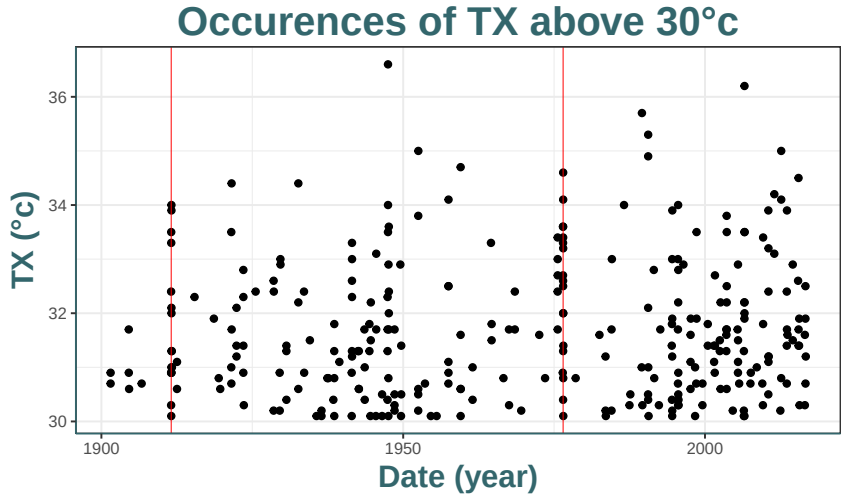

Figure C.9: *Plot of all daily TX that exceeded $30°c$ in the period [1901,2016] in Uccle. Red lines highlights two periods of heavy heat waves during summers 1911 and 1976.*

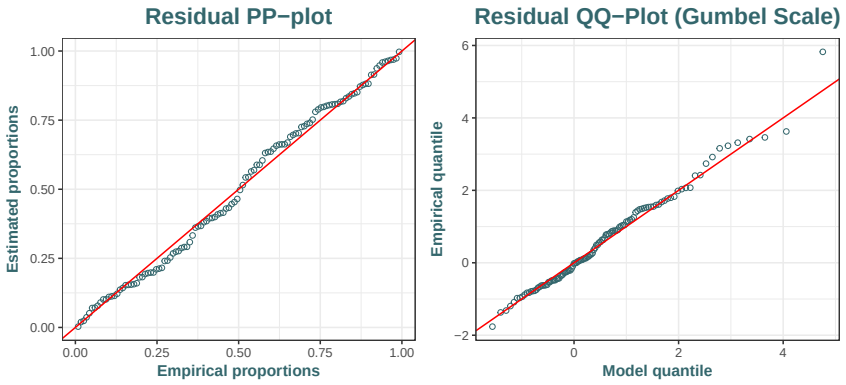

Figure C.10: *(left) Residual probability plot and probability. (right) Residual quantile plot on the Gumbel scale. Both for the nonstationary GEV model allowing for a linear trend in the location parameter and fitted by MLE.*

Appendix D

Github Repository Structure

The Github repository build for this thesis can be found on this address :

`https://github.com/proto4426/PissoortThesis`

where the R package **`PissoortThesis`** is located. It has the following **structure** :

- **/R/** : contains the scripts with all the functions that are made available through the package. The functions are located in the script by "category", i.e.
  - **1UsedFunc.R** : some functions created for the introduction, the stationary analysis of yearly maxima in GEV, analysis in POT, nonstationary analysis in GEV and POT,...
  - **BayesFunc.R** : functions created for the Bayesian Analysis, e.g. the Metropolis-Hastings algorithm and Gibbs Sampler (both stationary and nonstationary).
  - **BootstrapFunc.R** (not updated)
  - **NeuralNetsFunc.R** : functions slightly refined from Cannon [2010] to allow for better outputs for the nonstationary analysis with NN.
  - **runExample.R** : contains the function allowing to run the Shiny applications directly through the package (put the name of the application in ' ' in the function to load the application).

  The documentation of the functions are directly made through the package, by typing `?Function_Name`.

- **/Scripts-R/**
  - **1GEV_plots_(chap1).R** and **1GEV_ggplot_(chap1).R** : contain the plots made for the chapter 1, but only te last latter scripts contain the code to construct the final plots (made with `ggplot2`)
  - **1intro_stationary.R** introduction and preprocessing + descriptive analysis, stationary analysis of yearly maxima in GEV, analysis in POT, analysis with other time scale and with minima, ...
  - **1intro_trends(splines).R**
  - **1intro_stationary.R**

  - **1intro_stationary.R**
  - **1intro_stationary.R**
  - **1intro_stationary.R**
  - **1intro_stationary.R**
  - **1intro_stationary.R**
  - **1intro_stationary.R**

- **/Shiny_app_visu/** (not updated)
- **/data/**
- **/inst/**
- **/man/**
- **/stan/**
- **/vignettes/**

Bibliography

K. H. S. S. S. S. Amir AghaKouchak, David Easterling. *Extremes in a Changing Climate*, volume 65 of *Water Science and Technology Library*. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-4478-3 978-94-007-4479-0. URL http://link.springer.com/10.1007/978-94-007-4479-0.

E. Angelino, M. J. Johnson, and R. P. Adams. Patterns of Scalable Bayesian Inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000052. URL http://www.nowpublishers.com/article/Details/MAL-052.

A. A. Balkema and L. d. Haan. Residual Life Time at Great Age. *The Annals of Probability*, 2(5):792–804, Oct. 1974. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176996548. URL http://projecteuclid.org/euclid.aop/1176996548.

J. Beirlant, J. L. Teugels, and P. Vynckier. *Practical Analysis of Extreme Values*. Leuven University Press, 1996. ISBN 978-90-6186-768-5. Google-Books-ID: ylR3QgAACAAJ.

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, Mar. 2006. ISBN 978-0-470-01237-6. Google-Books-ID: jqmRwfG6aloC.

M. Betancourt. Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. *arXiv:1604.00695 [stat]*, Apr. 2016. URL http://arxiv.org/abs/1604.00695. arXiv: 1604.00695.

M. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:30, 2015.

A. Bolívar, E. Díaz-Francés, J. Ortega, and E. Vilchis. Profile Likelihood Intervals for Quantiles in Extreme Value Distributions. *arXiv preprint arXiv:1005.3573*, 2010. URL http://arxiv.org/abs/1005.3573.

L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

A. J. Cannon. A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, 24(6):673–685, Mar. 2010. ISSN 08856087. doi: 10.1002/hyp.7506. URL http://doi.wiley.com/10.1002/hyp.7506.

A. J. Cannon and I. G. McKendry. A graphical sensitivity analysis for statistical climate models: application to indian monsoon rainfall prediction by artificial neural networks and multiple linear regression models. *International Journal of Climatology*, 22(13):1687–1708, 2002. ISSN 1097-0088. doi: 10.1002/joc.811. URL http://dx.doi.org/10.1002/joc.811.

M. Carney, P. Cunningham, J. Dowling, and C. Lee. Predicting probability distributions for surf height using an ensemble of mixture density networks. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 113–120, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi: 10.1145/1102351.1102366. URL http://doi.acm.org/10.1145/1102351.1102366.

M. Charras-Garrido and P. Lezaud. Extreme Value Analysis : an Introduction. *Journal de la Societe Française de Statistique*, 154(2):pp 66–97, 2013. URL https://hal-enac.archives-ouvertes.fr/hal-00917995.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer London, London, 2001. ISBN 978-1-84996-874-4 978-1-4471-3675-0. URL http://link.springer.com/10.1007/978-1-4471-3675-0.

S. G. Coles and M. J. Dixon. Likelihood-Based Inference for Extreme Value Models. *Extremes*, 2(1):5–23, Mar. 1999. ISSN 1386-1999, 1572-915X. doi: 10.1023/A:1009905222644. URL http://link.springer.com/article/10.1023/A:1009905222644.

C. Cunnane. A note on the Poisson assumption in partial duration series models - Cunnane - 1979 - Water Resources Research - Wiley Online Library, 1979. URL http://onlinelibrary.wiley.com/doi/10.1029/WR015i002p00489/abstract.

A. C. Davison, S. A. Padoan, and M. Ribatet. Statistical Modeling of Spatial Extremes. *Statistical Science*, 27(2):161–186, 2012. ISSN 0883-4237. URL http://www.jstor.org/stable/41714789.

L. de Haan. *On regular variation and its application to the weak convergence of sample extremes*. Mathematisch Centrum, 1970. Google-Books-ID: sQ3vAAAAMAAJ.

L. de Haan and S. Resnick. On asymptotic normality of the hill estimator. *Communications in Statistics. Stochastic Models*, 14(4):849–866, 1998. doi: 10.1080/15326349808807504. URL http://dx.doi.org/10.1080/15326349808807504.

A. L. M. Dekkers and L. D. Haan. On the Estimation of the Extreme-Value Index and Large Quantile Estimation. *The Annals of Statistics*, 17(4):1795–1832, Dec. 1989. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176347396. URL http://projecteuclid.org/euclid.aos/1176347396.

A. L. M. Dekkers, J. H. J. Einmahl, and L. D. Haan. A Moment Estimator for the Index of an Extreme-Value Distribution. *The Annals of Statistics*, 17(4):1833–1855, Dec. 1989. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176347397. URL http://projecteuclid.org/euclid.aos/1176347397.

J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics, Philadelphia, Jan. 1987. ISBN 978-0-89871-364-0.

D. K. Dey and J. Yan. *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press, Jan. 2016. ISBN 978-1-4987-0131-0. Google-Books-ID: PYhUCwAAQBAJ.

B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979. doi: 10.1214/aos/1176344552. URL http://dx.doi.org/10.1214/aos/1176344552.

S. El Adlouni, T. B. M. J. Ouarda, X. Zhang, R. Roy, and B. Bobée. Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, 43(3):W03410, Mar. 2007. ISSN 1944-7973. doi: 10.1029/2005WR004545. URL http://onlinelibrary.wiley.com/doi/10.1029/2005WR004545/abstract.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events: for Insurance and Finance*. Springer Berlin Heidelberg, Feb. 1997. ISBN 978-3-642-08242-9.

M. Falk and F. Marohn. Von Mises Conditions Revisited. *The Annals of Probability*, 21(3):1310–1328, July 1993. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176989120. URL http://projecteuclid.org/euclid.aop/1176989120.

C. A. T. Ferro and J. Segers. Inference for Clusters of Extreme Values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(2):545–556, 2003. ISSN 1369-7412. URL http://www.jstor.org/stable/3647520.

R. A. Fisher and L. H. C. Tippett. Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *ResearchGate*, 24(02):180–190, Jan. 1928. ISSN 1469-8064. doi: 10.1017/S0305004100015681.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press, Nov. 2013. ISBN 978-1-4398-4095-5. Google-Books-ID: ZXL6AQAAQBAJ.

A. Gelman, J. Hwang, and A. Vehtari. Understanding Predictive Information Criteria for Bayesian Models. *Statistics and Computing*, 24(6):997–1016, Nov. 2014. ISSN 0960-3174. doi: 10.1007/s11222-013-9416-2. URL http://dx.doi.org/10.1007/s11222-013-9416-2.

E. Gilleland and R. W. Katz. **extRemes** 2.0: An Extreme Value Analysis Package in *R*. *Journal of Statistical Software*, 72(8), 2016. ISSN 1548-7660. doi: 10.18637/jss.v072.i08. URL http://www.jstatsoft.org/v72/i08/.

B. Gnedenko. Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, 44(3):423–453, 1943. ISSN 0003-486X. doi: 10.2307/1968974. URL http://www.jstor.org/stable/1968974.

J. A. Greenwood, J. M. Landwehr, N. C. Matalas, and J. R. Wallis. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5):1049–1054, Oct. 1979. ISSN 1944-7973. doi: 10.1029/WR015i005p01049. URL http://onlinelibrary.wiley.com/doi/10.1029/WR015i005p01049/abstract.

S. D. Grimshaw. Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution. *Technometrics*, 35(2):185, May 1993. ISSN 00401706. doi: 10.2307/1269663. URL http://www.jstor.org/stable/1269663?origin=crossref.

L. d. Haan. *On regular variation and its application to the weak convergence of sample extremes*. Mathematisch Centrum, 1970. Google-Books-ID: sQ3vAAAAMAAJ.

L. d. Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer series in operations research. Springer, New York ; London, 2006. ISBN 978-0-387-23946-0. OCLC: ocm70173287.

M. Hartmann and R. Ehlers. Bayesian Inference for Generalized Extreme Value Distributions via Hamiltonian Monte Carlo. *Communications in Statistics - Simulation and Computation*, pages 0–0, Mar. 2016. ISSN 0361-0918, 1532-4141. doi: 10.1080/03610918.2016.1152365. URL http://arxiv.org/abs/1410.4534. arXiv: 1410.4534.

T. Hastie and R. Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297–310, 1986. ISSN 0883-4237. URL http://www.jstor.org/stable/2245459.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970. ISSN 0006-3444. URL https://academic.oup.com/biomet/article-abstract/57/1/97/2721936/Monte-Carlo-sampling-methods-using-Markov-chains.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8. URL http://dx.doi.org/10.1016/0893-6080(89)90020-8.

J. R. M. Hosking and J. R. Wallis. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 29(3):339, Aug. 1987. ISSN 00401706. doi: 10.2307/1269343. URL http://www.jstor.org/stable/1269343?origin=crossref.

J. R. M. J. R. M. Hosking and J. R. Wallis. *Regional frequency analysis : an approach based on L-moments*. Cambridge ; New York : Cambridge University Press, 1997. ISBN 0521430453 (hardbound).

W. W. Hsieh and B. Tang. Applying neural network models to prediction and data analysis in meteorology and oceanography., Sep 1998. URL https://open.library.ubc.ca/cIRcle/collections/52383/items/1.0041821.

M. N. Khaliq, T. B. M. J. Ouarda, J.-C. Ondo, P. Gachon, and B. Bobée. Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. *Journal of Hydrology*, 329:534–552, Oct. 2006. doi: 10.1016/j.jhydrol.2006.03.004.

V. V. Kharin and F. W. Zwiers. Estimating extremes in transient climate change simulations. *Journal of Climate*, 18(8):1156–1173, 2005. doi: 10.1175/JCLI3320.1. URL https://doi.org/10.1175/JCLI3320.1.

A. M. G. Klein Tank and Wijngaard. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22(12):1441–1453, Oct. 2002. ISSN 1097-0088. doi: 10.1002/joc.773. URL http://onlinelibrary.wiley.com/doi/10.1002/joc.773/abstract.

A. N. Kolmogorov, N. Morrison, and A. T. Bharucha-Reid. *Foundations of the theory of probability*. Chelsea Publishing Company, New York, 1956. OCLC: 751236060.

M. R. Leadbetter. On extreme values in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 28(4):289–303, Dec. 1974. ISSN 0044-3719, 1432-2064. doi: 10.1007/BF00532947. URL http://link.springer.com/article/10.1007/BF00532947.

M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1983. ISBN 978-1-4612-5451-5 978-1-4612-5449-2. URL http://link.springer.com/10.1007/978-1-4612-5449-2.

F. Leisch. Creating r packages: A tutorial. 2008. URL https://epub.ub.uni-muenchen.de/6175/.

A. A. Lindsey and J. E. Newman. Use of Official Wather Data in Spring Time: Temperature Analysis of an Indiana Phenological Record. *Ecology*, 37(4):812–823, Oct. 1956. ISSN 1939-9170. doi: 10.2307/1933072. URL http://onlinelibrary.wiley.com/doi/10.2307/1933072/abstract.

Y. Liu, A. Gelman, and T. Zheng. Simulation-efficient shortest probability intervals. *Statistics and Computing*, 25(4):809–819, July 2015. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-015-9563-8. URL http://link.springer.com/10.1007/s11222-015-9563-8.

D. J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

J. Maindonald and J. Braun. *Data analysis and graphics using R: an example-based approach*, volume 10. Cambridge University Press, 2006.

G. Marra and S. N. Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74, 2012. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2011.00760.x/full.

E. S. Martins and J. R. Stedinger. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3):737–744, Mar. 2000. ISSN 1944-7973. doi: 10.1029/1999WR900330. URL http://onlinelibrary.wiley.com/doi/10.1029/1999WR900330/abstract.

P. C. D. Milly, J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer. Climate change. Stationarity is dead: whither water management? *Science (New York, N.Y.)*, 319(5863):573–574, Feb. 2008. ISSN 1095-9203. doi: 10.1126/science.1151915.

M. Mudelsee. *Climate Time Series Analysis*, volume 51 of *Atmospheric and Oceanographic Sciences Library*. Springer International Publishing, Cham, 2014. ISBN 978-3-319-04449-1 978-3-319-04450-7. URL http://link.springer.com/10.1007/978-3-319-04450-7.

R. M. Neal and others. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

S. Ni and D. Sun. Noninformative priors and frequentist risks of bayesian estimators of vector-autoregressive models. *Journal of Econometrics*, 115(1):159–197, July 2003. ISSN 0304-4076. doi: 10.1016/S0304-4076(03)00099-X. URL http://www.sciencedirect.com/science/article/pii/S030440760300099X.

P. J. Northrop and N. Attalides. Posterior propriety in Bayesian extreme value analyses using reference priors. *Statistica Sinica*, 26(2), Apr. 2016. ISSN 1017-0405. URL http://dx.doi.org/10.5705/ss.2014.034.

M. C. Peel, Q. J. Wang, R. M. Vogel, and T. A. McMAHON. The utility of L-moment ratio diagrams for selecting a regional probability distribution. *Hydrological Sciences Journal*, 46(1):147–155, 2001. URL http://www.tandfonline.com/doi/abs/10.1080/02626660109492806.

J. Pickands. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1): 119–131, Jan. 1975. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176343003. URL http://projecteuclid.org/euclid.aos/1176343003.

E. C. Pinheiro and S. L. P. Ferrari. A comparative review of generalizations of the Gumbel extreme value distribution with an application to wind speed data. *arXiv:1502.02708 [stat]*, Feb. 2015. URL http://arxiv.org/abs/1502.02708. arXiv: 1502.02708.

R.-D. Reiss and M. Thomas. *Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields ; [includes CD-ROM]*. Birkhäuser, Basel, 3. ed edition, 2007. ISBN 978-3-7643-7230-9 978-3-7643-7399-3. OCLC: 180885018.

S. I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, 1987. ISBN 978-0-387-75952-4 978-0-387-75953-1. URL http://link.springer.com/10.1007/978-0-387-75953-1.

M. Ribatet. A User's Guide to the POT Package (Version 1.4). *month*, 2006. URL http://www.unalmed.edu.co/~ndgirald/Archivos%20Lectura/Archivos%20curso%20Riesgo%20Operativo/POT.pdf.

M. Ribatet, C. Dombry, and M. Oesting. Spatial Extremes and Max-Stable Processes. In *Extreme Value Modeling and Risk Analysis: Methods and Applications*, pages 179–194. Chapman and Hall/CRC, 2015.

P. Ribereau, E. Masiello, and P. Naveau. Skew generalized extreme value distribution: Probability-weighted moments estimation and application to block maxima procedure. *Communications in Statistics - Theory and Methods*, 45(17):5037–5052, Sept. 2016. ISSN 0361-0926. doi: 10.1080/03610926.2014.935434. URL http://dx.doi.org/10.1080/03610926.2014.935434.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, Feb. 1997. ISSN 1050-5164, 2168-8737. doi: 10.1214/aoap/1034625254. URL http://projecteuclid.org/euclid.aoap/1034625254.

G. Rosso. Extreme Value Theory for Time Series using Peak-Over-Threshold method. *arXiv preprint arXiv:1509.01051*, 2015. URL http://arxiv.org/abs/1509.01051.

D. Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric Regression. Cambridge Books, Cambridge University Press, 2003. URL http://econpapers.repec.org/bookchap/cupcbooks/9780521785167.htm.

C. Scarrott and A. MacDonald. A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60, 2012. URL https://www.ine.pt/revstat/pdf/rs120102.pdf.

J. Segers. Generalized Pickands Estimators for the Extreme Value Index: Minimal Asymptotic Variance and Bias Reduction. 2001. URL http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=0C9C7CAE3938CA7A9B280DEA739EFDA9?doi=10.1.1.7.1713.

C. Sherlock, G. Roberts, and others. Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15(3):774–798, 2009. URL http://projecteuclid.org/euclid.bj/1251463281.

V. P. Singh and H. Guo. Parameter estimation for 3-parameter generalized pareto distribution by the principle of maximum entropy (POME). *Hydrological Sciences Journal*, 40(2):165–181, Apr. 1995. ISSN 0262-6667, 2150-3435. doi: 10.1080/02626669509491402. URL http://www.tandfonline.com/doi/abs/10.1080/02626669509491402.

R. L. Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90, Apr. 1985. ISSN 0006-3444. doi: 10.1093/biomet/72.1.67. URL https://academic.oup.com/biomet/article-abstract/72/1/67/242523/Maximum-likelihood-estimation-in-a-class-of.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353/full.

M.-S. Suh, S.-G. Oh, D.-K. Lee, D.-H. Cha, S.-J. Choi, C.-S. Jin, and S.-Y. Hong. Development of New Ensemble Methods Based on the Performance Skills of Regional Climate Models over South Korea. *Journal of Climate*, 25(20):7067–7082, May 2012. ISSN 0894-8755. doi: 10.1175/JCLI-D-11-00457.1. URL http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-11-00457.1.

A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, Aug. 2016. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-016-9696-4. URL http://arxiv.org/abs/1507.04544. arXiv: 1507.04544.

R. Von Mises. La distribution de la plus grande de n valeurs. *Rev., Math, Union Interbalcanique,*, 1: pp.141–160, 1936.

J. L. Wadsworth. Exploiting Structure of Maximum Likelihood Estimators for Extreme Value Threshold Selection. *Technometrics*, 58(1):116–126, Jan. 2016. ISSN 0040-1706. doi: 10.1080/00401706.2014.998345. URL http://dx.doi.org/10.1080/00401706.2014.998345.

S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010. URL http://www.jmlr.org/papers/v11/watanabe10a.html.

R. Yang and J. O. Berger. *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University, 1996. URL http://www.stats.org.uk/priors/noninformative/YangBerger1998.pdf.

C. Zhou. The extent of the maximum likelihood estimator for the extreme value index. *Journal of Multivariate Analysis*, 101(4):971–983, Apr. 2010. ISSN 0047-259X. doi: 10.1016/j.jmva.2009.09.013. URL http://www.sciencedirect.com/science/article/pii/S0047259X0900178X.