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 Kirilliouk

Michel Crucifix

Mémoire présenté en vue de l'obtention du

Master en statistiques, orientation générale

par : **Antoine Pissoort**

Aout 2017

Abstract

This thesis provides an in-depth nonstationary analysis of recent official records of annual maximum temperatures in order to statistically evaluate global warming. Relying on various modeling techniques based upon the Extreme Value Theory, this thesis demonstrates that a nonstationary extreme value Weibull model with a linear model on the location parameter is the preferred model to explain data within a vast number of parametric and nonparametric models. This thesis proves that the scale is not significantly affected. Whilst this lead to consideration that global warming is linear in the location of the annual maximum temperatures, other modeling techniques using splines's derivatives in a generalized additive model have previously shown that the trend on the annual maxima are not simultaneously significant over time. While developing theoretical backgrounds on these state-of-the-art methods, this thesis carefully presents the application of these methods that are gathered into a repository and R package, enabling efficient automated analysis for the future.

Keywords • Extreme Value Theory • Generalized Extreme Value • Generalized Pareto Distribution • Nonstationary analysis • Bayesian inference • Generalized Additive Models • Splines smoothing • Conditional Density Networks • Bootstrapping methods • Parallel computing • R package • Shiny applications

Acknowledgements

I would first like to thank my thesis supervisor Johan Segers for all his help and his guidance during this whole year. These repeated appointments have made this thesis very interesting.

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, but I would like to thank the open source community such as R or LaTeX can benefit. For me and for a great number of students, it has been a non-negligible source of ideas and an effective practical source of learning.

Also quite less usual, but I would like to thank in particular this thesis that permits me to acquire a vast knowledge on various subjects and to develop expertises in the R language.

Finally, I want to thank my family and my friends, but also Bernadette and Michel for their support and all the time I have spent writing in their house 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.

BMA . . . . . Bayesian Model Averaging

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

MC(MC) . . . . . Marko Chain (Monte Carlo)

MH . . . . . Metropolis-Hastings

ML . . . . . Maximum Likelihood

MLE . . . . . Maximum Likelihood Estimator

NN . . . . . Neural Network

PP(D) . . . . . Posterior Predictive (Distribution)

TN . . . . . Temperature miNimum

TX . . . . . Temperature maXimum

Introduction

Extreme Value Theory (EVT) is concerned with the statistical characterization of *extreme* events. This thesis aims at applying the EVT on broad environmental data, in particular with a meteorological application. EVT could also be applied to other areas such as risk analysis, finance or insurance. Although these are relevant on people's lives, in terms of the number of persons impacted and lives lost. Moreover, the increase in weather extremes over the past decade has raised awareness of the importance of this subject.

There have been a number of major drought events recorded in the past few years in Ethiopia, Australia, United Stated, Eurasia, etc. Pakistan, for example, experienced the worst flooding in 80 years resulting in over two thousand casualties and twenty million people affected. The 2003 European heatwave killed approximately seventy thousand people, more than ten times the number of killed people in the September 11 attacks.

Hence, the need for EVT to analyze these extreme events is compelling. The data are provided by an official intitude, the "Institut Royal of Météorologie" (IRM) of Uccle. As the problem facing climate change evidence is often the lack of reliable past data to compare with, this thesis is based upon a homogeneous dataset with annual temperatures from 1901 to 2016.

Following Kharin et al. [2007], this thesis anticipated that climate change affects extreme weather. Aware of the existence of the temporal increase of the annual maxima but having no prior idea on the form of this evolution, the research problem of this thesis can be summarized as :

> To statistically evaluate the nonstationarity of the sequence of annual maxima in order to assess climate warming and predict the extreme events associated with this climate warming.

To this end, we can model these *extremes* using two different approaches :

- as maximum observations that occur inside nonoverlapping *blocks* of equal size. In Chapter **1**, this thesis presents the method of *block-maxima* and derive the Generalized Extreme Value (GEV) distribution; or-
- as observations that *exceed* a certain high *threshold*. This method, presented in Chapter **2**, is often regarded as more efficient as it can use more observations. It is called the *peaks-over-threshold* approach, from which is derived the Generalized Pareto Distribution.

Since this thesis deals with an annual sequence, i.e. maxima over a block of 1 year, the first approach will be given more attention. While the first chapters restrict the analysis to independent sequences, in Chapter **3** the thesis will consider dependent sequences and present tools to model nonstationary sequences. To go beyond the somewhat restrictive parametric inferential methods in EVT, Chapter **3**

will also consider a novel flexible model relying on a Neural Networks framework, in order to assess the nonstationarity with a significantly higher number of possible models with a state-of-the-art method. Chapter **4** will present the Bayesian methods that can be applied for this purpose, and to take the estimation and predictive uncertainty into account.

The second part will apply these methods on the annual maxima in Uccle. Chapter **5** will start with an in-depth introductory trend analysis on the maxima in order to strictly evaluate their evolution over time, both with pointwise and simultaneous confidence intervals. Chapter **6** will present the application of the EVT with the block-maxima approach with a strong emphasis on methods to make a nonstationary GEV analysis. Chapter **7** will conclude the analysis using Bayesian methods, and using newly implemented methods to introduce flexibility to the model.

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. This approach aims at modeling the extremes inside a predefined block. 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 $X_{(n)}$

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

which is also a degenerate distribution.

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

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

$$\Pr\Big\{a_n^{-1}(X_{(n)} - b_n) \le z\Big\} \approx G(z).$$

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

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

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

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

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

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

- 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 density distribution of 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 σ increases, 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 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

[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.

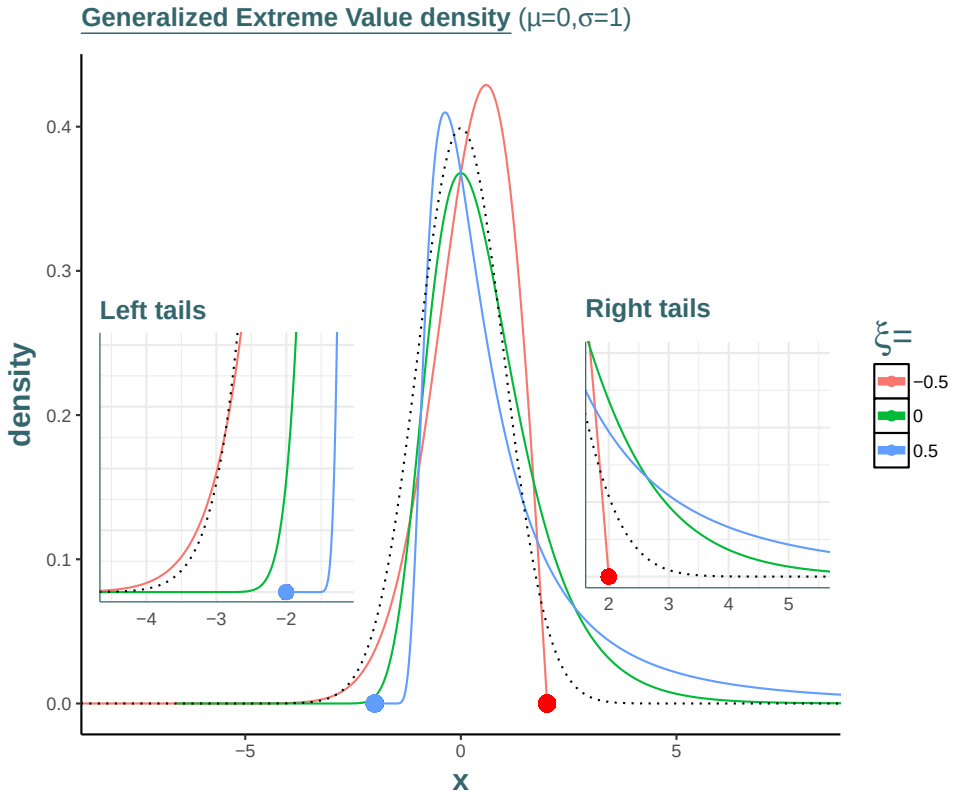


Figure 1.1: *GEV distribution with the normal as benchmark (dotted lines) and a zoom on the parts of interest to better visualize the behaviour in the tails. In green, we retrieve the Gumbel distribution ($\xi = 0$). In red, we retrieve the Weibull-type ($\xi < 0$) while in blue, we get the Fréchet-type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

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 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

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

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 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

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

Then, by setting $b_n = 0$, 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, we see that by setting $a_n = n^{1/\alpha}$ and keeping $b_n = 0$, we have

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

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 for heavy-tailed distribution (see Appendix **A.1**) such as the Pareto distribution, a correction for loca-

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

tion is not necessary to obtain a non-degenerate limiting distribution, see Beirlant et al. [1996, pp.51]. More generally, we can state the 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 helps to understand the shape of the tails of this kind of distributions. We define the concepts of *regularly varying functions*, and *slowly varying functions* in Appendix **A.3**

Convergence to EV Weibull distribution

The **type** **III** or **EV Weibull** family of distributions $G_3(x)$ arises 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 $\neq [0, 1]$. When choosing $a_n = R$ and $b_n = (R - L)/n$, we find the unit Reversed Weibull distribution $We(1, 1)$ in the limit

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

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

Conditions and Comments

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 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**.

It is 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 parameters (μ, σ) will implicitly take the normalization into account, while the shape parameter ξ is not affected. More details on this are in Chapter **3** with methods to estimate (μ, σ, ξ) in Section **1.6**.

1.4 Maximum Domain of Attraction

The preceding results can be 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 only write "*domain of attraction*" (DA) in the following for convenience.

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 $\xi_i = \xi_j$.

At this point, we have all the necessary tools to the pertaining DA. But, before proceeding, it is necessary to point out 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 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 derive their *von Mises conditions*, coming from Von Mises [1936] but revisited in Falk and Marohn [1993]. These conditions are important in practice and 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 + δx given that the subject has "survived" up until 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,*

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

□

More technical characterizations and conditions together with proofs can be found 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 $F \in D(G_1)$. □

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 with tails that are exponentially decaying (light-tailed i.e. the exponential, Gamma, Weibull, logistic, ...) but also distributions that are moderately heavy-tailed such as the lognormal. To see this, consider a Taylor expansion :

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

where "$\sim$" refers to the asymptotic equivalence function. Hence, we 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 theorem informs us that all dfs $F \in D(G_{2,\alpha})$ 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 $F \in D(G_{2,\alpha}) \iff \bar{F} \in R_{-\alpha}$.

Falk and Marohn [1993] present the revisited ***Von Mises condition*** for this DA :

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

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

By setting $K = x_m^{\alpha}$, we obtain $\bar{F}(x) = Kx^{-\alpha}$. Therefore, $a_n = (Kn)^{\alpha^{-1}}$ 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, for example the rainfall process in some tropical zones. Common distributions include Pareto, Cauchy, Burr, etc. Another example is (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 the relation between the Fréchet and the EV Weibull distributions

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

A certain symmetry occurs for these two types (e.g. 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, 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 preferable in an extreme analysis context as it allows for arbitrarily large values, most phenomena are typically bounded, hence we will use EV Weibull for the most attractive and flexible class for modelling extremes. For example, our case of modelling a process of maximum temperatures seems to be a perfect candidate.

1.4.2 Closeness under tail equivalence property

An interesting property of all three types of DA $D(G_k)_{k=1,2,3}$ we have derived, is that of those *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 a "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 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 the block-length is one year -as usually assumed in meteorological analysis-, let 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 on average once 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

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

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 address this issue, it is important to distinguish stationary from nonstationary sequences. Section **3.3** investigates the return periods and return levels more precisely by relaxing the independence assumption (stationary) and then under a climate change environment (nonstationary). In order to check the accuracy of the fitted model, Section **1.7.1** present the *return level plot.*

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, this section presents 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 other methods that are widely used to estimate GEV parameters (probability weighted moment estimator in Section **1.6.2**). There exist estimators for the EVI ξ only, and these will be addressed in Section **A.5**. After all, we will heavily rely on Bayesian inference in Chapter **4**.

1.6.1 Likelihood-based Methods

The most usual inference we will consider is Maximum Likelihood (ML). It is intuitive to understand though its implementation could bring some problems. Depicted by Smith [1985], the potential difficulty when using likelihood methods for 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.

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 see what defines the borders of the problematic case. The 'bell' of the curve becomes very narrow. In the problematic

[4] See paragraph below equation (1.17).

cases, Bayesian inference that 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 use this concept in Section **??** to circumvent issues of these likelihood computations in nonstationary sequences, and to add more flexibility through a deep 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.
$$\ell(\mu, \sigma, \xi = 0\,; \mathbf{z}) = -m\log\sigma - \sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\exp\left\{-\left(\frac{z_i-\mu}{\sigma}\right)\right\}. \tag{1.34}$$

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 poor. *Profile likelihood* inference 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. Note that $\boldsymbol{\theta_j}$ can be of dimension

higher than one (e.g. in a nonstationary context).

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, we 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. 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 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

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

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

A problem for the so-computed standard errors when considering long-range return levels is that they can increase so drastically with the return period such that the confidence intervals of the *return level plot* can become difficult to work with. To address this issue 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 is that confidence intervals computed in the usual method, with standard errors computed by the Delta method in (1.37), relying on the normal approximation, is 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

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

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 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 $\xi > 0$. Figure 6.2 can already help to visualize shape of the plot in our case ($\xi < 0$).

Chapter 2

Peaks-Over-Threshold Method

Contents

This chapter will focus on a 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, numerous techniques have emerged and we will navigate the main ideas.

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

Unfortunately, the resulting analysis of this chapter will not be presented in Part **II** because it would have made the text too voluminous. Empirical results (code and html reports) will remain available in the repository[1](see 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 mostly 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 this is 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.

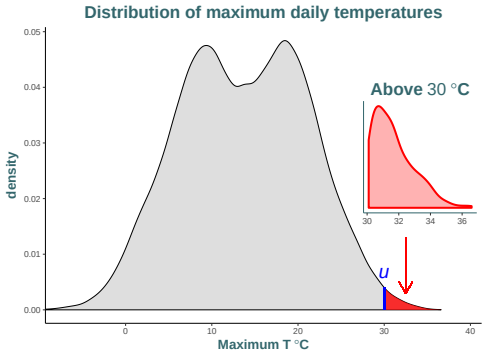


Figure 2.1: *Kernel density estimates for the whole series (grey) and for the series of excess over the threshold of $u = 30°c$ (red). More information on the data will be given in Part **II**.*

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 (2.1) could be computed. But, similarly as in the method of block maxima, F is unknown in practice, and we will still rely on approximations[2]. Theorem **1.1** was used to find an approximate distribution for block maxima; here 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 will 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 on rates of convergence come from Reiss and Thomas [2007, pp.27-28]. 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]. Similarly as for the GEV, the scale parameter σ_u for the GPD differs from the standard deviation since it governs the "size" of the excess, as mentioned in AghaKouchak et al. [2012, 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 note 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 it is invariant. In block

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

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).

Similarly as in the block-maxima approach, there are 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 notices that it corresponds to an **exponential** df, i.e. $Y \sim \exp(\sigma_u^{-1})$ and hence light-tailed.

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

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

  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$(0, \sigma_u)$ following Grimshaw [1993, pp.186].

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

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

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

2.3 Return Levels

Presented in Section **1.5** in the block-maxima approach, return levels are also useful in POT to bring valuable information. However, unlike for block maxima, the quantiles of the GPD cannot be as easily interpreted as return levels because the observations no longer derive from predetermined *blocks* of equal length. Instead, it is necessary 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

Whilst 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 resembles that of GEV in Section **1.6**, and also requires numerical techniques. The two approaches we have encountered so far -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. Appendix **A.5** gives some of the often used methods to estimate the EVI.

Finally, methods dedicated to POT exist to estimate the parameters of the GPD. For example, they 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, whereas 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* :

- *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. 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 such 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 note that $mrl(u)$ is linearly increasing in u, with gradient $\xi \cdot (1-\xi)^{-1}$ and intercept $\sigma_{u_0} \cdot (1-\xi)^{-1}$. 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 gives 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 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 being 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. The Point process approach that has not been 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 :

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

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 as they lead 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 have 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 has made great progress but 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 processes as stated in Milly et al. [2008] to temperature data to demonstrate climate warming, such theoretical assumptions that have been previously made are not sustainable in practice. 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 that will enable us to study the possible increasing behavior of the maximum temperatures. Section **3.3** will redefine return levels under those less restricted cases, and finally Section **3.4** will introduce a new flexible way to model nonstationary extremes under a redefined Multi Layer Perceptron framework with inferential techniques minimizing the risk 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 $X_{(n)} = \max_{1\leq i\leq n} X_i$ 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 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. Typical ingredients that define stationarity are the trend and the seasonality. In this text, we will be interested by the trend. Other formulations of *stationarity* exist but we will use 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 it 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 compared to this of an independent series.

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 $\theta = 1$ leads 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 results of Theorem **3.2** would not hold.

Formally, it can be defined in the POT approach as

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

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 clusters is increased. This problem of temporal dependence makes inference based on the likelihood invalid. Two methods 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 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 if the goal of the application is to analyze 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 $\xi^* = \xi$ will induce that the two distributions G^* and G 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 more generally in Section **3.3**, including 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^* remains 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 by making sure the model fit is convincing with the diagnostic tools presented in Section **1.7**.

3.2 Non-Stationary Extremes

Whereas the previous Section relaxed the first "i" of the "iid" assumption made during Chapter **1** and **2** by allowing temporal dependence under certain conditions leading to a stationary process, this section will now tackle the "**id**" part, i.e. the 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 of 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, 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 climate warming. We will hence focus more 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 theories a for nonstationary processes and so we will now use a pragmatic approach of combining standard EV models with statistical modeling. We will only discuss the approach in block maxima; extrapolation to POT is straightforward for inference, model comparison and diagnostics.

3.2.1 Block-Maxima

By considering yearly blocks, we face nonstationary concerns for the trend which could probably be imputed to Global Warming. The evidence of seasonality arising when we decrease the blocks' length is an issue in block maxima. For example if one considers daily maximum temperatures, seasonality will be present. On the other hand, we lose information by taking yearly blocks only as we do not use all the information. Indeed, at least one half of the daily temperatures from October to March will not be used since they will never include days that will be an annual maximum. To overcome this issue, the dataset could be divided and different models can be applied on sub-blocks of the data, for example on 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 with 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 the nonstationary parameters $\mu(t), \sigma(t), \xi(t)$ 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, β denote the complete vector of parameters, and X is the design matrix. In our example, Z_t will describe the annual maximum temperature of year t for $m = 116$ years. As already stated, it is not 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 $\mu(t), \sigma(t), \xi(t)$ evalu-

ated at z_t, the log-likelihood of the model (3.7) is, provided $\xi(t) \neq 0 \; \forall t$,

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

where the notation $y_+(t) = \max\{y(t), 0\}$ holds for all t. The parameters $\mu(t), \sigma(t)$ and $\xi(t)$ 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, which is the difference of complexity of two models. In the second method, it is directly included in the criterion since their partial derivative with p is positive for both criterion(3.11) that are to minimize.

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 relevant models in Table 6.6 in the following Section by means of BIC and AIC_c.

Model Diagnostics

When the best model is selected, it is still necessary to assess that the model fits well the data so we can infer conclusions about some aspects of the population. We will use tools seen in the independent (or stationary) case, i.e. the quantile and the probability plots presented in Appendix **A.4**. From the inhomogeneous distribution across years, the data needs to be standardized. For instance, 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* :

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

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

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

The choice of the Gumbel as reference distribution can be discussed but this is a reasonable choice regarding its place in EVT. Figure 6.3 in Appendix **C** will present such plots.

3.3 Return Levels : Definitions

Section **1.5** defined return levels for independent sequences, and so 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. By recalling $F(r_m) = \Pr\{X_{(n),y} \leq r_m\} = 1 - 1/m$, we write

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

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 :** 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 can be seen as 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 that $N \sim \text{Bin}(m, 1/m)$ and hence N has an expected value of $m \cdot m^{-1} = 1$.

Non-stationarity

Mathematical derivations go beyond the scope of this thesis, however, as demonstrated in Amir AghaKouchak [2013, Section 4.2], we can retrieve the same two interpretations of return period as for a stationary process. Moreover, from the definition of non-stationary processes, as parameter(s) are a function of time, return levels will also change over time. This will have a big impact on modeling, since an inappropriate model will lead to inappropriate return levels. We will see an example of these nonstationary return levels in Figure 6.4. 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, as they cannot address the possibly more complex temporal relationships with the 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

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

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 us 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 processes 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 not overfit the data. Another pitfall is the 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 the GEV 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 sit in the gray area between a *parametric* and a *nonparametric* model, in the sense that they assume a GEV distribution from the output layer defined by the three parameters of interest (see Figure 3.1 below), while they allows for great flexibility coming from the hidden layers which lead to think that these are 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 will not impact the modeling.

3.4.1 Generalized Maximum Likelihood

In Section **1.6.1** we introduced the concept of penalized likelihood. The issue 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 only the most probable values of the parameter are included, excluding the divergent cases. This method extends 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*

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

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 MLE'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 :

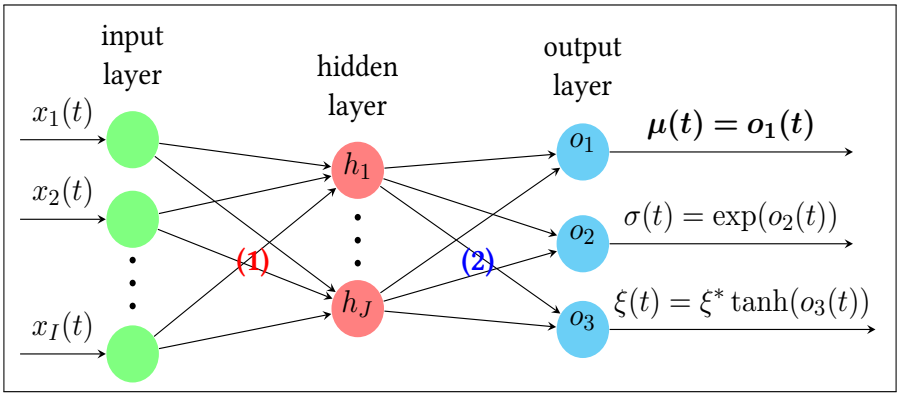


Figure 3.1: *General framework of the fully-connected nonstationary GEV-CDN based on Cannon [2010]. The input layer will be the time itself in our application, i.e. $x_i(t) = t,\ \forall i = 1, \ldots, I$ but it can be other covariates. The hidden layer represent additional complexity incorporated in 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 (i.e., $m(x) = x$) 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 reduces 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, increasing the size of original data cannot improve the model predictive force, but can 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 allows a reduction in 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 should 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 the original sample is iid. There exists a large panel of different bootstrap procedures to construct confidence intervals. For example, we will follow the steps of Khaliq et al. [2006] for the residual bootstrap :

1. Fit a nonstationary GEV model to the data.

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

2. Transform residuals from the fitted model to be identically distributed :

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

   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 :

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

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 *residual* bootstrap significantly outperformed the *parametric* bootstrap in the GEV-CDN framework (and we also will verify it in, 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 upcoming chapter will introduce other methods to construct intervals in a strictly Bayesian approach relying also on a Monte Carlo sampling.

Chapter 4

Bayesian Extreme Value Theory

Contents

Whilst sometimes criticized for the introduction of some subjectivity, the Bayesian paradigm is increasingly adopted by practitioners for its benefits. In this case, Bayesian inference is interesting to perform an EV analysis and to explore this framework. This chapter focuses on GEV analysis on annual maxima. Extension to POT or Point Processes is straightforward with tools provided in the text or in the code.

Section **4.1** introduces the Bayesian paradigm and presents its characteristics. Section **4.2** presents an essential component of the Bayesian sphere, the priors that we will describe for GEV applications. Section **4.3** discusses several Monte Carlo algorithms that can be used to obtain a sample from the posterior. Then, Section **4.4** presents diagnostics to assess the convergence of the algorithms in order to check i the reliability of the posterior's sample. If this is positive, Section **4.5** provides the Bayesian techniques to make inferences. Section **4.6** presents the posterior predictive distribution thats allows to handle predictions. Then, Section **4.7** discusses tools for model's selection in a Bayesian setting. Finally, Section **4.8** introduces tools to validate models through information criteria. Useful tools regarding basic Bayesian inference are left in Appendix **B**.

This chapter is mostly based on Coles [2001, sec. 9.1], Beirlant et al. [2006, chap.11], Dey and Yan [2016, chap.13] and Amir AghaKouchak [2013, chap.3], and other relevant articles.

4.1 Preliminaries : Motivations

This chapter will rely on a multidimensional set of parameters θ of a GEV model, i.e. $\theta' = (\mu, \sigma, \xi)$ in a stationary context, where $\nu = \log \sigma$ is often used instead of σ to eliminate the positivity constraint. Unless stated differently, we will write θ as this 3D vector to facilitate readability.

We already argued that likelihood-based methods have desirable properties for inference, and hence we have adopted in Chapter **1** and **3** the methods of (generalized or penalized) maximum likelihood to compute our estimators. Bayesian inference introduces a flexible alternative that also relies on likelihood. First, we present the most fundamental result for Bayesian inference which is a direct interpretation of the Bayes' Theorem.

Definition 4.1 (Posterior distribution). *Let $\boldsymbol{x} = (x_1, \dots, x_n)$ denote the vector of n independent observations of a random variable X with density $f(x|\theta)$, and let $\pi(\theta)$ denote the density of the prior distribution for θ. According to Bayes' theorem, the **posterior distribution** of θ is*

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

where $L(\theta|\boldsymbol{x}) = \prod_i f(x_i|\theta)$ is the likelihood since the x_i's are independent. △

This result (4.1) provides a probabilistic framework to convert an initial set of beliefs about θ, represented by the prior $\pi(\theta)$, into a posterior distribution that will be convenient to use for inference. Relying on the posterior distribution to compute an estimate $\hat{\theta}$ of θ provides interesting characteristics :

- Whenever possible, it introduces an other source of knowledge coming from the domain, by the elicitation of a prior (see Section **4.2**). However, it also introduces subjectiveness and antagonists contend that, since different analysts would specify different priors, all conclusions become meaninglessly subjective.
- Accounting for uncertainty is handled easily in the Bayesian paradigm relying on distribution's properties of $\pi(\theta|\boldsymbol{x})$, for example its variance. Now, the parameters are treated as random variables. It also permits an elegant way of making future predictions, an important issue in EVT.
- Bayesian framework can overcome the regularity conditions of the likelihood inference (see Section **1.6.1**) by providing a viable alternative in cases when ML breaks down. We will see that the annual maximum temperatures are not far from the problematic cases ($\xi < -0.5$). Moreover, Bayesian confidence intervals (i.e. *credible* intervals, see Section **4.5.2**) do not require asymptotic theory.

In EVT, Bayesian inference is a wide expanding domain that has nearly infinite possibilities. With strong probabilistic foundations, this method now strongly relies on computational capabilities (see Section **4.3**).

4.2 Prior Elicitation

The amount of information that can be retrieved from the prior is sometimes viewed as the greatest strength of Bayesian inference and this is also sometimes viewed as its main pitfall due to the un-

quantifiable subjectivity that is introduced. However, the construction of the prior is at the center of Bayesian analysis since they are necessary in the to compute (4.1). To make this viewed as the least subjective, it requires the legitimate statement of an expert.

Priors may not be of importance if the number n of observations is large. From (4.1) we see that the amount of information contained in the data through $L(\theta|\boldsymbol{x})$ will be prominent compared to that contained in the prior $\pi(\theta)$. However, this configuration is rare in EVT where we are using small constructed datasets. For this reason, incorporating external information through the prior should be rigorously studied. This type of prior is called *informative priors*, while those which give most weight to data are called *non-informative* (or *objective*) *priors*. We focus on priors that are widely applied in EVT.

4.2.1 Trivariate Normal Distribution

The trivariate normal prior distribution on $\theta' = (\mu, \nu, \xi)$ leads to the following prior density :

$$\pi(\theta) \propto \sigma^{-1} \cdot \exp\Big\{ -\frac{1}{2}(\theta - \boldsymbol{m})'\Sigma^{-1}(\theta - \boldsymbol{m})\Big\}, \tag{4.2}$$

where the mean vector $\boldsymbol{m}$ and the symmetric positive definite $[3\times 3]$ covariance matrix Σ must be specified. This approach was used by Coles and Powell [1996] but other parametrizations exist. For example, changing μ to $\log \mu$ can be useful if a physical lower bound for this parameter must be specified.

4.2.2 Gamma Distributions for Quantile Differences

This method constructs priors on the quantile space, for fixed probabilities. Let $\Pr(X > q_p) = p$, where X has a GEV distribution (1.12). Then,

$$q_p = \mu + \sigma \cdot \xi^{-1} \cdot (x_p^{-\xi} - 1),$$

where $x_p = -\log(1-p)$. Indeed, we notice that it is a m-return level (1.32) with $p = m^{-1}$. The prior distribution is constructed in terms of the quantiles $(q_{p_1}, q_{p_2}, q_{p_3})$ for specified probabilities $p_1 > p_2 > p_3$. It is easier to work with the differences $(\tilde{q}_{p_1}, \tilde{q}_{p_2}, \tilde{q}_{p_3})$, with $\tilde{q}_{p_i} - \tilde{q}_{p_{i-1}}$, $i = 1, 2, 3$, where q_{p_0} is the physical lower endpoint of the process variable. We can take priors on the quantile differences to be independent with

$$\tilde{q}_{p_i} \sim \text{Gamma}(\alpha_i, \beta_i), \qquad \alpha_i, \beta_i > 0, \quad \text{for } i = 1, 2, 3.$$

The differences $(\tilde{q}_{p_2}, \tilde{q}_{p_3})$ only depend on (σ, ξ). Therefore, prior information on μ arises only through $\tilde{q}_{p_1}$. Hyperparameters $(\alpha_1, \alpha_2, \alpha_3)$ and $(\beta_1, \beta_2, \beta_2)$ and probabilities $p_1 >_2> p_3$ must all be specified. Constructed by Coles and Tawn [1996], this leads to

$$\pi(\theta) \propto J(\theta) \prod_{i=1}^{3} \tilde{q}_{p_i}^{\alpha_i - 1} \exp\big\{ -\tilde{q}_{p_i} \cdot \beta_i^{-1}\big\}, \qquad q_{p_1} < q_{p_2} < q_{p_3}, \tag{4.3}$$

where $J(\theta)$ is the Jacobian of the transformation from $(q_{p_1}, q_{p_2}, q_{p_3})$ to θ, i.e.

$$J(\theta) = \sigma \cdot \xi^{-2} \cdot \left| \sum_{\substack{i,j \in (1,2,3) \\ i<j}} (-1)^{i+j} (x_i x_j)^{-\xi} \log\left\{ x_j \cdot x_i^{-1} \right\} \right|, \tag{4.4}$$

where $x_i = -\log(1 - p_i)$ for $i = 1, 2, 3$.

4.2.3 Beta Distributions for Probability Ratios

This method is the opposite of the last one, in the sense that it constructs priors on the probability space, for fixed quantiles. Let $\Pr(X > q) = p_q$, where $X \sim \text{GEV}(\theta)$ with $\theta = (\mu, \sigma, \xi)$. Then,

$$p_q = 1 - \exp\left\{ -\left[1 + \xi \cdot \sigma^{-1}(q - \mu)\right]_+^{-\xi^{-1}} \right\}.$$

A prior can be defined in terms of $(p_{q_1}, p_{q_2}, p_{q_3})$ for $q_1 < q_2 < q_3$, with $p_{q_0} = 1$ and $p_{q_4} = 0$. Since $p_{q_1} > p_{q_2} > p_{q_3}$ it easier to work with $(\tilde{p}_{q_1}, \tilde{p}_{q_2}, \tilde{p}_{q_3})$, where $\tilde{p}_{q_i} = p_{q_i}/p_{q_{i-1}}$ for $i = 1, 2, 3$. We can take

$$\tilde{p}_{q_i} \overset{\perp\!\!\!\perp}{\sim} \text{Beta}\left(\sum_{j=i+1}^{4} \alpha_j, \alpha_i \right), \qquad i = 1, 2, 3. \tag{4.5}$$

where $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ are positive hyperparameters. This construction from Crowder [1992] leads to

$$\pi(\theta) \propto J(\theta) \prod_{i=1}^{4} (p_{q_{i-1}} - p_{q_i})^{\alpha_i - 1}, \quad p_{q_1} > p_{q_2} > p_{q_3} \text{ and } 1 + \xi\sigma^{-1}(q_i - \mu) > 0 \; \forall i = 1, 2, 3.$$

$J(\theta)$ is the Jacobian of the transformation from $(p_{q_1}, p_{q_2}, p_{q_3})$ to θ. It can be shown that

$$J(\theta) = \sigma \cdot \xi^{-2} \cdot \prod_{i=1}^{3} g(q_i) \cdot \left| \sum_{\substack{i,j \in (1,2,3) \\ i<j}} (-1)^{i+j} (x_i x_j)^{-\xi} \log\left\{ x_j \cdot x_i^{-1} \right\} \right|, \tag{4.6}$$

where $x_i = -\log(1 - p_{q_i})$ for $i = 1, 2, 3$ and $g(q_i) = x_i^{1+\xi} \cdot e^{-x_i} \cdot \sigma^{-1}$ is the density of the GEV. Specific example to construct this prior in practice can be found in Dey and Yan [2016, pp.272].

4.2.4 Non-informative Priors

To receive correct and acceptable advice from an expert is often difficult. In many cases, it is not possible to inject information through the prior. Hence, priors must be constructed to represent this lack of knowledge and not influence posterior inferences.

Parameters of the prior distribution are often called *tuning parameters* or *hyperparameters* to emphasize that it is possible to tune the amount of information provided through the prior. In practical applications with Markov Chains (see Section **4.3**), these values are often tuned to maximize the convergence of the posterior to its stationary target distribution. However, adjustments of these priors must always be thought of in practical applications since there can be multiple objectives. For example, if we take $\pi(\theta) \sim \mathcal{N}_d(\mu, \Sigma)$, the hyperparameter Σ will be tuned to represent the amount of knowledge

we want to incorporate. It will tend to infinity to be non-informative (vague prior). However, note that it can be a huge help with computation to have less diffuse priors, even if they are not very informative and will not have a noticeable impact on the posterior.

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

Jeffrey's prior

Discovered by Jeffreys [1961], this prior is specified as

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \qquad \text{where} \quad I_{ij}(\theta) = \mathbb{E}_\theta\left[-\frac{\partial^2 \log f(X|\theta)}{\partial\theta_i \partial\theta_j}\right];\ \ i,j = 1,\dots,d. \tag{4.7}$$

This prior is a standard starting rule for an objective analysis. This prior is invariant to reparametrization, but has a complex form for GEV models, and it exists only when $\xi > -0.5$ in GEV models, where it is function of ξ and σ only.

Maximal Data Information prior

Maximal Data Information (MDI) priors are defined to provide maximal average data information on θ. These are not invariant under reparametrization but are easy to implement. For GEV models, it is defined as

$$\pi(\theta) = \exp\Big\{\mathbb{E}\big[\log f(X|\theta)\big]\Big\} \propto \sigma^{-1}\exp\Big\{-\gamma(1)\cdot(1+\xi)\Big\}.$$

where $\gamma(1)$ denotes the Euler's constant. 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, and hence 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 preferred alternative is often to construct uninformative priors by using proper priors which are near flat, e.g. which are uniform or which exhibits very large variance. Taking a trivariate normal distribution as prior as in Section **4.2.1** is often difficult as it involves 9 hyperparameters in total (μ, ξ and 7 in the $[3\times 3]$ symmetric matrix Σ). In GEV, we will often take univariate independent normally distributed priors each with a large (tuned) variance. When these variances increase, we get at the limit

$$\pi(\theta) = \pi(\mu,\nu,\xi) \overset{(\perp\!\!\!\perp)}{=} \pi(\mu)\cdot\pi(\nu)\cdot\pi(\xi) \ \propto 1, \tag{4.8}$$

where $\nu = \log\sigma$.

4.3 Bayesian Computation : Markov Chains

The main difficulty of Bayesian inference has been the computation of the normalizing integral in the denominator of (4.1), which is often intractable. The prior $\pi(\theta)$ could be chosen so that the integral could be evaluated analytically, but this is exceptional. Moreover, it is not possible in EV models since they involve distributions that are not within the exponential family. More complex cases where the likelihood $L(\theta|\boldsymbol{x})$ is intractable can be dealt with approximate Bayesian computation (Beaumont et al. [2002]), but we will not consider this subject in this text.

Hence, simulation-based methods have been developed to sample from an arbitrary posterior distributions $\pi(\theta|\boldsymbol{x})$. Simulations of N iid samples $\theta_1, \theta_2, \ldots, \theta_N$ 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 we need *Markov Chain Monte Carlo* (MCMC) techniques.

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, i.e.*

$$\Pr\{\theta_{t+1} = k_{t+1} \mid \theta_t = k_t,\ \theta_{t-1} = k_{t-1}, \ldots\} = \Pr\{\theta_{t+1} = k_{t+1} \mid \theta_t = k_t\}. \tag{4.9}$$

In other words, given θ_t, we have that θ_{t+1} is independent of all the states prior to t. △

This in one of the most important results that defines MCMC techniques (see Angelino et al. [2016, section 2.2.3] for more results). We use this techniques to simulate a Markov chain $\theta_1, \theta_2, \ldots, \theta_N$ that converges to the target distribution $\pi(\theta|\boldsymbol{x})$. This means that, after some *burn-in period* B, the chain $\theta_{B+1}, \theta_{B+2}, \ldots, \theta_N$ can be treated as random sample from $\pi(\theta|\boldsymbol{x})$ that has reached its target **stationary** distribution. This will be achieved if the so-generated chain is

1. *aperiodic*, i.e. it does not have period η for any $\eta > 1$.
2. *irreducible* or *ergodic*, i.e. we can get from any state to any other state (possibly in several steps).

4.3.1 Metropolis–Hastings

Named after Metropolis et al. [1953] and Hastings [1970] , the *Metropolis–Hastings* algorithm is the pioneering MCMC algorithm for Bayesian analysis. It is a conceptually simply method with specific update proposals that are accepted or rejected according to an appropriate rule. Details on the algorithm are in Appendix **B.1.1**. We can summarize the *pros* and *cons* of this algorithm :

- *PROS* : Very easy to program and works for relatively complex densities.
- *CONS* : Can be inefficient, in the sense that it will require many iterations before the stationary target distribution will be reached.

This algorithm requires some tuning through the proposal distribution. It is recommended to target an acceptance rate of ≈ 0.25 since all components of θ are updated simultaneously, in order to facilitate convergence, following Bédard [2008]. The idea is that if the variance of the proposal distribution is too large, most proposals will be rejected, i.e. jumps through the chain will be too large, and the converse if the variance of the proposal distribution is too small.

4.3.2 Gibbs Sampler

Originated from Geman and Geman [1984], the *Gibbs Sampler* can be seen as a special case of the MH algorithm. Suppose our parametric vector θ can be divided into d subvectors $(\theta_1, \ldots, \theta_d)$, and suppose that each of these subvectors represent a single parameter, i.e. (μ, σ, ξ) for the stationary case, so that $d = 3$ in this model. For each $t = 1, \ldots, N$, the Gibbs sampler samples the subvectors $\theta_t^{(j)}$ conditionally on both the data $\boldsymbol{x}$ and the remaining subvectors at their current values. More details on the algorithm are left in Appendix **B.1.2**.

It is recommended to target an acceptance rate of ≈ 0.25 when all components of θ are updated simultaneously, and ≈ 0.45 when the components are updated one at a time, that is when θ_j is univariate (see e.g. Gelman et al. [2013, chap.11]), so that the so-generated Markov-chain has desirable properties.

We must also note the increase of complexity of this sampler compared to the Metropolis-Hastings, since the nested loop in Algorithm 3 implies that there are d iterations with each simulation. Finally, we summarize the *pros* and *cons* of this sampler :

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

Note that it is possible to combine Gibbs sampling with some Metropolis steps as in [Gelman et al., 1995, chap.11], when a conditional distribution cannot be identified.

4.3.3 Hamiltonian Monte Carlo

Betancourt [2016] emphasizes that the performance of the MCMC 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 [2017] said "guess-and-check strategy of Random Walk Metropolis is doomed to fail in high dimensional spaces where there are an exponential number of directions in which to guess but only a singular number of directions that stay within the typical set and pass the check (...)". This is the *curse of dimensionality* issue since there is much more volume outside any given neighborhood than inside of it. In an EV framework where the sample size is limited and the dimension can easily increase with nonstationary models, it is necessary to consider other forms of sampling to address computational efficiency. Hartmann and Ehlers [2016] have demonstrated in a GEV application that HMC (and Riemann manifold HMC) are more computationally efficient than traditional MCMC algorithms such as MH.

Following Stan Development Team [2017, chap.32], the *Hamiltonian Monte Carlo* (HMC) algorithm starts at specified or random initial values for θ. Then, it uses derivatives of the density function being sampled to generate efficient transitions spanning the posterior. It uses an approximate Hamiltonian dynamics simulation based on numerical integration. Finally, a Metropolis acceptance step is applied, and a decision is made whether to update to the new state or keep the existing state. HMC permits to better exploit properties of the target distribution to make informed jumps through neighborhoods of high probability, while avoiding neighborhoods of low probability. We leave interesting details of this algorithm in Appendix **B.1.3**.

This new MC method is build from differential geometry and topics such as *Riemannian manifolds*, *microcanonical Disintegration*, *kinetic energy*, etc. Hence,this thesis does aim to theoretically explaining all this algorithm. In trying to implement it we faced difficulty to build the (nonstationary) GEV model in STAN[1]. This was probably due to a problem of misspelled constraints in the boundaries of the GEV distribution, since these are function of parameters in the EV-Weibull family. More information is in Section **7**. The *pros* and *cons* of this algorithm used within STAN are :

- *PROS* : Easy to program as we only 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. The theory of which can be very difficult to understand. There is less control over the sampler (but maybe it is for the best?).

Pros and cons that are specific to compiled languages must also be noted.

4.4 Convergence Diagnostics

When applying MCMC algorithms to estimate posterior distributions, it is vital to assess convergence of the algorithm to be confident that we have reached the stationary target distribution. Below enumerates some key steps for the convergence in order to obtain reliable results.

1. A sufficient *burn-in period* $B < N$ must be chosen to ensure that the convergence to the stationary posterior distribution has occurred, as it will eliminate the influence of starting values. For the same reason, a sufficient number of simulations N is required to ensure accurate MCMC estimates.
2. Several dispersed starting values must be simulated to ensure all the regions of high probability of the parameter space have been explored. This is particularly important when the target distribution is complex. The chains must then have good mixing properties. A common technique is to run different chains several times with different starting values, and then combine a proportion of each chain (typically 50%) to get the final chain.
3. A good choice of the proposal is crucial since a poorly chosen distribution may drastically reduce the speed of the algorithm. To a certain extent, the proposal should be similar to the target distribution. Our preference is for Gaussian proposals for their convenience and because they are "easy-to-tune", but note that fat-tailed distributions could be interesting as they will occasionally generate a candidate far from the previous one, reducing the risk of being stuck in a local mode.
4. The *autocorrelations* within single parameter's chains must be small to keep a good effective sample size. This is the same for *cross-correlations*, i.e. the correlations between the chains of different parameters in the vector θ. Both impacts the speed of convergence of the algorithm.

Some important diagnostics are defined in Appendix **B.2**. We must keep in mind that no convergence diagnostics can prove that convergence really happened and validate the results. However, a combined use of several relevant diagnostics will be required to increase our confidence that convergence actually happened.

[1]High level probabilistic programming language which incorporates HMC with Nuts.

4.5 Bayesian Inference

Assuming the posterior has been correctly explored using a MCMC algorithm, leaving us with N realizations $\theta_1, \dots, \theta_N$ from the posterior. Since θ is d-variate, marginal estimates are simply obtained by considering the realizations $\theta_1^{(j)}, \dots, \theta_N^{(j)}$ from the marginal posterior of the parameter $\theta^{(j)}$. It is possible to obtain point-estimates from the posterior, for example by computing the marginal mean or other quantiles from $\theta_1^{(j)}, \dots, \theta_N^{(j)}$. It is also possible to derive a modal estimate by selecting the parameter vector corresponding to the largest posterior value.

4.5.1 Distribution of Quantiles : Return Levels

The Markov chains generated can be transformed to estimate quantities of interest such as quantiles or return levels. When F is GEV, the distribution of quantiles can be retrieved in the same manner as in the frequentist setting in (1.32) with $y_m = -\log(1-m^{-1})$. Let $r_m = q_p(\theta)$ and let $F(q_p) = 1-p$. Then, for each p, the samples $\theta_1, \dots, \theta_N$ can be substituted into (1.32) to yield $q_p(\theta_1), \dots, q_p(\theta_N)$. These values can be used to estimate features of the prior and posterior distributions of $q_p(\theta)$ in the same way that the values $\theta_1, \dots, \theta_N$ were used to estimates features of $\pi(\theta|\boldsymbol{x})$.

4.5.2 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. A 95% Bayesian credible intervals contains 95% of the posterior probability for θ.

There exist two kinds of credible intervals in the Bayesian framework :

- The *quantile-based* credible intervals or *equal-tailed interval* picks an interval to ensure that the probability of being below this interval is as likely as of being above it. For some posterior distribution which are not symmetric, this could be misleading, and is then not recommended in this case. The advantage is that they are easily obtained from a random sample of the posterior.

- The *Highest Posterior Density* (HPD) interval 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, we calculate :

$$\int_{\xi_0}^{\xi_1} \pi(\xi|\boldsymbol{x}) \cdot d\xi = 0.95 \qquad \text{with} \qquad \pi(\xi_0|\boldsymbol{x}) = \pi(\xi_1|\boldsymbol{x}). \tag{4.10}$$

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

For symmetric densities, HPD and central interals are the same while HPD is shorter for asymmetric densities, see e.g. illustrations in Liu et al. [2015].

4.6 Posterior Predictive Distribution

Together with parameter estimation, prediction is the ultimate aim of inference in EVT, and it is neatly handled within the Bayesian paradigm. It permits a straightforward quantification of the uncertainty associated. It is made through the *posterior predictive distribution*

Definition 4.3 (Posterior Predictive Density)**.** *Let $\tilde{X}$ denotes a future observation (e.g. $n+1$) with density $f(\tilde{x}|\theta)$. We define the **posterior predictive density** of $\tilde{X}$ given $\boldsymbol{x}$ as*

$$\begin{aligned} f(\tilde{x}|\boldsymbol{x}) &= \int_{\Theta} f(\tilde{x}, \theta|\boldsymbol{x}) \cdot d\theta = \int_{\Theta} f(\tilde{x}|\theta) \cdot \pi(\theta|\boldsymbol{x}) \cdot d\theta \\ &:= \mathbb{E}_{\theta|\boldsymbol{x}}\big[f(\tilde{x}|\theta)\big], \end{aligned} \tag{4.11}$$

with $f(\tilde{x}|\theta, \boldsymbol{x}) = f(\tilde{x}|\theta)$, where $\boldsymbol{x}$ is the n-dimensional vector of observations. △

The second line of (4.11) emphasizes that we can evaluate $f(\tilde{x}|\boldsymbol{x})$ by averaging over all possible parameter values. Uncertainty in the model is reflected through $\pi(\theta|\boldsymbol{x})$, and uncertainty due to variability in future observations through $f(\tilde{x}|\theta)$.

Definition 4.4 (Posterior Predictive Distribution)**.** *The **posterior predictive distribution** of a future observation $\tilde{X}$ is*

$$\begin{aligned} \Pr\{\tilde{X} < x \mid \boldsymbol{x}\} &= \int_{\Theta} \Pr\{\tilde{X} < x \mid \theta\} \cdot \pi(\theta|\boldsymbol{x}) \cdot d\theta \\ &= \mathbb{E}_{\theta|\boldsymbol{x}}\big[\Pr(\tilde{X} < x \mid \theta)\big]. \end{aligned} \tag{4.12}$$

△

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

$$\Pr\{\tilde{X} < x \mid \boldsymbol{x}\} \approx r^{-1} \sum_{i=1}^{r} \Pr\{\tilde{X} < x \mid \theta_i\}, \tag{4.13}$$

where $\Pr\{\tilde{X} < x \mid \theta_i\}$ follows directly from $f(x|\theta)$.

4.7 Model Comparison

In a nonstationary GEV context, we would like to compare models in a similar way as in the frequentist setting in Chapter **3.2.1** with the deviance statistic, the BIC, etc. There may exist different nonstationary models for the location or for the scale, and we would like to assess which model is the most relevant to describe the data. Although the posterior distribution quantifies estimation uncertainties, it remains conditional on the modeling assumptions. And as in any other estimation approach, it is a vital task to select the best possible model and validate it, otherwise the underlying inferences could be drastically different.

Let assume that $\mathcal{M}_1, \ldots, \mathcal{M}_q$ is a set of q candidate models to describe $\boldsymbol{x}$. Each $\mathcal{M}_i$ uses a parameter vector $\theta^{\mathcal{M}_i}$ whose dimension may differ across models. Let $\Pr(\mathcal{M}_1), \ldots, \Pr(\mathcal{M}_q)$ denote prior probabilities assigned to each model such that $\sum_{i=1}^{q} \Pr(\mathcal{M}_i) = 1$. It is then possible to compute models' posterior probabilities given the observations, from the discrete form of the Bayes' Theorem :

$$\Pr(\mathcal{M}_i|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|\mathcal{M}_i)\cdot\Pr(\mathcal{M}_i)}{\sum_{j=1}^{q}\Pr(\boldsymbol{x}|\mathcal{M}_j)\cdot\Pr(\mathcal{M}_j)}, \tag{4.14}$$

where $\Pr(\boldsymbol{x}|\mathcal{M}_i)$ is the marginal likelihood of observations an can be computed as follows :

$$\Pr(\boldsymbol{x}|\mathcal{M}_i) = \int \Pr(\boldsymbol{x}|\theta^{\mathcal{M}_i},\mathcal{M}_i)\cdot\Pr(\theta^{\mathcal{M}_i}|\mathcal{M}_i)\cdot d\theta^{\mathcal{M}_i}. \tag{4.15}$$

However, computing the marginal likelihood is difficult and suffers from the curse of dimensionality. Bos [2002] provides various methods to compute the marginal likelihood when the size of the parameter vector is moderate. Model posterior probabilities (4.14) can be used for model comparisons and predictions.

The Bayes Factor

One approach is to use the *Bayes factor* to compare models. It measures the relevance of one model compared to another. For two models $\mathcal{M}_i$ and $\mathcal{M}_j$, it is defined as :

$$B_{i,j} = \frac{\Pr(\mathcal{M}_i|\boldsymbol{x})}{\Pr(\mathcal{M}_j|\boldsymbol{x})}\cdot\left(\frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)}\right)^{-1} = \frac{\Pr(\boldsymbol{x}|\mathcal{M}_i)}{\Pr(\boldsymbol{x}|\mathcal{M}_j)}. \tag{4.16}$$

High values of $B_{i,j}$ give stronger confidence in model $\mathcal{M}_i$. The Bayes factor balances quality of fit against model complexity. Formal guidelines have been developed (see e.g. Kass and Raftery [1995]). It is also possible to compare several models via a composite Bayes factor.

Bayesian Model Averaging

Instead of choosing between models, it is also possible to perform multi-model predictions by computing a weighted average of individual model predictions weighted by the posterior model probabilities. *Bayesian model averaging* (BMA) accounts for model uncertainty. Interesting similarities can be found with the bagging ensemble method discussed in Chapter **3.4.3** which makes an average of bootstrapped models to improve the prediction accuracy. Both rely on the same data but BMA uses different models, and hence correlations between predictions of the different models will be important. It is particularly useful when distinct models provide an acceptable description of the data, but yield different predictions.

4.8 Bayesian Predictive Accuracy for Model Checking

If we have a large amount of data, we can use validation techniques by dividing the dataset between a training set (e.g. containing 75% of the whole set), and a test set containing the remaining observations. For example, having N draws $\theta_1,\theta_2,\ldots,\theta_N$ from the posterior $\pi(\theta|\boldsymbol{x}_{\text{train}})$, we can score each value using

$$\log\left[N^{-1}\sum_{t=1}^{N} f(x^*|\theta_t)\right],$$

and that can be summed for any value x^* in $\boldsymbol{x}_{\text{test}}$. This can be used to compare models, with larger values indicating better models. However, we often do not have large amounts of data in an EVT context. Henceforth, we can rather use *cross-validation* techniques, but this is a computationally demanding method. There exists several variants, and Vehtari et al. [2016] provide an excellent summary.

Or, one could rather prefer information criteria that avoid these computations. This approach is to use a penalized loss function of the form

$$\log f(\boldsymbol{x}|\bar{\theta}) - p^*, \qquad \text{with } \theta_1, \dots, \theta_N \text{ drawn from } \pi(\theta|\boldsymbol{x}), \tag{4.17}$$

where p^* represents the effective number of parameters and $\bar{\theta}$ is the posterior mean. Since $\log f(\boldsymbol{x}|\bar{\theta})$ is the density of the observations given the posterior mean of the parameters, it indicates a good fit. We will see two existing methods that use this idea.

1) Deviance Information Criterion

The *Deviance Information Criterion* (DIC) was first proposed by Spiegelhalter et al. [2002] and use the following to compute the effective number of parameters :

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

Since the DIC is defined on the deviance scale, we have

$$\text{DIC} = 2 \log f(\boldsymbol{x}|\bar{\theta}) - \frac{4}{N} \sum_{t=1}^{N} \log f(\boldsymbol{x}|\theta_t). \tag{4.18}$$

Hence, smaller DIC values indicate better predictive models. It can be computed simply by storing $f(\boldsymbol{x}|\theta_t)$ at the end of each iteration in the main loop of the MCMC algorithm. DIC handles uncertainty in inferences within each model, whilst not depending on models' aspects that don't affect inferences within each model. Despite a poor theoretical foundation, DIC is shown to be a good approximation of the penalized loss function (4.17), and is valid only when the effective number of parameters in the model is much smaller than the number of independent observations.

2) Widely Applicable Information Criterion

The *Widely Applicable Information Criterion* (WAIC) is a more recent approach proposed by Watanabe [2010], giving the criterion

$$\begin{aligned} \text{WAIC} &:= 2 \sum_{i=1}^{n} \left[\log \left\{ \mathbb{E}_{\theta|x} f(x_i|\theta) \right\} \right] - \mathbb{E}_{\theta|x} \log f(x_i|\theta) \\ &= \sum_{i=1}^{n} \left[2 \log \left(N^{-1} \sum_{t=1}^{N} f(x_i|\theta_t) \right) - \frac{4}{N} \sum_{t=1}^{N} \log f(x_i|\theta_t) \right]. \end{aligned} \tag{4.19}$$

As the DIC, smaller values indicate better predictive accuracy. It can be calculated by computing $\log f(x_i|\theta_t)\ \forall i = 1, \dots, N$ at the end of each iteration in the main loop of the MCMC algorithm. **?** argued that, asymptotically, the WAIC is equivalent to the leave-one-out predictive fit.

Any predictive accuracy measure involves two definitions, following **?**:

- The choice of what part of the model to label as *the likelihood*, which is directly connected to which potential replications are being considered for out-of-sample prediction.
- The factorization of the likelihood into *data points*, which is reflected in the later calculations of expected log predictive density.

This involves some confusion and the real interpretations of the predictive criteria is still a bit unclear. There exists other interesting methods, as proposed by Gelman et al. [2014] who give in-depth developments of the predictive information criteria.

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 have already showed the distribution of this data in Figure 2.1 to introduce the POT approach. This chapter will present an introductory analysis of the *annual* maxima of these data, which will provide a convenient GEV modeling. We will then use several techniques to analyze the 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, this chapter briefly present the shiny applications created to enhance visualizations. Section **5.2** will present data and its source. Section **5.3** 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 the significance of the trend with correction for simultaneous tests and provide the first aspects in the issue of Climate Warming without using extreme models.

This chapter will 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 **5.3.3**.

5.1 Repository for the code : R Package

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

`https://github.com/proto4426/PissoortThesis`

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

For the reader's convenience, an external folder has been created in the above repository containing all the scripts created during this thesis. It allows reproduction of the results often with more details, tables or plots. It is located in the **/Scripts-R/** folder of the repository, but note that each script will be mentioned in its corresponding section. Further details on the repository structure can be found in Appendix **D**.

Visualization Tool : Shiny Application

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

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

and by choosing one applications's name between the displayed propositions and write it inside (' '). So far, we have

- ('GEV_distributions') : application based on Figure 1.1 smoothly displaying the GEV and the influence of its parameters.
- ('trend_models') : application dedicated to annual maxima that can be visualized with some preliminary methods (see Section **5.3.2**).
- ('splines_draws') : application that simulates splines of a GAM model to visualize the difference in coverages between pointwise and simultaneous confidence intervals. (see Section **5.3.3**).
- ('neural_networks') : application for GEV`cdn` allowing bagging to demonstrate the effects of all the (hyper)parameters, since these choices can be subjective...

A dynamic overview of the applications is available in the repository's README.

All-in-one : Dashboard Still for the reader's convenience, all applications have been placed in a smooth *dahsboard*. Moreover, it is uploaded on a server at the following URL[1] in order to provide a quick user-friendly overview through the `rsconnect` interface.

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/`

5.2 Presentation of the Analysis : Temperatures from Uccle

Data used in this thesis comes from the "Institut Royal de Météorologie" (IRM) in Uccle in order to have reliable data provided by Belgian meteorologists. IRM provided a set of databases this thesis focuses on temperature analysis in order to assess climate warming.

The fact that the dataset begins only at year 1901 is due to homogeneity reasons since the measurement shelters evolved substantially compared to previous periods, especially in terms of measurement 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 helpful advice from C.Tricot, climatologist at the IRM, temperatures in closed shelters have the advantage of not being overly influenced by solar radiation which artificially increases temperatures, especially in periods of extreme heats and high solar activity, and therefore in the studied case. For example, the maximum temperature of $36.6°c$ that occurred on 27th June 1947 in closed shelters was 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 analyze this data for reliability reasons as mentioned above.

Afterwards, we compared these two datasets for years 1901 to 2001. Note 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. This gives confidence that the public dataset has temperatures measured in open shelters, which is not recommended. Hence, we conclude that large measurement errors can easily occur in unofficial data emphasizing the necessity of reliable data in order to yield a reliable analysis.

5.3 First Analysis : Annual Maxima

5.3.1 Descriptive Analysis

The goal of this section is to display a global overview of the data. The code provided most of the descriptive analysis. To summarize most of the information, we have 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 season to highlight the differences. For example, we can see that the spread is higher for autumn and spring.

We will use the block-maxima approach by taking yearly blocks leaving us with $n = 116$ years of data, which seems justifiable for further GEV analysis and the convergence to hold. We will discuss the choice of the block length further in the next Chapter.

5.3.2 First visualization with simple models

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

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

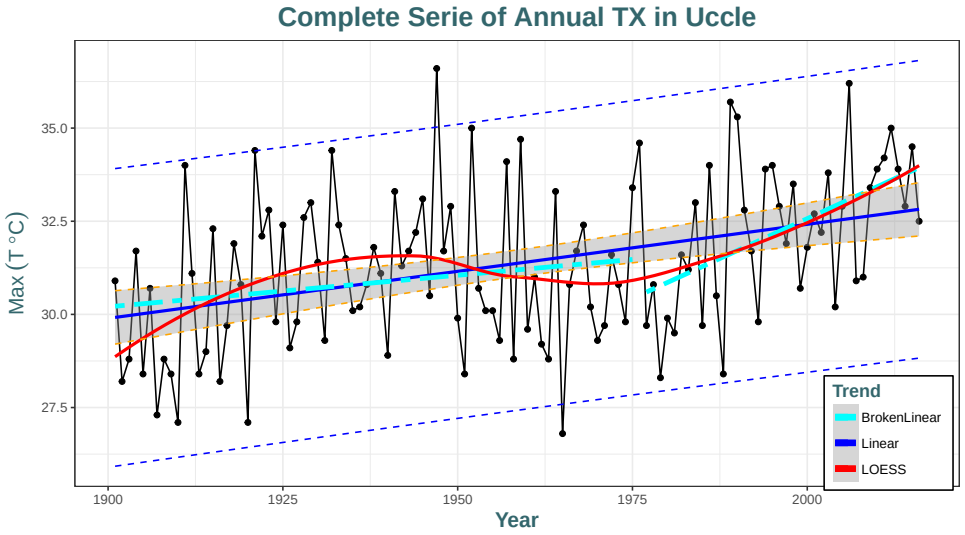


Figure 5.1: *Yearly maxima together with three first models that represent the trend. Note the shaded grey area around the linear regression fit is* 95% *pointwise confidence interval on the fitted values (i.e.,* $\pm 1.96 \times \sigma_{pred}(year)$*) while blue dotted lines are prediction intervals taking into account the prediction uncertainty.*

- *Linear regression* is a **parametric** fit. Note that it is slightly but significantly increasing over time (p-value $\approx 10^{-5}$). From this model, one rough deduction is that each year we expect the annual maximum temperature to increase by $\hat{b} = 0.025°c$ (with $\hat{\sigma}_{\hat{b}} = 0.005$).
- *Local Polynomial regression* or LOESS is a **nonparametric** fit which smooth the data. 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. 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 trend difference between the period [1901-1975] and [1976-2016]. These two periods have been chosen arbitrarily and we note that period [1901-1952] has also a very high positive slope. We will study that in more details in the 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 aim of this thesis rather we are interested in assessing the trend and its statistical significance over time using more sophisticated methods.

5.3.3 Trend Analysis : Splines derivatives in GAM

Apart from the significant pointwise increasing linear trend, we did not find concrete results. Additionally, we see in the series in Figure 5.1 that a linear trend to the entire series could be too restrictive. The LOESS or the broken-linear trend highlight that the trend is not constant over time. We would like here to assess more formally this difference in the slope of the trend over time.

This section assumes reader knowledge of *Generalized Additive Models* (GAM) developed by Hastie and Tibshirani [1986] and on *penalized splines*. Theoretical explanations of these concepts can be found

in Ruppert et al. [2003, chapter 3, 6 and 11], 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 compares coverage properties of intervals.

Pointwise or Simultaneous confidence intervals ?

For this analysis, we deem important to differentiate the two types of intervals for a better understanding of the actual meaning of "confidence interval". A special example is the grey area around the linear fit in Figure 5.1 which represents a 95% interval for the regression line, taking into account only variability in the data. For instance, if we repeatedly sample 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

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

In the linear regression example of Figure 5.1, from the pointwise confidence intervals we can say that, $f(1920)$ has 95% chance to lie within $[29.8, \ 31.1]$ and $f(2016)$ has also 95% to lie within $[32.1, \ 33.6]$ but it would be false to say that both are contained in these intervals at the same time with 95% confidence. The corresponding simultaneous intervals (blue dotted lines) are quite larger, $[26.4, \ 34.3]$ and $[36.8, \ 28.8]$ for years 1920 and 2016 respectively. 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 have fitted a GAM model on the annual maxima relying on the `mgcv` package from Maindonald and Braun [2006].

- Figure C.3 in Appendix **C** shows the correlation structure of the serial normalized residuals. As the 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, therefore 2 additional degrees of freedom are sufficient. Results are 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 validated our choice. Diagnostics of the model presented in Figure C.4 in Appendix **C** are correct.

- 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 significantly.

- 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 **4.5.2** 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.

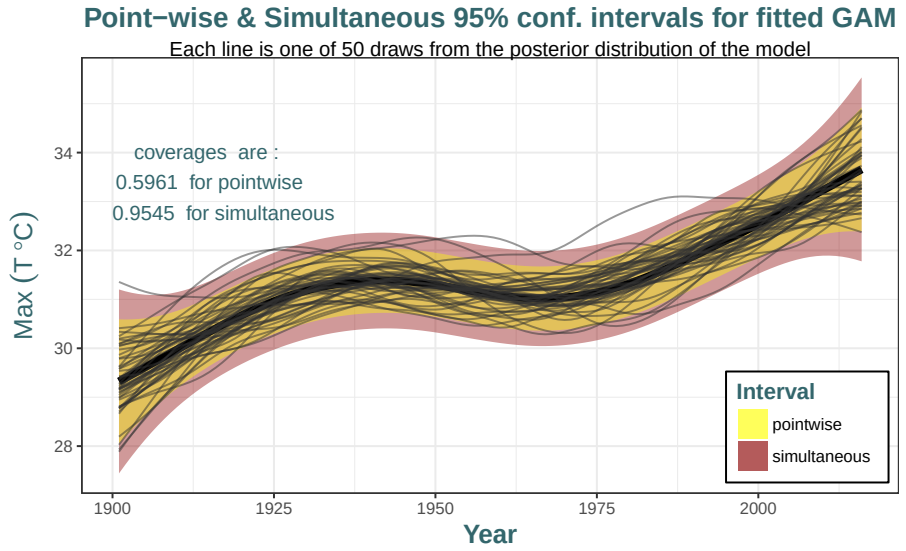


Figure 5.2: *Draws from the posterior distribution of the model (5.3). Notice the large uncertainty associated to the posterior draws (grey lines). The coverages are calculated for $M = 10^4$ simulations.*

Coverage Analysis

To highlight the inadequacy of the pointwise intervals, we computed posterior draws of the fitted GAM and we counted how many draws lied within each interval.

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

A Shiny application has been build from this graph for better visualization of the impact of the number of simulations on the confidence intervals and how coverage varies, both visually and quantitatively. Finally, 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 themselves.

Final Results

We see two significant increases in the trend fitted by GAM in Figure C.5 in Appendix **C** when we do not use simultaneous intervals. The decrease pointed out in Figure 5.1 with LOESS between 1945 and 1970 is hence not significant and probably due to randomness.

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

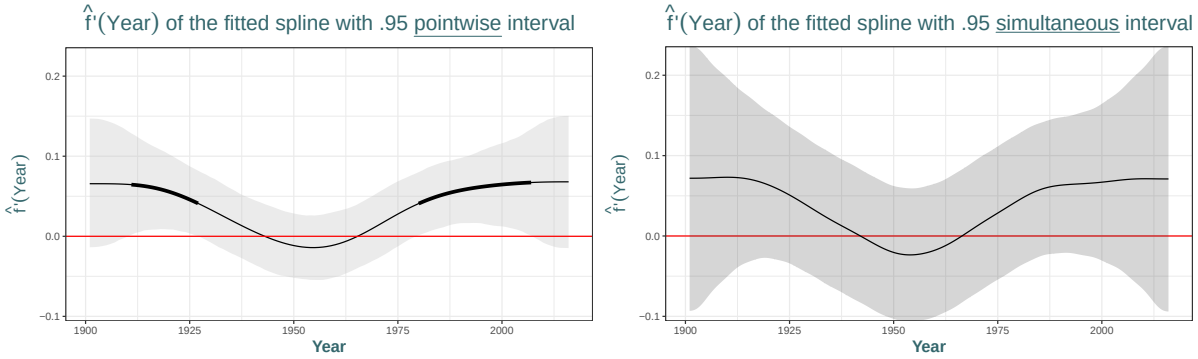


Figure 5.3: *Plots of the first derivative $f^{'}_{(20)}(year)$ of the estimated splines on the retained GAM model. Grey area represents 95% confidence bands. Sections of the spline where the confidence interval does not include zero are indicated by thicker lines.*

- Together with the series in Figure 5.1, we can make the link between the trend behavior of the series and the splines' first derivatives which accurately models the slope behavior of the trend.
- Note the increasing trend with decreasing slope, i.e. $f^{''}_{(20)}(\text{year}) < 0$, until year ≈ 1945 where it becomes negative. The slope then increases and the trend starts to increase again near 1962. This upward trend in the series of annual maxima for this last period brings light to the climate change we are currently facing. However, we see that we are now at a "critical point" for several years, i.e. $f^{''}_{(20)}(\text{year}) \approx 0$, meaning that the trend is likely linear.
- Whilst the pointwise confidence interval includes significant regions, i.e. 0 is not included in the interval, the simultaneous interval that is accounting for the increase in uncertainty has no significant regions. Therefore, we cannot conclude that there has been any significant increase in the annual maximum temperatures in any period.

5.4 Comments and Structure of the Analysis

We have found that there is an upward (linear) trend for the series of yearly maxima, but it is not significant once corrected for simultaneous intervals. Hence, a nonstationary analysis is worth continuing with. After this introductory analysis, we will continue with the specific subject of this thesis,

the extreme value analysis. Hence, next Chapter will present a stationary GEV analysis and then allow nonstationarity in various ways. Bayesian analysis will finally try to make additional improvements in Chapter **7**, e.g. with quantification of uncertainty or in computational efficiency.

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 for ease of reading. Henceforth, we will focus on the GEV analysis only. Note that the POT analysis is available on the repository presented above (or see Appendix **D**).

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

Chapter 6

Analysis in Block Maxima

Contents

This chapter introduces the core subject of this thesis, the stationary and nonstationary GEV analysis for the annual maximum temperatures in Uccle. Concepts borrowed from Chapter **1** and Chapter **3** will be used. POT analysis of Chapter **2** will be mentioned but all results are kept in the code.

In Section **6.1**, we will present GEV inferences assuming stationarity which relies on `1intro_stationary.R` code from the **/Scripts-R/** folder of the repository. Section **6.2** will introduce the nonstationary analysis with parametric models and relies on `2Nonstationary.R` code, while Section **6.3** allows for more complex models from deep architectures and relies on `2NeuralsNets.R` code.

Although some Bayesian concepts will be used in this section such as use of *prior distributions* for the EVI to limit its domain when estimating it with a likelihood-based method, or to regularize weights of Neural Networks in Section **6.3**, we will still call all these methods as "frequentists". A strict Bayesian analysis will be made in next Chapter **7** and comparisons will be provided.

Block-length

The block-length selection is crucial in a GEV 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, or else a large bias in the estimates could occur. Since a large block-length implies less data to work with and thus a large variance of the estimates, a compromise must be found as noted in Chapter **1**. Here, we chose yearly blocks, justifiable for the above reason and for their interpretability and ease of use. Note that this will cause wastage of data since we will not be using 6 months (from October to March) in the analysis since it did not have an annual maximum. One way to prevent this wastage could be by fitting several GEV models on some subsets of the dataset divided by a seasonal criterion, but this approach would generate other issues.

R packages in EVT

A bunch of packages exist for modeling extreme values in R. We have explored and used most of them to do the following analysis, and made some comparisons. 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`, ...

Since a number of packages are doing the same analysis but with different methods, we decided to rely mostly on `ismev` as it is the package used in the book of Coles [2001].

6.1 Inference of the Stationary Model

Chapter **1** is important to understand the concepts used in this section but 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

We estimate the MLE's relying on packages cited above, but also by checking it manually, that is by numerically solving the minimization of the negative log-likelihood. This can be done with the `nlm` routine using the Newton-Raphson algorithm of Dennis and Schnabel [1987]. This algorithm is based on an approximation of the log-likelihood by a quadratic function which is the second order Taylor series approximation of the log-likelihood for a given point. Estimates and standard errors are shown in Table 6.1.

Table 6.1: *MLE's of the three GEV parameters assuming an independent (or rather stationary) context.*

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

Table 6.1 the negative value of the **shape** parameter, meaning 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 which has a finite

right endpoint. A likelihood ratio test will confirm this in Table 6.4, comparing the fitted EV-Weibull with a Gumbel distribution.

The Weibull-type implies that the fitted distribution has an estimated right endpoint of $\hat{x}_* = \hat{\mu} - \hat{\sigma} \cdot \hat{\xi}^{-1} = 38.77°c$. Comparing this value with the maximum value of the annual series ($36.6°c$) shows that this model takes into account the uncertainty since there are only 116 years of data. Hence, it allows return levels, i.e. quantiles of the fitted distribution, to go beyond $36.6°c$, for very long return periods. This will be highlighted in Figure 6.2 (right plot) where we notice a probability mass beyond the minimum and the maximum values of the series from the fitted model.

Profile log-likelihood intervals These 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 left in Appendix **C**. These intervals are constructed by searching for the horizontal line and then subtracting the maximum log-likelihood by half the corresponding upper quantile of the χ^2_{df} for $\text{df} = 1$ parameter of interest. We notice that :

- Even at 99% confidence, the interval for $\hat{\xi}$ does not contain 0 supporting 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 many asymetries. In fact, these will be more relevant for return levels as we will see in the next Section.

Probability-Weighted-Moments

It is always 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: *Stationary GEV parameters estimated by PWM.*

	Location μ	Scale σ	Shape ξ
Estimates	30.552	2.115	$\mathbf{-0.232}$

We immediately see that these results are very close to the MLE's of Table 6.1, in particular for the EVI ξ. This gives us confidence that we are indeed under a Weibull-type GEV model. For convenience and for their properties, we will only keep the MLE's to work with in the following.

6.1.1 Return Levels

First presented in Section **1.5**, return levels are appreciated by the practitioners for inference in an environmental EVT 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 in Table 6.3.

Table 6.3 demonstrates that, for example, the estimated 100-year return level is $36.23°c$ which is the temperature that will be exceeded on average once every 100 years. However, note that very long term extrapolation should be tempered by caution since we only have 116 years of data and predicting

Table 6.3: *The m-year return level estimates and 95% intervals. Last line computes the length's differences between the normal and the profile likelihood intervals.*

	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

far beyond this value is unreliable. We can 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 see blue dots going higher than red lines for higher return periods. We also see that the profile likelihood intervals are more precise for small return periods, i.e. approximately below half the total number of annual data, and then profile likelihood intervals become wider than normal intervals. This illustrates how profile likelihood intervals better 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 which, surprisingly, are not used by default in EV packages such as `ismev`.

6.1.2 Diagnostics

Section **1.7** provided the tools to check the accuracy of the fitted model. The goal here 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.

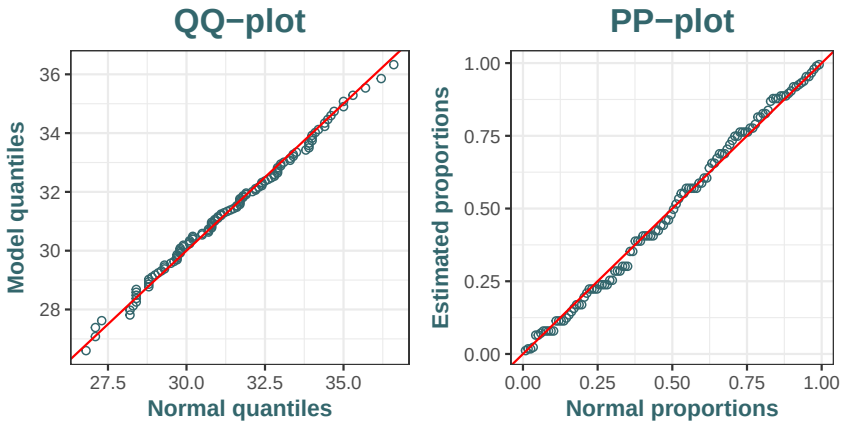


Figure 6.1: *Quantile (left) and probability (right) plots for the stationary GEV model fitted by MLE.*

Both plots show points lying very close to the unit diagonal, showing that the empirical df is very close to the fitted model and hence putting confidence that our model fits our data accurately. The right plot in Figure 6.2 has a similar interpretation and leads to the same conclusion, although the fitted model has a higher peak.

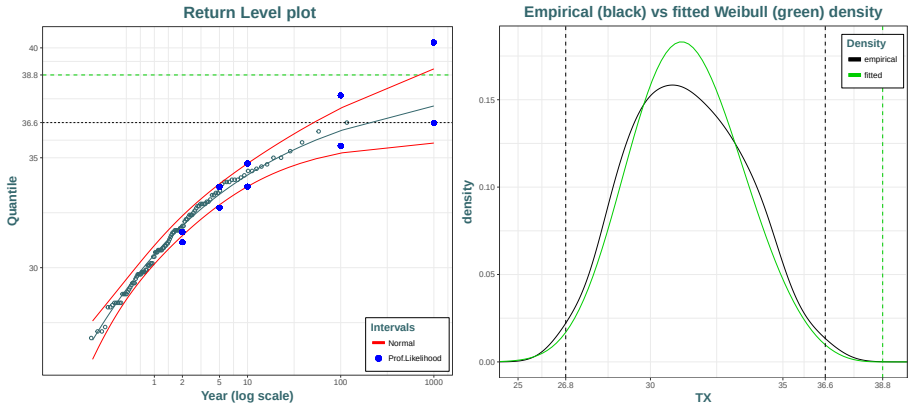


Figure 6.2: *(Left) Return level plot with red lines representing normal confidence intervals, and blue points the individual profile likelihood intervals for return levels. The horizontal dotted line represents the right endpoint of the fitted model in green and the maximum of the series in black. (Right) kernel density in black compared with the density of the fitted model in green, with black dotted lines representing the* ***endpoints*** *of the empirical distribution, and green doted lines still represent the same right endpoint of the fitted EV-Weibull. A Gaussian kernel with a bandwidth calculated by the Silverman [1986, pp.48, (3.31)] "rule of thumb" has been used.*

Return Level Plot

Another tool that is available in EVT to check the fit of a model is the return level plot. It allows for comparisons of observations with return levels coming from the EV-Weibull model fitted by MLE. The left plot in 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 model since $\xi < 0$. Note that all points are very close to the estimated return level and hence we have confidence that our model is suitable. Moreover, all these points are inside the normal confidence intervals (recall that profiled 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 of the fitted model. Again, this is justified since for such far periods, both intervals are allowed to go beyond the domain of the fitted model, with profile likelihood ones taking more uncertainty into account.

Profile Likelihood Intervals for Return Levels

Figure C.7 left in Appendix **C** represents the profile log-likelihood intervals for three return periods, at the x-intersections between the blue line and the curve. We can see the asymmetries (positive skew) which are increasing at higher values of the return period. This was expected since the data at hand provides increasingly weaker information about high return levels of the process. We also displayed the return levels from Table 6.3 represented by green lines on the plots in Figure C.7, and which were computed relying on another method from the package `extRemes`. It is interesting to see how the results can be slightly different.

6.1.3 Stationary Analysis

Section **3.1** proved that a dependent sequence can still have the GEV distribution in the limit of normalized maxima. This will only induce different location and scale parameters compared to the sequence as if it were independent. We can visualize dependence in the series for example with the autocorrelation functions. Corresponding plots are shown in Figure C.8 let in Appendix **C** where we see that temporal dependence is light, but present. Actually, estimates shown in Table 6.1 implicitly take this dependence into account.

POT

Dependence is not a big concern for GEV, however it is more problematic for POT. Indeed, by only analyzing data that are above a threshold (say $30°c$), serial dependence in the data can be strong and observations will have the tendency to occur in clusters. We illustrated this in Figure C.9 in Appendix **C** which highlights this dependence with red lines corresponding to the historic heat waves of summers 1911 and 1976 in Uccle. Indeed, observations lying on the red lines are serially correlated since they occurred during a same period of extreme heat. It is not difficult to assume that *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 get an idea of 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 around 2. The next step (not displayed here) is to decluster the series.

6.2 Parametric Nonstationary Analysis

As depicted in the introductory Figure 5.1, even the assumption of stationarity is likely to be poor for our data. Whereas the oscillatory behavior caught by the LOESS model is probably due to noise rather than a true characteristic of the underlying process, the increasing trend is more alarming. Indeed, our flexible modeling of the trend in Section **5.3.3** confirmed that the trend is statistically significant for two periods when doing pointwise comparisons but it is inadequate since we proved that this method leads to coverages that did not match with the assumed confidence level. When controlling for simultaneous tests, significance of the trend completely disappeared. However, one could argue that these intervals are very large (looking back at Figures 5.2 or 5.3) and thus not precise. Hence, results are not clear regarding this trend and a nonstationary GEV analysis is worth undertaking.

6.2.1 Comparing Different Models

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

1. *Gumbel* : most simple EV-model with unrestricted domain and only 2 parameters as $\xi = 0$.
2. *stationary* : EV-Weibull model fitted in Table 6.1 by MLE.

3. *linear in* $\mu(t)$: allowing a nonstationary location parameter with $\mu(t) = \beta_0 + \beta_1 \cdot t$.

4. *quadratic in* $\mu(t)$: allowing a nonstationary location parameter with $\mu(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2$.

5. *linear in* $\mu(t)$ *and* $\sigma(t)$: same as Model 3 regarding the location but with nonstationary scale $\sigma(t) = \exp(\alpha_0 + \alpha_1 \cdot t)$. The use of the inverse link $b(\cdot) = \exp(\cdot)$ is to ensure positivity of σ $\forall t$.

6. *cubic in* $\mu(t)$: nonstationary location parameter with $\mu(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3$. Note that computational singularities occurred for the likelihood computation of this model (or more complex ones). This came from a problem of solving the hessian to compute the covariance matrix of the parameters since there were too much linear dependencies in the hessian. We solved this by decreasing the tolerance, but this shows us that a linear model with so many polynomial terms is not recommended.

Section **3.2** explains that it is not recommended to model ξ as a smooth function of time, since the data provides little information about the shape of the distribution relative to the location and the scale. The p-values from the sequential pairwise comparisons have been computed with respect to the best retained model. Results are shown in Table 6.4.

Table 6.4: *Comparisons of nested GEV models with nonstationary parameters. Significant p-values at level* 5% *are shown in bold.*

Model	ℓ	df	p-value
Gumbel	-256.84	2	
stationary	-251.75	3	$\mathbf{0.14}\%$
linear in $\mu(t)$	$\mathbf{-241.81}$	**4**	$\mathbf{0.001}\%$
quadratic in $\mu(t)$	-241.48	5	42%
linear in $\mu(t)$ and $\sigma(t)$	-241.69	5	63%
cubic in $\mu(t)$	-241.37	6	65%

The model that is chosen by this procedure in Table 6.4 is Model 3 which allows a linear model for the location parameter. The choice is clearly emphasized by likelihood ratio tests and hence we have confidence in a nonstationary GEV model with a linearly increasing trend from the location. In the next Section we will allow for more flexibility in the model's parameters.

6.2.2 Selected Model : Diagnostics and Inference

Having selected the best model, we will now give estimates of the new parameters before assessing the accuracy of this new model in order to provide reliable inferences through return levels.

Model

Table 6.5 shows the parameters and standard errors for the selected nonstationary GEV model estimated by MLE.

We notice that ξ is in the same range as thr previous stationary model. The parameter β_1 says that the location parameter $\mu(t)$ is expected to increase by 0.025 each year, starting at $\beta_0 = 29.13$ until

Table 6.5: *MLE's of the nonstationary GEV parameters of Model 3 with a linear trend on* $\mu(t) = \beta_0 + \beta_1 \cdot t$.

	Location β_0	Location β_1	Scale σ	Shape ξ
Estimates (s.e.)	29.13 (0.38)	0.0254 (0.005)	1.87(0.14)	-0.21 (0.06)

$29.13 + 0.025 \cdot 116 = 32.07$ where 116 is the number of years of data. Extrapolation, say in year 2050 for example, would then lead to a distribution GEV$(32.93, 1.87, -0.21)$, other parameters being held constant.

Diagnostics

Section **3.2.1** showed how diagnostic tools such as quantile and probability plots can still be used in a nonstationary context, with some transformations, i.e. the observations must be standardized to have a Gumbel distribution. Results are shown in Figure 6.3 for the selected model.

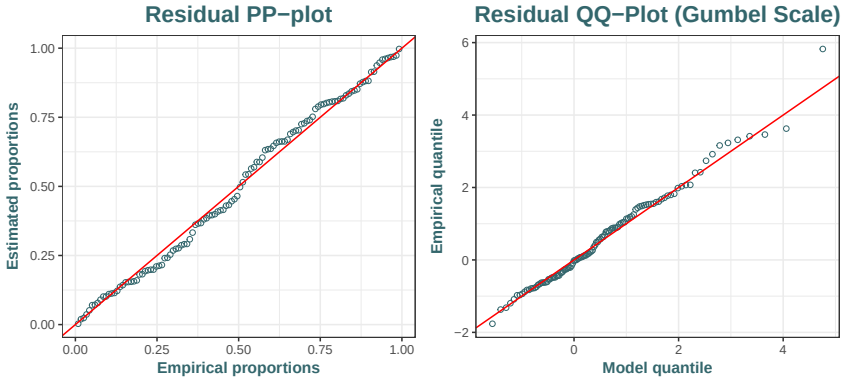


Figure 6.3: *(left) Residual probability plot and (right) residual quantile plot on the Gumbel scale for the nonstationary GEV model allowing for a linear trend in the location parameter fitted by MLE.*

We appreciate that the fit seems accurate. Note that problems for large quantiles in the residual QQ-plot is not problematic.

Return Levels

We will still use the return levels as our reference tool to make inference in EVT, even in the presence of nonstationarity. Since the distribution will change every year (by a right shift of the location), the return levels will also change each year. Hence, we plot return levels against years by choosing a 10-year return period which corresponds to the 90% quantile of the fitted GEV distribution. This is shown in Figure 6.4.

We clearly see the linear pattern of the return levels with t explained by the linear model for $\mu(t)$. The x-axis has been carefully chosen to have the best interpretation. 36 years are in the range of data and then

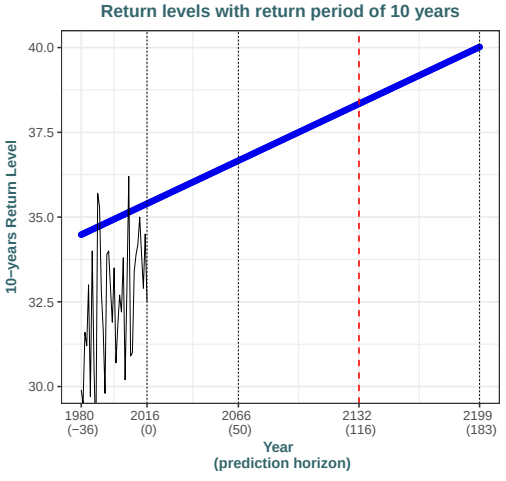


Figure 6.4: *Return levels (in blue) of the nonstationary GEV model allowing linear trend on* $\mu(t)$. *Dotted red line represent horizon after* $n = 116$ *years and black lines represent the series* $[1980 - 2016]$.

return levels are extrapolated from the model. The maximum prediction horizon of length 183 corresponds to the reach of $40°c$ for the 10-year return level.

Quite interestingly, we see that the fitted return level after $n = 116$ years, where n is the number of data, is $38.3°c$, which is higher than the maximum $36.6°c$ of the series. This return level means that, in year 2132, we will expect a maximum temperature of $38.3°c$ to be reached in a period of 10 years. Note that such far extrapolation, especially those right to the number of data (dotted red line), should be made with caution. Moreover, we cannot know in practice whether or not it is believable for the upward linear trend in annual maximum temperatures to continue to be valid beyond the range of data.

6.3 Improvements with Neural Networks

New advances in EVT enables consideration of more flexible approaches. Indeed, we have seen that a linear model in the location parameter is significant against the other parametric models of Table 6.4, but we did not consider enough models to be able to sustain this conclusion.

As discussed in Section **3.4**, Neural Networks (NN) allow through a complex process to capture complex relationships between the input (time) and the outputs (the 3 GEV parameters). We will follow the approach of Cannon [2010] and its package GEVcdn using an extension of the multi layer perceptron, the GEV Conditional Density Network (GEV-CDN).

Model parameters are estimated from weights and biases via GML using a quasi-Newton BFGS optimization algorithm and appropriate relationships, i.e. refer to Figure 3.1 for the complete architecture. To avoid convergence to a shallow local minimum of the error surface, the algorithm is run 100 times with different initial values. The selected model is the one that minimizes an appropriate cost-complexity criteria (AIC_c or BIC). Different structures are tested with combinational cases of stationary and nonstationary parameters of the GEV distribution, or linear and nonlinear architecture of the CDN.

6.3.1 Models and Results

NN's are meant to approximate any functions with very good accuracy. Thus, it incorporates all the parametric models considered so far. We allowed the shape parameter ξ to vary with time in order to consider all possible models, though this is not advised for EV models as it allows the family of the fitted distribution to change with time. In any case, it does not change the final result.

Figure 3.1 represented the fully connected architecture. The hierarchy of models we will consider is, by ascending complexity :

1. Stationary GEV model as in Table 6.1, but now fitted by **G**ML.
2. Nonstationary GEV model with linear trend on the location.
3. Nonstationary GEV model with linear trend on the location and the scale parameters.
4. Nonstationary GEV model allowing complete linear relationships for the location, scale and shape parameters.

5.-9. The preceding nonstationary models allowed linear relationships between time and the parameters since the activation function $m(\,\cdot\,)$ used to compute the hidden layer nodes (3.17) was the identity link. Now, since $m(\,\cdot\,)$ is a (logistic) sigmoid function, the relationships can be nonlinear.

The degree of complexity of the nonlinear relationships, i.e. number of degrees of freedom of the model, is controlled by the number of hidden layers. We also implemented the hyperbolic tangent as activation function but it has not improved the results.

Models 2 to 4 thus allow **linear** relationships of the parameters with time while the following Models 5 to 9 can handle nonlinear relationships. We did not consider nonstationary models for the sole $\sigma(t)$ parameter as it would not correspond to the goal of this thesis, in the sense that allowing changes in scale but not in location would be irrelevant.

We picked the recommended values of 6 and 9 for the parameters c_1 and c_2 (resp.) of the Beta prior (3.15) for ξ to compute the generalized likelihood (3.16). Moreover, Cannon [2010] recommended to use between 1 and maximum 3 hidden layers due to the relatively small sample size $n = 116$ and the danger to quickly overfit. Table 6.6 displays the results for maximum 2 hidden layers.

Table 6.6: *Comparisons of nonstationary GEV-CDN models fitted by GML. Linear models have the identity activation function and the nonlinear models have the logistic sigmoid activation function.*

	model	AIC_c	BIC	hidden	df
	stationary	-19.6	-11.5	0	3
Linear	$\mu(t)$	-37.4	**-26.7**	0	4
Linear	$\mu(t), \sigma(t)$	-35.4	-22.2	0	5
Linear	$\mu(t), \sigma(t), \xi(t)$	-34.2	-18.4	0	6
Nonlinear	$\mu(t)$	-35.4	-19.6	1	6
Nonlinear	$\mu(t), \sigma(t)$	-36.2	-17.9	1	7
Nonlinear	$\mu(t), \sigma(t), \xi(t)$	-34	-13.3	1	8
Nonlinear	$\mu(t)$	-37.4	-14.3	2	9
Nonlinear	$\mu(t), \sigma(t)$	-32.5	-4.7	2	11
Nonlinear	$\mu(t), \sigma(t), \xi(t)$	**-38.4**	3.9	2	13

Relying on the BIC, Table 6.6 reinforces the findings of Table 6.4. This criterion is justified because the AIC_c does not penalize sufficiently complex models and we want a parsimonious model since NN models are likely to overfit. Indeed, the AIC_c selected the most complex model with 2 hidden layers and 13 df, and the likelihood ratio tests always chose Model 2 ; p-value = 10^{-4} when compared with the complex model chose by AIC_c.

For Model 2 selected by BIC, Table 6.7 shows the parameters estimated by GML from the estimation of the weights $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ with (3.17)-(3.18).

Table 6.7: *Estimation by GML of the nonstationary parameters from the GEV-CDN Model 2 allowing a linear trend on the location* $\mu(t) = \beta_0 + \beta_1 \cdot t$.

	Location β_0	Location β_1	Scale σ	Shape ξ
Estimates	29.11	0.0253	1.84	-0.185

We notice that Table 6.5 leads to very similar results. Small differences are due to the method of estimation since we used GML here while the parametric nonstationary analysis in Table 6.5 used ML. Scale and shape parameters are of the same order of magnitude, and the same conclusions can be drawn since we used the identity activation function with no hidden layers. This model does not use all the

power of NN's, being more convenient to use. Nonlinear relationships retrieved from hidden layers are indeed difficult to interpret. Some quantiles of the model are shown on the left plot of the following Figure 6.5.

6.3.2 Bagging

A solution presented in Section **3.4.3** to prevent overfitting is *bagging*. Since the linear model is not complex enough to require bagging, and the support for the linear model was not straightforward when looking at the AIC_c, we will now allow for nonlinearities that will be handled by the bagging process in order to limit overfitting. From Table 6.6, we chose 2 hidden layers to keep the model relatively simple, and we only allowed the location and scale parameters to vary with time for the reasons above.

We made use of parallel computing through the `foreach` and `doParallel` framework. This method is computationally intensive and it is important to have a reliable number M of boostrap resamples for the method to be effective. It decreased computation time by a factor of ≈ 2.5 and allowed us to have $M = 1000$ resamples in only ≈ 158 seconds[1].

We tried to estimate the parameters in Table C.2 in Appendix **C**. But, from the nonlinear relationships introduced in the model, it was not possible to accurately estimate the parameters $\mu(t)$ and $\sigma(t)$. Sensitivity analysis methods such as proposed by Cannon and McKendry [2002] are recommended to yield more accurate estimates of the parameters. We note that, the scale parameter is (slightly but) constantly decreasing with time, whilst we expected that the climate warming would lead to more variable extremes.

Results of the generated quantiles are shown on the right plot of Figure 6.5. We also made available in Figure C.10 in Appendix **C** a graph which gathers results to better visualize differences in quantiles.

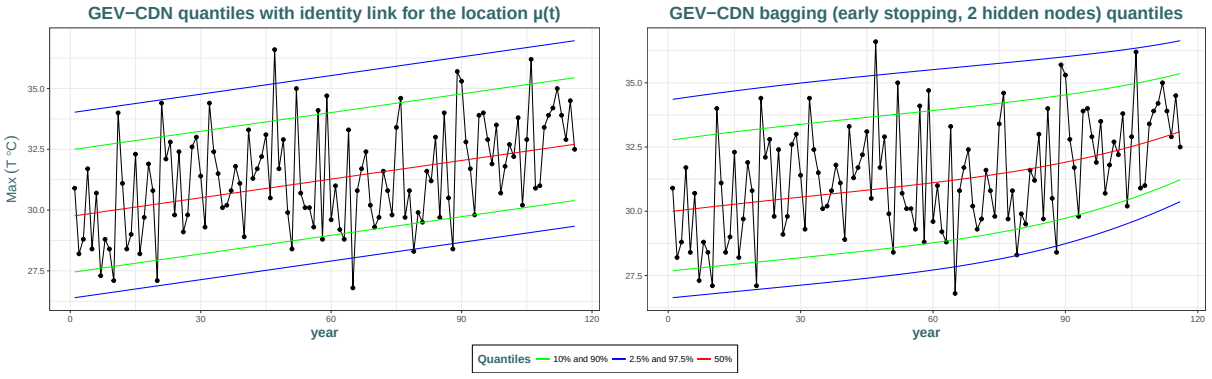


Figure 6.5: *(Left) observations with quantiles from Model 2 estimated in Table 6.5. (Right) observations with same quantiles coming from the nonstationary and nonlinear bootstrap aggregated model with* $M = 1000$ *resamples. Note that in a stationary context, the* 50%*,* 90% *and* 95% *quantiles represent return levels with return periods of* 2*,* 10 *and* 20 *years respectively.*

In Figure 6.5 we can clearly see the introduced nonlinear relationships in the two parameters $\mu(t)$ and $\sigma(t)$ with time from the two hidden layers. Compared with the left plot where the nonstationary link is linear and in $\mu(t)$ only, it seems that the ensemble model leads to an accelerate increasing rate at the end of the series, especially for low quantiles. This seems appropriate since the warming seems to accelerate for this period, and it fits with the findings of Section **5.3.3**. More formal comparisons could

[1] All computations are made on a `i7-7700HQ 2.8GHz 20gb ddr4`.

be made, for example Cannon [2010] compared predicted quantiles of the model with the empirical quantiles and computed the root mean squared error.

Note that the displayed values of the quantiles that are different for each year and represent return levels, can also be seen as confidence intervals, similar to credible intervals seen in the Bayesian paradigm in Section **4.5.2**.

In order to reduce the risk of overfitting, another technique that is commonly used is *weight penalty regularization*. In the GEVcdn framework, it works by controlling the variance σ_w^2 of a gaussian prior on the weights. Ideally, this value should be controlled by cross validation. It means that we have to find the value of σ_w that provides the best fit relying on some criterion but this method rules out the use of the selection criterion such as BIC or AIC_c since the effective number of model parameters will no longer be equal to the number of weights in the GEV CDN model. We chose $\sigma_w = 2$ to limit the risk of overfitting. At the limit when σ_w goes to zero, we checked that results are the same as for nonstationary GEV-CDN in location with no hidden layers, i.e. left plot of Figure 6.5.

6.3.3 Inference : Confidence intervals by Bootstrap

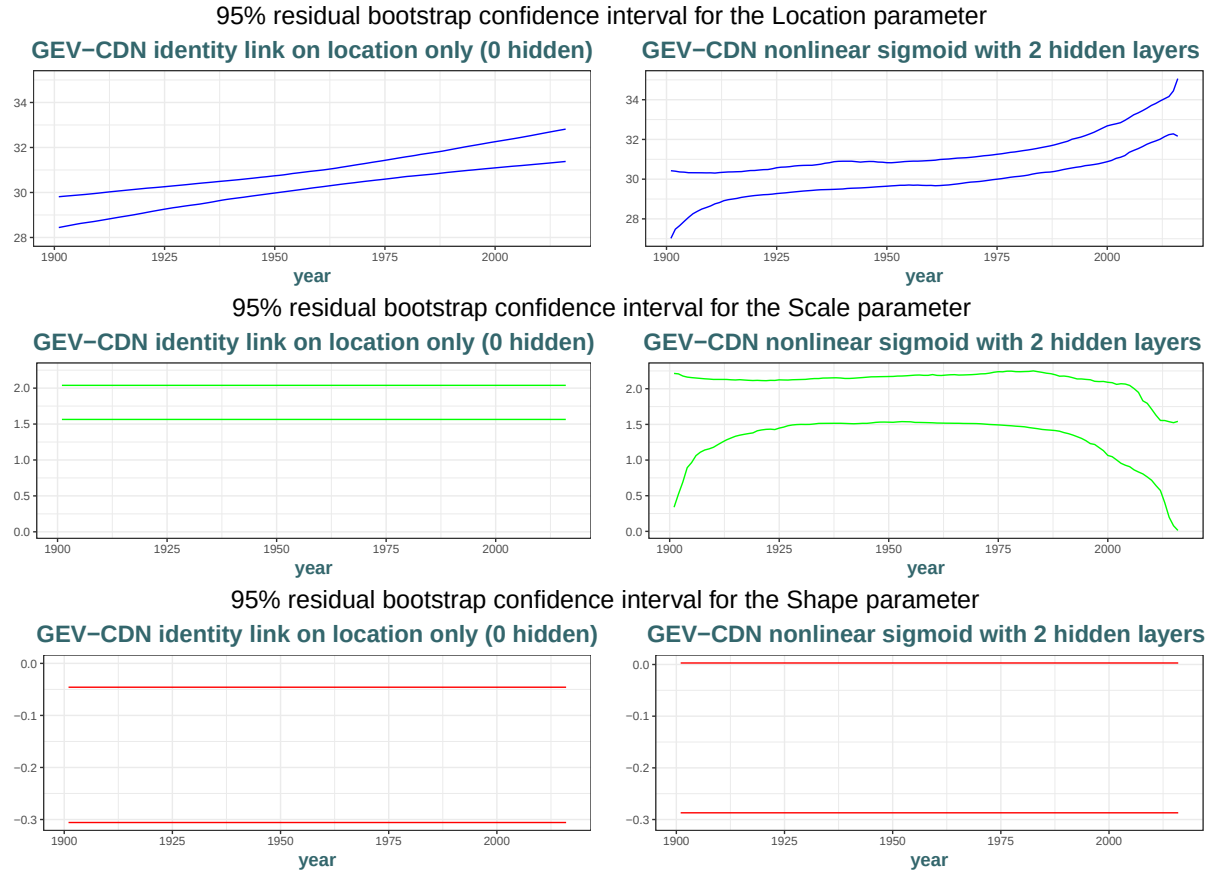


Figure 6.6: Residual *bootstrap* 95% *intervals computed with the GEV-CDN model allow a linear nonstationary location parameter only (left), and for the nonlinear nonstationary model in location and scale parameters only with 2 hidden layers (right).*

Bootstrap is an effective technique to compute accurate quantile-based confidence intervals when

we cannot rely on approximations. Similarly as for bagging, we did the computations in parallel. We took $B = 1000$ resamples in 74 seconds for the residual bootstrap. We used the residual bootstrap as presented in Section **3.4.4** since it is the method recommended by Kharin and Zwiers [2005]. The results are shown in Figure 6.6 for all three GEV parameters, where left plots come from the 2.5% and 97.5% quantiles of Model 2 (Table 6.7) and right plots are for the nonlinear GEV-CDN model considered in the previous section (Table C.2) but with no bagging since it is not relevant to combine two bootstrap methods. We compared results with the parametric bootstrap from Figure C.11 left in Appendix **C** but we note no significant differences. Graphs of the differences between the two methods did not yield additional concerns and are thus left at the end of `/Scripts-R/2NeuralsNets.R`. We think a coverage analysis is thus not relevant. These models have been chosen from previous sections. We did not extrapolate beyond the range of data.

We clearly notice the flexibility introduced in the location and scale parameters intervals with the incorporated hidden layers. Stationary parameters have constant confidence lines as expected. Regarding the shape parameter ξ, we see that the value of 0 is near to be outside the EV-Weibull type for the complex model. Moreover, we notice that this interval is wider for the complex model, but also regarding the location and scale parameters. To a certain extent, this loss of accuracy is due to overfitting, emphasized by the fact that we did not use bagging to estimate the model. We also note the decreasing behavior of the intervals for the scale parameters. This explains, to a certain extent, the decreasing quantiles intervals widths at the end of the series of right plot of Figure 6.5.

6.4 Comments and Comparisons with POT

There are many ways to analyze and we considered some of them. Sometimes our choices were somewhat subjective and thus debatable since we did not perform any sensitivity analysis. This is the reason why we wanted to provide the user automated methods to easily redo analysis and visualize the effects of changing some of the (hyper)parameters.

The other approach we have seen in Chapter **2** is the POT, and results were consistent with GEV for the EVI ξ in the stationary and nonstationary analysis. For the nonstationary analysis, we considered e.g. varying thresholds and seasonal models. A few results of nonstationary analysis in a POT approach are presented in Section 4.1 of the `Summary1_intro.html` file in the **/vignettes** folder of the repository. As expected, threshold selection were an issue.

Chapter 7

Bayesian Analysis in Block Maxima

Contents

Bayesian analysis is used more and more everywhere. For example in this text, Chapter **5** used a simulation-based Bayesian approach to compare coverage properties of pointwise and simultaneous confidence intervals, while Section **6.3** used priors to constrain the domain of the shape parameter or to penalize weights in a neural network's architecture.

After briefly presenting the available methods in Section **7.1** and discussing our choice, Section **7.2** will present the GEV stationary model in order to compare the three leading algorithms by Bayesian's practitioners for sampling with Markov Chain Monte Carlo. Section **7.3** will focus on the comparison of nonstationary models with predictive accuracy criteria in order to select the model that will be analyzed in Section **7.4**. Convergence issues will be addressed, with both quantitative criteria and visual demonstrations, and comparisons with the results obtained in the last chapter will be provided. Section **7.5** will conclude the chapter by discussing other possibles techniques and will also provide a visual summary of all the credible (or confidence) intervals obtained for the parameters during this whole text obtained during this thesis.

This analysis relies on all the scripts with filenames starting with "`Bayes_`" in the **/Scripts-R/** folder of the repository. `Bayes_own_gev.R` is especially useful to retrieve the outputs of the chapter. All the created functions are used and made available in **/R/**`BayesFunc.R` and can be used from the package. Moreover, the `/vignettes/Summary_Bayesian.html` file in the repository presents the analysis in more details.

7.1 Methods

There exists a number of methods to make Bayesian inference, including different algorithms but also different implementations of these algorithms, i.e. different packages, programming interfaces or languages. We have looked for methods that are the most suited for our nonstationary GEV analysis.

The pioneering and reference Bayesian R package in EVT is `evdbayes` from Ribatet [2006], launched more than 10 years ago, is still the only available package on CRAN to do Bayesian inference in EVT with MCMC techniques. However, problems were encountered opening the "black-box" and understanding its structure. For example, it was not possible to tune correctly the algorithm in order to reach the target posterior distribution for a nonstationary model. Moreover, Hartmann and Ehlers [2016] states that `evdbayes` only implements the Metropolis-Hastings (MH) (Algorithm 2 in Appendix **B.1.1**). Hence, we decided to develop our own methods in the developed R package. We decided to implement not only the MH, but also the Gibbs sampler (Algorithm 3 in Appendix **B.1.1**). Below is a brief summary of the methods we have used (or considered) for this chapter :

- The `evdbayes` package discussed above. Although the documentation is complete, the methods are rather not intuitive nor very flexible. Moreover, the R language has evolved this last decade.

- From *our functions* in the R package, where we implemented both the MH algorithm and the Gibbs sampler. This allowed us to have a thorough understanding of these algorithms. We also used it to carefully tune the hyperparameters and obtain relevant and reliable results, that we could visually plot in the best manner.

- From *HMC algorithm* using *STAN language*, which is the most efficient but also the most complex to use. While offering state-of-the-art methods for both execution and diagnostic of the generated Markov chains, it also presents more theoretical and practical challenges. This explains the limited amount of results we have been able to produce with this method.

- ▷ The *Ratio of Uniform* method from the `revdbayes` package (see Northrop [2017]), which internally makes use of the `rust` package to simulate a random sample from the required posterior distribution, is an other acceptance-rejection type of simulation algorithm. Even if this method is different from the traditional MCMC, the package also makes use of a low-level language (C++) to perform the most time-consuming tasks. Hence, in terms of efficiency, this method is worth studying for Bayesian inference with extreme values.

Using a compiled language, either in preference to R or as a subroutine called from R, typically results in a two-fold or three-fold speed increase over optimized R code for iterative simulation algorithms such as in MCMC simulation, see e.g. Stan Development Team [2017].

Regarding the prior $\pi(\theta)$, the different methods discussed in Section **4.2** are available in the `evdbayes` framework. Since we will mostly rely on our functions and we have no experts' advice, we will use vague priors as in Section **4.2.4**, and hence not implement all these priors in our package.

Since learning a new language (STAN) for such a complex task was limiting, we will only present the results that we obtained for the selected nonstationary model (Section **7.3**) with the functions we have constructed.

7.2 Stationary GEV Model : Algorithms' Comparison

This section presents the results obtained for the stationary GEV model in order to compare the different methods discussed above and in Section **4.3**.

We note that for this particular model, the results are very similar to those obtained with the `evdbayes` framework and are not shown here. We empirically verified that this package uses the MH algorithm, even if we think that `evdbayes` can use the Gibbs sampler, contrary to Hartmann and Ehlers [2016, pp.6] statement in their article.

Since it is not possible to evaluate analytically the complete posterior in (4.1) for this model, we will use MCMC approximations. The same non-informative (near-flat) normal priors will be used in the three methods.

7.2.1 Comparison of the Methods

Figure B.1 in Appendix **B.3** shows the parameters' chains generated for each parameters by the different algorithms. We see the chains updating one at a time in the MH algorithm while the updates are individual in the Gibbs sampler and in the HMC, implying traceplots that have the same shape for each parameters with the MH algorithm. This is related to the acceptance rates, where we can clearly visualize their effects : they are increasing from the left to the right plots, i.e. from chains that have a succession of constant episodes (left plots) to chains that resemble to those of a white noise process. This impacts the efficiency of the underlying algorithm since the parameter space of the target distribution will be more quickly visited when the acceptance rate(s) is (are) high, depending also on the structure and on other parameters of the algorithm.

Table 7.1: *Comparison of three Bayesian MCMC algorithms with $N = 2000$ samples with a Burn-in period $B = 500$ with the frequentist MLE for the stationary model. Parameters are estimated by their posterior mean, effective sample sizes (N_{eff}) (B.8) for estimating the mean are displayed for Bayesians, and standard errors for frequentist.*

	MH (N_{eff})	**Gibbs** (N_{eff})	**HMC** (N_{eff})	**MLE** (s.e.)
Location μ	30.596 (183)	30.567 (232)	30.5978 $(\mathbf{769})$	30.587 (0.216)
Scale σ	2.101 (183)	2.117 (145)	2.100 $(\mathbf{621})$	2.081 (0.155)
Shape ξ	-0.2445 (183)	-0.2449 (144)	-0.2433 $(\mathbf{541})$	-0.254 (0.067)
Computation time	0.39 sec.	2.23 sec.	0.72 sec.	

We can see from Table 7.1 that all the estimates are very close to each other and vary in a very low order of magnitude. Even if all the Bayesian methods are giving satisfying results, it is important to discover which method is the most efficient.

Metropolis-Hastings and Gibbs Sampler

Regarding the choice of the proposals, we took multivariate (MH) and univariate (Gibbs) Normal distributions since symmetric distributions have better properties and are more easily handled. We tuned the standard deviations of the proposals by a *trial-and-error* approach. Even with an univariate proposal (Gibbs), it is a difficult task to tune each standard deviations separately to achieve desirable acceptance probabilities for all parameters.

The MH algorithm is the best in terms of computation time since it only requires one update within each iterations, and therefore only one log-posterior computation (the most computationally time-consuming task in the algorithms). As the Gibbs sampler evaluates an update for each parameter separately, it takes more time to execute. This can be seen on Algorithms 2 and 3, where the latter involves a nested loop.

Since the autocorrelations in the chains is the same for the MH algorithm, the number of effective samples will be the same for each parameters. It will also highlight the convergence issues with Gibbs sampling, and we see that the shape parameter has more difficulty to attain the target posterior stationary distribution than the other parameters such as μ.

However, other technicalities have to be taken into account when choosing between these algorithms, and even these two "accept-reject" methods have lots of similarities. We have chosen to use Gibbs sampling since it produces a sequence of low dimension simulations that still converge to the right target. This is easier to handle when considering nonstationary models with increasing number of parameters (Section **7.3**).

Hamiltonian Monte-Carlo

As discussed in Section **4.3.3**, the efficiency of the HMC algorithm can be visualized in Table 7.1. Indeed, the model executes in only 0.82 seconds since the log-posterior evaluation is handled into a lower-level of abstraction, and yields a drastically higher number of effective samples (N_{eff}). It comes from the jumping rules that are much more efficient because they learn from the gradient of the log-posterior density (see step (c) in Algorithm 4), and so they know better where to jump to.

We can see in the following URL[1] that the chains are very poorly autocorrelated. This URL presents a Shiny application provided by the package `rstan` together with `shinystan` that allow to locally deploy this kind of visualization from an executed model in a few lines of code only, see **/Scripts-R/**`Shiny_stan_test.R` on the repository.

This tool was particularly useful in our case since we ran this algorithm a high number of times, and it was extremely difficult for us to attain convergence. An impressive number of diagnostics, often difficult to understand but well documented inside the application, helped us to resolve the convergence issues. In fact, this problem is linked to the property of the GEV distribution. Since the support depends on the parameter values, i.e. $x_* = \mu - \sigma \cdot \xi^{-1}$ and ${}_*x = -\infty$ as we are under an EV-Weibull model, we needed to be clever with the declarations of the parameters. Otherwise, zero or undefined density values could be picked as valid parameter values by STAN. The difficulty was also probably related to Section **1.6.1** where we have shown that with a parameter ξ being relatively near the problematic region, the likelihood computations could cause convergence problems in the STAN program.

In addition to being a lower-level language, the STAN language has several benefits :

- Although the HMC algorithm is complex to implement, the STAN community offers its help and proposes tools that help for the modeling, such as the automatically generated Shiny applications shown above. These applications generate an enormous amount of information. The `bayesplot` also provides outstanding visualizations from a STAN model. Moreover, an active STAN community driven by the founders of STAN offer technical support.

- It allows for more flexibility through the direct mathematical formulation of the desired model,

[1] `https://proto4426.shinyapps.io/ShinyStanGEV_converged/`

and the probabilistic computations that will result in the STAN program. Moreover, a large library of distributions is available.

- While the program structure is more straightforward and user-friendly for this kind of problems, every STAN programs can be called in R, and we can then benefit from all the tools and packages that R proposes.

However, the major drawback is that it is a new language methodology, with all the problems associated. Together with the convergence issues discussed above, we were unfortunately not able to compute the nonstationary models with this method and so we continued with the Gibbs sampler.

7.3 Nonstationary GEV : Model Comparisons

Although information criteria such as the BIC informed us in Chapter **6** that the nonstationary model with a linear trend on the location parameter is the best suited among all the parametric models considered for the location and the scale parameters, we decided to adopt the same sequential methodology, starting with the simplest model (Gumbel) to more complex parametric models and by using the Bayesian techniques from Section **4.7** to compare models. We hoped that then, we would be able to compare the results between the frequentist and the Bayesian frameworks, and perhaps reinforce our confidence that the model selected in Chapter **6** is the best suited. The decision could be based on the frequentist's analysis of the previous chapters to select some (subset of) more relevant models to analyze, but we decided to take the same models in order to facilitate the comparisons. Note that the number of parametric models that we propose is still limited.

The Gibbs sampler was then used for each models. To limit the influence of the four randomly selected starting values, we took a burn-in period of length $B = N/2$. Taking $N_{\text{tot}} = 4 \cdot 1000$ for each parameter's chains, this leaves us with $N = 2000$ samples. With tools from Section **4.4**, we diagnosed the convergence of each chains separately, for each model.

Bayes Factor

The Bayes Factor is the most straightforward statistic in the Bayesian paradigm to compare models. Unfortunately, we did not find any ways to compute the marginal likelihood (4.15) which is an intractable 2 to 6-dimensional integral for the models considered. This is a necessary and often considerable issue when computing the Bayes factor in practice. We were not able to compute for our models even with the efficient method proposed by Chib [1995].

Predictive Information Criteria

Another way to compare models is by using the predictive information criteria discussed in Section **4.8**. Rather than evaluating the quality of fit of a model in terms of its accuracy with the observed data, these criteria evaluate the predictive accuracy which is, to a certain extent, relatively similar .

For each of the four generated chains with dispersed starting values, we evaluated separately the predictive information criteria to check that the variability of the criteria is reassuringly small between chains. The discrepancies between the chains are very small for most models, which is a good sign.

Except for the simpler models, where we observed much more variability in these predictive criteria. We obtained the following standard deviations for the most simple and the most complex models :

$$\sigma(\text{DIC}|\text{Gumbel}) = 157, \qquad \sigma(\text{WAIC}|\text{Gumbel}) = 78;$$
$$\sigma(\text{DIC}|\text{Cubic}) = 5, \qquad \sigma(\text{WAIC}|\text{Cubic}) = 3.$$

The complete results are displayed in Table 7.2, by averaging the criteria for each parameter chain with a different starting value.

Table 7.2: *Comparisons of nested (non)stationary GEV models computed by the Gibbs sampler for $N = 2000$, by means of predictive accuracy criteria.*

Model	DIC	WAIC	Computation Time (sec.)
stationary Gumbel	517.4	515.6	0.98
stationary GEV	576.6	540.1	1.9
linear in $\mu(t)$	**720.9**	**602.5**	3.9
quadratic in $\mu(t)$	484.6	483.9	4.9
linear in $\mu(t)$ and $\sigma(t)$	486.6	485.1	5.4
cubic in $\mu(t)$	450.3	487.8	7.8

The model that is the most predictive accurate is the nonstationary GEV model with a linear trend on the location, as previously chosen with other criteria in Chapter **6**. We noted that the difference with the other models is large. The more complex models have a very low amount of predictive accuracy, probably because they overfit the data and they do not capture the signal. Since we were not able to compare models with the Bayes factor, one could argue that carefully using Bayesian model averaging could improve the results. We discuss this issue at the end of this Chapter, but at first thought averaging with those less informative complex models would not be beneficial. The computation time is obviously higher for models that have more parameters since each parameters requires one update (see Algorithm 3).

7.4 Nonstationary GEV with Linear Model on Location

From the above Bayesian model selection with predictive criteria, but also the more straightforward selection made during Chapter **6** (see e.g. Tables 6.4 and 6.6) in a frequentist setting, we decide to explore this selected nonstationary model to see how a Bayesian analysis can help with inferences.

We write our model for the annual maxima X_t as

$$X_t \sim \text{GEV}\big(\mu(t), \sigma, \xi\big), \qquad \text{with } \mu(t) = \mu_0 + \mu_1 \cdot t,$$

where $t = [1901, 2016]$ but in practice, we will take the rescaled values of t by $t^* = \frac{t-\bar{t}}{|t|}$ that yield better properties of the Gibbs sampler. We then have four parameters $\theta = (\mu_0, \mu_1, \nu, \xi)$, where $\nu = \log\sigma$. We take independent vague prior distributions for each parameters :

$$\pi(\mu_0) \sim \mathcal{N}(30, 40^2), \quad \pi(\mu_1) \sim \mathcal{N}(0, 40^2), \quad \pi(\nu) \sim \mathcal{N}(0, 10^2), \quad \pi(\xi) \sim \mathcal{N}(0, 10^2).$$

Values are recommended by Dey and Yan [2016, chap.13]. We verified that relatively small changes in

the hyperparameters do not significantly influence the results. For numerical reasons, will compute the unnormalized posterior on the log scale, i.e.

$$\log \pi(\theta|\boldsymbol{x}) \propto \log \pi(\theta) \cdot \ell(\theta|\boldsymbol{x}), \tag{7.1}$$

in order to facilitate the numerical computations while not affecting the results.

MCMC Algorithm : Gibbs Sampler

Since it is not possible to evaluate analytically (4.1) for this model, we will use the MCMC method to approximate the true target posterior distribution.

We will use the Gibbs sampler (Algorithm 3) as it is the most effective, especially in a nonstationary setting, i.e. when the number of parameters to evaluate is high. As the computation time is not a real issue for our application, this sampler is the perfect candidate.

Starting Values

Chapter **6** numerically optimized the log-likelihood to compute the MLE in the frequentist setting. Now, we will use the same technique but on the log-posterior (7.1). To control the influence of the starting values, we run $c = 4$ chains, each having different starting values. This $[4 \times 4]$ matrix $\boldsymbol{\theta_0}$ of starting values are randomly selected from a multivariate normal distribution that is over-dispersed relative to the target. To make this clear, we present this in Algorithm 1.

Algorithm 1: Compute c starting values $\boldsymbol{\theta_0}$

1. Take the optimized values from log-posterior from (7.1), say $\hat{\theta}$.
2. **For** $i = 1, \ldots, c$ **do**
   - Sample the i-th starting value $\theta_{0,i}$ from

$$\theta_{0,i} \sim \mathcal{N}_4\Big(\hat{\theta},\ k \times \Sigma^l\Big)$$

   where Σ^l is the estimated covariance matrix of the log-posterior optimized values $\hat{\theta}$.

Note that $\hat{\theta}$ could be the MLE if the priors are near-uniform. To obtain over-dispersed values, k must be high. We take $k = 50$ in our application to increase our confidence that the whole parameter space has been visited. This value can be tuned to obtain more desirable starting values. We display the obtained starting values in Table 7.3. We see the great dispersion in the starting values. The shape parameter includes both positive and very negative values, while the trend parameter μ_1 includes negative values but also (very) high values. The generated chains will be displayed in Figure 7.1.

Proposal Distribution

The implemented Gibbs sampler (Algorithm 3, or `gibbs.trend.own` in our R package) has univariate normal distributions as proposal distributions. This choice could be questioned but we will take

Table 7.3: *Starting values taken for the Gibbs sampler, from Algorithm 1 with* $k = 50$.

	μ_0	μ_1	$\log\sigma$	ξ
$\theta_{0,1}$	30.193	5.730	0.960	-0.349
$\theta_{0,2}$	31.849	4.402	0.447	0.257
$\theta_{0,3}$	31.487	0.478	1.156	-0.529
$\theta_{0,4}$	31.636	-0.011	-0.024	0.130

it for its convenience since it is symmetric, and results would not significantly change as long as the acceptance rates are acceptable and the algorithm converges : it would only impact the algorithm's efficiency. Hence for each parameter $j = 1, \dots, 4$, we have at iteration t of the algorithm

$$p_{t,j}\left(\theta_*^{(j)} | \theta_{t-1}^{(j)}\right) \sim \mathcal{N}\left(\theta_{t-1}^{(j)}, \left(\sigma_p^{(j)}\right)^2\right), \tag{7.2}$$

where $\boldsymbol{\sigma_p} = (0.5, 1.9, 0.15, 0.12)'$ is the 4-dimensionnal vector containing the proposal's standard deviations of each parameter. This vector is a hyperparameter that we have tuned by trial-and-error to obtain the recommended acceptance rates for each parameter's chains : all between 36% and 43% for the $4 \cdot 4 = 16$ computed chains. If generalization is required, it is possible to find an automatic way to achieve this instead of relying on trial-and-error.

Computations

It took only 3.5 seconds for $N_{\text{tot}} = 4 \cdot 1000$ iterations. For each of the 4 chains with different starting values, we took a burn-in period of 500, leaving us with $N = (4000 - 4 \cdot 500) = 2000$ samples. It would be better to take more iterations as it is quite fast, but for better visualizations, we kept $N = 2000$. Note that running the different chains in parallel is straightforward and should decrease the computation time by a factor of ≈ 2, but it is not necessary here for such a small N_{tot}.

7.4.1 Diagnostics

Before any inference on the generated chains, it is vital to check whether the samples reached their target stationary distribution or not. We note that no diagnostics can prove convergence, but their combined use will increase our confidence that it indeed happened. This section will sometimes rely on the `coda` package for some diagnostics. These diagnostics are described in Appendix **B.2**

Figure 7.1 shows the traceplots of the chains after burn-in. Even if the convergence seems to occur quite fast from the complete traceplots of the chains, we preferred removing 50% to increase our confidence that the starting values have no influence. Each parameter show chains that are mixing well, reinforcing our confidence that the over-dispersed starting values we used have no influence on the target stationary distribution. Thence, we now gather all the chains with different starting values into a single chain, to obtain one single chain of size $N = 2000$ for each parameter.

Gelman-Rubin

The Gelman-Rubin diagnostic compares the behavior of the randomly initialized chains. It computes the $\hat{R}$ from (B.7) or *shrink factor* which measures the mixing of the chains. If convergence occurred, the

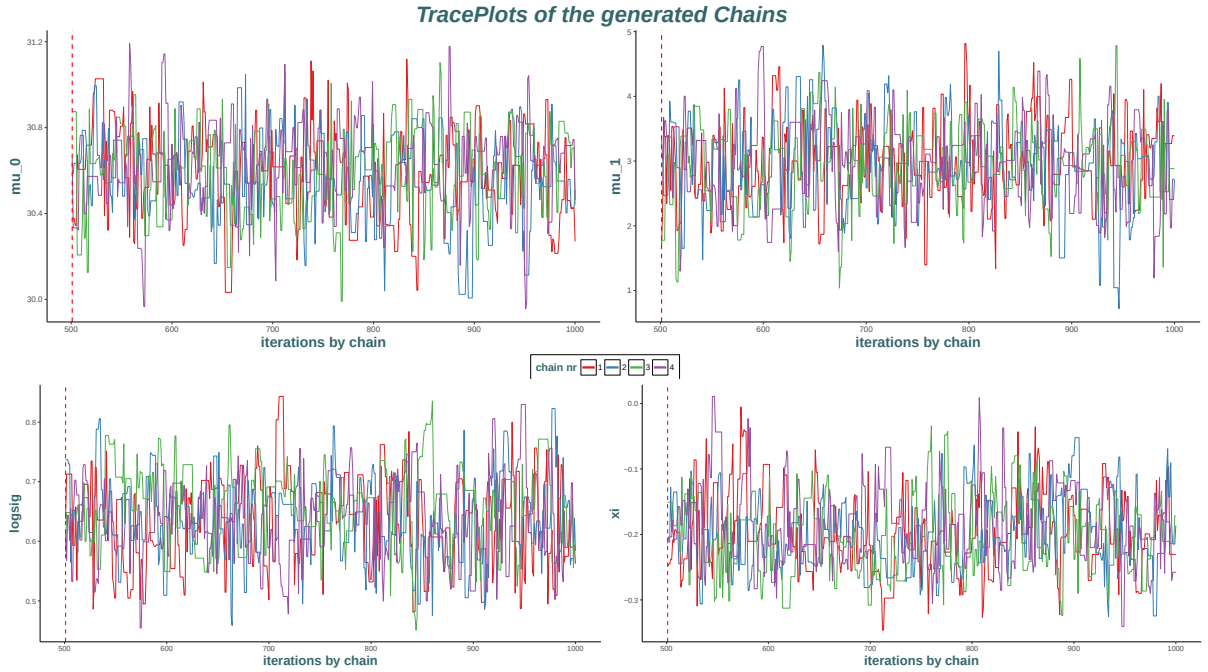


Figure 7.1: *Traceplots of the chains with 4 different starting values obtained with our Gibbs sampler for the nonstationary model with linear model on location. Note that the location parameter of the trend μ_1 is of different order as before because we are based on the rescaled values t^* of t. We will transform it back later for inferences (Table 7.4).*

$\hat{R}$ will be 1. The corresponding results are displayed in Figure B.2 in Appendix **B.3**. It clearly shows that the parameter μ_0 attains very quickly the values that indicate convergence, and somewhat surprisingly, it is yet faster for μ_1. As expected, it is more tedious for the shape parameter, but especially for $\log\sigma$ for which takes at least 500 iterations before $\hat{R} < 1.1$. However, since all parameters have $\hat{R}$ very close to 1 at the end of the series, the number $N = 2000$ of iterations should be sufficient.

Correlations within and between chains

- The **autocorrelations** of each parameter's Markov chain depicted in Figure B.3 in Appendix **B.3** quickly decrease with the lag. Again, it seems more problematic for $\log\sigma$ and ξ as the decrease is slower, showing more dependence inside their respective chain. We have computed the effective sample size N_{eff} from (B.8), i.e. the sample size after correction for the autocorrelation in the chains :

$$N_{\text{eff}}^{(\mu_0)} = 389, \qquad N_{\text{eff}}^{(\mu_1)} = 436, \qquad N_{\text{eff}}^{(\nu)} = 194, \qquad N_{\text{eff}}^{(\xi)} = 217. \tag{7.3}$$

  We can see the link of these values with the autocorrelations in Figure B.3.

- The **cross-correlation** between the parameters' Markov chains are depicted in Figure B.4 in Appendix **B.3**. High correlations go, as expected, for ξ with $\log\sigma$, but also for ξ and μ_0. Reparametrization could be considered. We have compared these values with the Fisher information matrix that we transformed in a correlation matrix, and we have found very similar values except for the correlation between the location parameters μ_0 and μ_1 (probably because of the rescaling of t).

Geweke

This diagnostic tests the equality of the first 10% of a chain with the mean computed from the second half of the chain. The chain has been partitioned in order to repeat the test 20 times. Figure B.5 in Appendix **B.3** shows the results. Quite surprisingly, μ_0 has 6 rejected values out of 20, but these are at the limit. Shape and log scale parameters are still somewhat problematic. However, we notice that this diagnostic is very diffuse, in the sense that results will be very different from one test to another with different generated chains with the same method.

Raftery and Lewis and Thinning

One last diagnostic comes from Raftery and Lewis [1992] and provides interesting values on the chains. These are shown in Table B.1 in Appendix **B.3**. Relatively high autocorrelations in the chains are again highlighted here. The advised values for N are slightly higher to those used, and the Burn-in period is very low.

Conclusion

From the above results, i.e. especially from the relatively high correlations within and between chains, it is recommended to increase the number of iterations N. Since the computational time is not of real concern, it would be easy to increase N. But in order to keep the same chains as presented above (by carefully setting the same seed for each random generation), and to keep the analysis homogeneous, we decreased the burn-in period by 50%. We are now left with $N = 3000$ samples for each chains.

7.4.2 Inference

Now that we have shown confidence that the Gibbs sampler has converged, in the sense that the Markov chains are supposed to have been sampled from their target posterior stationary distribution. It is now possible to make reliable inferences on the GEV parameters with these samples.

We first present the results in Figure 7.2 with the corresponding Markov chains' Kernel posterior densities for each parameters. We have added their corresponding 95% Bayesian intervals (see Section **4.5.2**. We can see the differences between the two intervals when the densities are not symmetric. In this case, the highest-posterior density (HPD) intervals are slightly more precise, and should be preferred.

This Figure 7.2 brings the advantage of displaying the complete probability distributions of each parameters. We see here that the probability mass for the shape parameter is below 0, i.e. $\Pr\{\xi > 0\} = 0$, reinforcing our confidence that we are under an EV-Weibull model. Nevertheless, this should be tempered by caution since by relaunching the same model a high number of times, we obtained results where $\Pr\{\xi > 0\} > 0$, with a very small probability. This should lead us to consider sensitivity analysis. These results are also summarized in Table 7.4.

Depending on the shape of the distribution, the median or the mean will generally be preferred as an estimate. Comparing these results with the frequentist in Table 6.5 we can appreciate that these are very similar. The value of ξ, σ and μ_0 are slightly higher in the Bayesian results, but not significantly. Moreover, the most important parameter for our nonstationary analysis, μ_1, has the same value at 3 decimal places. We see from Figure 7.2 that it has the most peaked density distribution. Hence, the interpretation made with this model still holds.

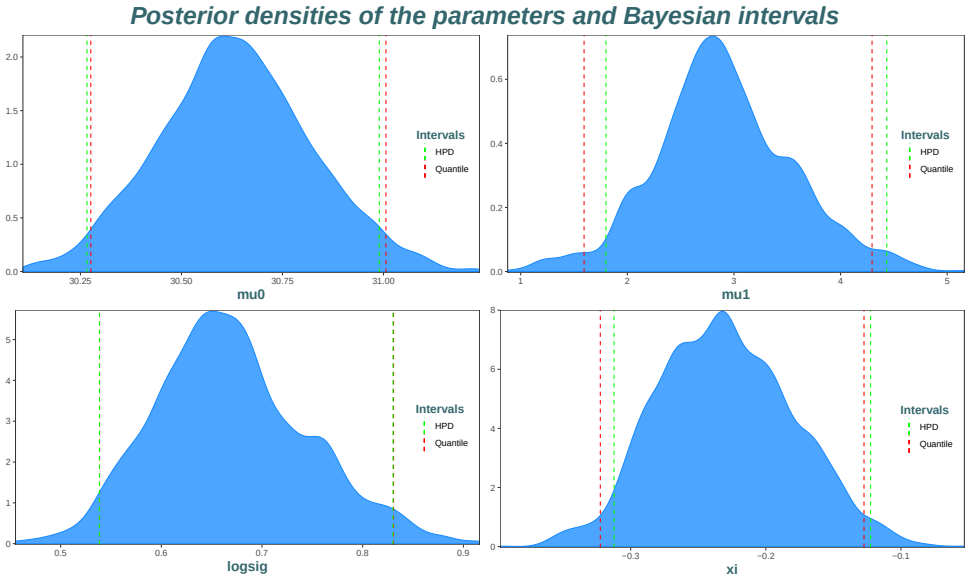


Figure 7.2: *Markov chains' Kernel posterior densities for the parameters with their corresponding Bayesian intervals. A Gaussian kernel with a bandwidth calculated by the Silverman [1986, pp.48, (3.31)] "rule of thumb" for each parameter has been used. Note that we did not make the back rescaling from t^* to t for this plot, which explains the values of μ_1.*

Table 7.4: *Table of quantiles and* ***mean*** *of $\pi(\theta|\boldsymbol{x})$. Here, we transformed back μ_1 from t^* to t in years for convenient comparisons with frequentist results.*

	2.5%	25%	50%	**mean**	75%	97.5%
μ_0	30.276	30.511	30.632	30.633	30.754	31.006
μ_1	0.0136	0.0214	0.0245	0.0248	0.0282	0.0367
$\log\sigma$	0.539	0.617	0.663	0.669	0.717	0.831
ξ	-0.323	-0.266	-0.232	-0.230	-0.195	-0.127

7.4.3 Posterior Predictive Distribution

The Posterior Predictive Distribution (PPD) is a pure Bayesian method of inference that allows to make inference for predictions relying on the posterior results, as presented in Section **4.6**. The primary objective is often prediction, in the sense that it is of interest to extrapolate the annual maxima in Uccle and for example predict the range of values that would take these maxima in order to have an idea, say, on the extreme heat waves that could occur in the future.

To compute this PPD from (4.11), we expressly took a range of 116 years in the future, as we know that it is not recommended to make very long-term extrapolations. We first depict the results in Figure 7.3 where we represent the PPD by its 95% credible intervals, together with the observed values from 1901 to 2016 and the values from 2016 to 2131 that have been simulated from this PPD (4.12).

We see in Figure 7.3 the linear trend from the linear model on the location parameter of the posterior distribution. Indeed, we see from (4.12) the contribution of the posterior to the PPD. This Figure 7.3 also clearly highlights that the PP intervals are not 95% credible intervals for the observed values but rather intervals for the posterior predicted values. Indeed, coverage analysis shows that the PP intervals cover

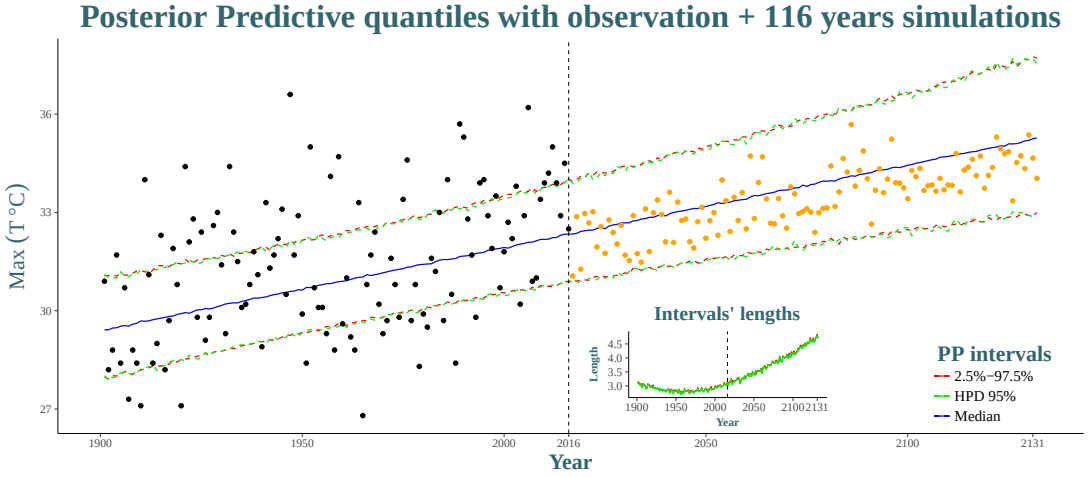


Figure 7.3: *Black dots represent the observations while orange dots represented the simulated values from the PPD. The printed plot aims to conveniently display the differences between the upper and the lower bounds of the two credible intervals considered.*

95% of the simulated values from the PPD as the number of simulations becomes very high, but only $\approx 50\%$ for the observed values. The HPD and the quantile-based credible intervals are very similar, for all the observations or simulations. However, we can see that these intervals are taking the uncertainty of predicting the future into account as they exponentially increase beyond the range of data. In fact, the evolution of the PP quantiles is linear when in the range of data, and so is the PP median even beyond the range of the data. But, in extrapolation, the PP upper quantiles will have an increasing slope over time while the PP lower quantiles will have a decreasing slope.

To better understand the PPD with a global view and in order to obtain a more convenient visual quantification of the predictive uncertainty over time, we present Figure 7.4.

From this Figure 7.4, we also notice the linear trend and we visualize more clearly the evolution of the quantiles over time with their gap increasing after 2016, i.e. from their deviation to the median. This comes from the predictive density distributions that become more and more flat after 2016, indicating an increasing variance and hence the inclusion of the prediction uncertainty through the PPD.

Interesting information can come from these PP plots when considering other models, and the shape of the predictive densities is sometimes interesting. But, the parametric models from Table 7.2 are rather limited, and other models need to be developed. For example, step-change models, or more flexible models by following the idea of Section **6.3**. It would then be interesting to consider *Bayesian Neural Networks*.

7.4.4 Return Levels

We note that quantiles of the fitted distribution can be more conveniently interpreted in EVT as returns levels. Here, for the quantiles depicted above it is different since we are under the fitted PP distribution and not the fitted distribution. Indeed, these quantiles are taking the uncertainty of future prediction into account and are thus not strictly linearly increasing. In fact, considering quantiles in extreme tails will increasingly raise the gap between these quantiles, as the PPD will become increasingly flat and will allow for increasingly more extreme values. Yet, we have seen in Figure 6.4 that the increase

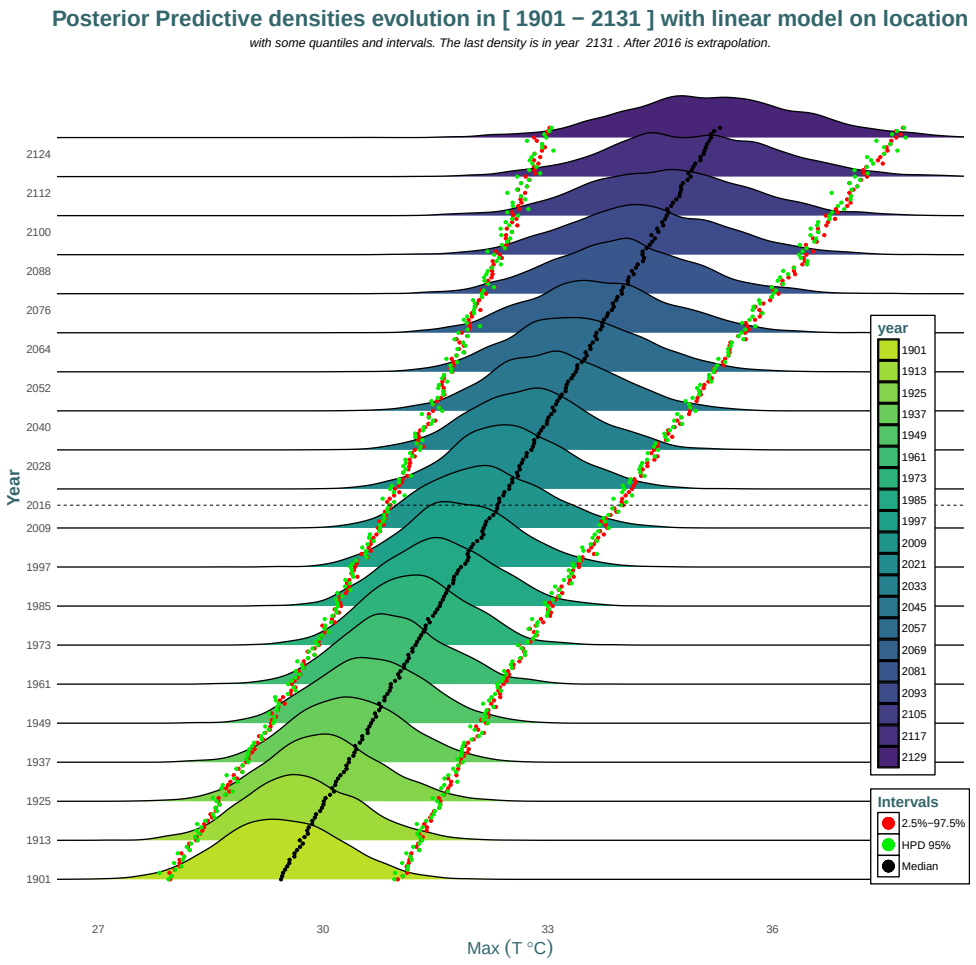


Figure 7.4: *Visualization of the PPD from* 1901 *and extrapolated until year* 2131. 20 *densities are drawn in this range by steps of* 12 *years. Quantiles are displayed for every years. A Gaussian kernel with a joint bandwidth calculated by the Silverman [1986, pp.48, (3.31)] "rule of thumb" is picked for each densities.*

of the return levels was strictly linear even beyond the range of the data, although it was the same nonstationary model in a frequentist setting. In fact, the quantiles of the PPD depicted in Figure 7.3 or 7.4 are commonly called *predictive return levels*. For example the above red line in Figure 7.3 is actually the 40-year predictive return levels.

It is still also possible to compute the returns levels as in Figure 6.4 by simply estimating the parameters (e.g. by their MC's posterior mean or median from Table 7.4). Then, the GEV model can be fitted and the corresponding return levels can be computed from this nonstationary GEV model. Relying on the posterior median, we verified that the line defining the return levels is nearly the same for all points since the parameter's values are very close ; Bayesian return levels having a slightly higher intercept and slightly smaller slope compared to the blue line in Figure 6.4.

7.5 Remarks and Comparison with Frequentist results

In this analysis, we relied on Bayesian inference to complement the nonstationary analysis started in Section **6**. As we noticed, comparing Table 7.4 with Table 6.5 yielded very similar results. This is expected and is comforting as we used non-informative priors.

As this is the last section, we wanted to gather all the results obtained over this thesis in one graph that could summarize the information in the most convenient way. The following Figure 7.5 compares several parameters' confidence intervals that have been computed during this thesis.

- The *frequentist* confidence intervals computed in Section **6.1**, comprising those coming from the normal approximation of the MLE and the profiled likelihood intervals.
- The *bootstrap* confidence intervals computed in Section **6.3.3** for the nonstationary model with a linear model on the location only. We show the residual but also the parametric bootstrap method. These intervals have not been computed for the location parameters since these were directly computed for $\mu(t)$ rather than μ_0 and μ_1. From the particular nature of these intervals which estimated parameters by GML and made use of prior distribution to compute the GEV-CDN weights, we will not consider these as truly "frequentist", but this is debatable.
- The *Bayesian* credible intervals seen and discussed above in Section **7.4.2**.

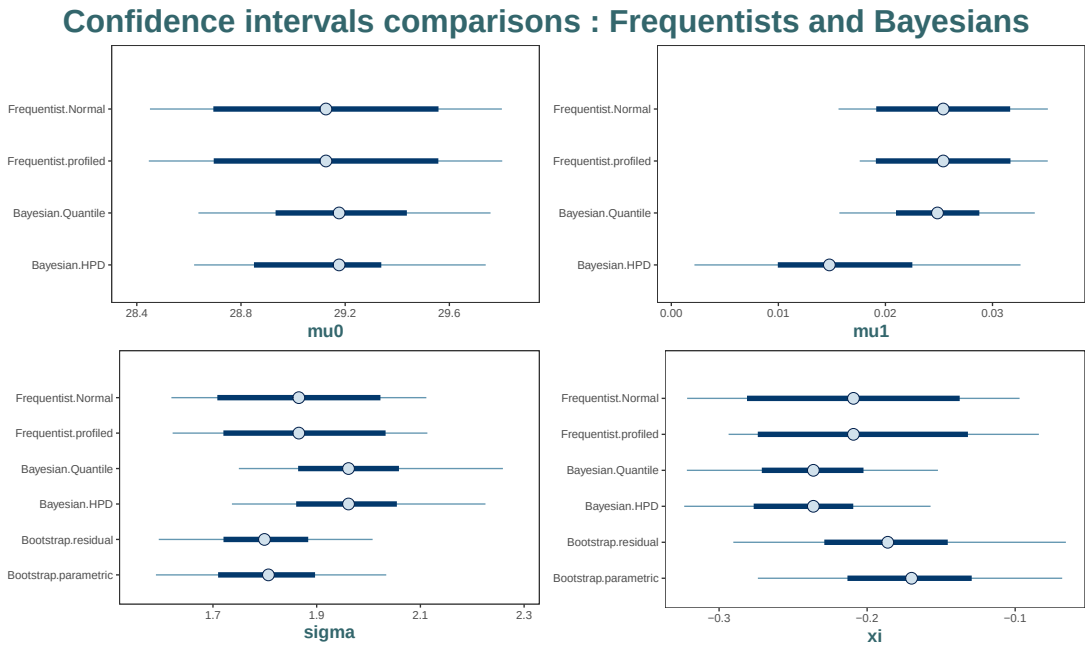


Figure 7.5: *The center of the interval is the MLE for the frequentist intervals and the posterior median for the Bayesian intervals. Thicker lines indicate* 50% *confidence (or credibility) intervals while thinner lines indicate* 95% *intervals.*

By using objective priors (i.e. priors with very large variance), we should have obtained the same results in the frequentist and in the Bayesian setting, which will be a proof of robustness of these results. However, the way of computing the intervals are intrinsically different and hence, the observed differences between these intervals is rather a reflection of the difference in the methods used.

Note that the large differences between the Bayesian and the frequentist intervals for the parameter μ_0 are probably due to the difficulties faced by the rescaling of the parameter μ_1 and has influence on the distribution of μ_0. The same argument goes for the Bayesian HPD for μ_1 where the rescaling with t brought problems.

Discussion on the Prior : Sensitivity Analysis

In analysis undertaken, we were not able to introduce knowledge through the specification of the prior. Hence, we have used non-informative priors (see Section **4.2.4**), and in particular *vague* (near-flat) independent normally distributed *priors* (4.8) for each parameter since this is the most convenient way to express our ignorance.

The basic method of sensitivity analysis is to fit several models to the same problem. From our prior ignorance, it was not relevant to analyze the sensitivity from the prior. Even if the considered nonstationary models were limited here, it could be the case that several models provide an adequate fit to the data. In this case, sensitivity analysis could determine by what extent posterior inferences change when alternative models are used. This includes posterior distributions of the parameters but also the PPD. This would be interesting to consider when we have adequate models at hand, but when looking at Table 7.2, it would be hazardous. Note that the sensitivity of the marginal posterior density of the shape parameter ξ is often of particular interest

Bayesian Model Averaging

Since we were not able to strictly select one model in a Bayesian fashion with the Bayes factor in Section **7.3**, *Bayesian Model Averaging* could an idea to prevent us from fitting one single model that could be wrong. This idea follows the one of the ensemble model such as bagging used in Section **6.3.2**. However, this method is beneficial when we have a large amount of models that could represent a great part of the process, or when the models at hand are really informative for the process, which is not the case here.

Conclusion

This thesis aimed at statistically assessing the presence of a nonstationary model using a sequence of annual maximum temperatures in Uccle from 1901 to 2016 and applying the Extreme Value Theory. Official data was used to evaluate the effects of climate warming. To this end, this thesis is divided into two parts to make a clear separation between the literature review (Part **I**) and the application of the methods (Part **II**).

Part **I** stars by presenting the approach of block maxima in Chapter **1** which models the extremes that occur inside a block. It introduces the GEV distribution and its parameters that form at the basis of the nonstationary models. The second approach of modeling extremes discussed in Chapter **2** aims to model the extremes that exceed a certain threshold. Since it was decided to work only with annual maxima due to space constraints, this thesis puts more emphasis on the first approach. In order to evaluate the form of the nonstationarity in the sequence, Chapter **3** delivers tools that allows the analysis to go beyond the restrictive independence assumption of the first two chapters. This provides the tools to consider different parametric models for the GEV parameters that will form the basis of the nonstationary analysis, as well as a flexible approach using Neural Networks to surpass the restrictive parametric assumption. Chapter **4** presents the Bayesian analysis applied to the world of extremes which provides us inferential methods that better account for both estimation and prediction uncertainty.

In Part **II**, Chapter **5** introduces the analysis with other methods not from EVT to provide initial insights on the analysis of the trend. Although the trend were significant from a simple linear regression, simultaneous intervals based on splines derivatives of a Generalized Additive Model highlighted that the trend is not significant over time, whilst pointwise intervals emphasized the accelerating trend of the annual maxima in the last 40 years. This accelerating behavior of the annual maxima at the end of the sequence was confirmed at the end of Chapter **6** by a bootstrap aggregated Neural Network which used a deep nonlinear model that address overfitting. However, prior to this, this thesis identifies that the best and most parsimonious parametric GEV model had only a linear trend on the location parameter. This is the same model selected by the flexible modeling with Neural Network, according to information criteria to prevent overfitting among a large number of models. Chapter **7** finally confirms with predictive accuracy criteria that the nonstationary GEV model with a simple linear model on the location parameter is the best among some carefully chosen parametric models. Hence,

> this thesis concludes, based upon the outcomes of three intrinsically different modeling techniques relying on the GEV distribution that there is a linear trend on the location of the annual maxima in Uccle.

It is important to highlight that a climatologist from IRM noted that artificial warming has been observed on cities' stations which were significantly less urbanized 100 years ago, including Uccle. This,

with the limitations we specified in Part **II**, means that the conclusions and interpretations drawn in this thesis should be tempered by great caution.

Time and space constraints did not allow this thesis to consider some methods :

- The *Bayesian Neural Networks*n from the pioneering thesis of Neal [1996] : these are now widely used for their high level of flexibility. For this thesis and to make it simple, it would have meant roughly combining the flexibility of the GEV-CDN framework of Sections **6.3** and **3.4** (in the sense of being capable of modeling any relationships) with the flexibility of the Bayesian analysis of Chapters **4** and **7** (in the sense of being, for example, able to tune a vast amount of information). Hence, it would have been able to surpass the restrictive assumption of parametric models that have been considered in the Bayesian chapters.

- This thesis considers time as the sole covariate to explain the changes in the extreme temperatures. Other covariates could have been included to improve the nonstationary analysis. For example, the *Southern Oscillation Index* (a proxy for El Niño), the CO_2 level, etc. However, these alternative covariates are (highly) time-dependent, and linking to a climatological covariate makes the extrapolation into the future more difficult as one would need to extrapolate the covariate as well.

- Since the IRM is able to provide temperature measurements in real-time, it would be feasible to automate all the methods and the analysis by using similar tools (e.g. in the R package or in a Shiny application) in order to automatically take into account each new observations, providing live updates of the statistical models.

- Naveau et al. [2017] noticed that winter is becoming warmer in the context of climate warming. Computationally speaking, it would require only a few changes in order to undertake the same analysis as was undertaken with maxima. Here again, an automated method would be very easy and worth implementing. These different analysis have been introduced a then end of `1intro_stationary.R` and the beginning of `2nonstationary.R`, but we did not use all the methods on these new data.

Finally, this thesis analyzes the increasing behavior of extreme temperatures, but it does study the causality of this change nor make an extensive analysis of the anthropogenic nature of the global warming, which is often the big issue.

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*

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

△

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,* $\forall t > 0$,

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

△

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.*

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

△

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*

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

△

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 Lafaye de Micheaux and Liquet [2009] 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** if*

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

△

We notice 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 and objective analysis can be derived, for example from the theory of linear regression ; see e.g. Beirlant et al. [2006, chap.1].
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 ordered sample (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.

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

Hill estimator ($\xi > 0$)

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 used with attention.

Moment estimator ($\xi > 0$)

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 under certain conditions.

Appendix B

Bayesian Methods

B.1 Algorithms

B.1.1 Metropolis–Hastings

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

Algorithm 2: 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_* coming 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
   $$\theta_t = \begin{cases} \theta_* & \text{with probability } \alpha = \min(r, 1); \\ \theta_{t-1} & \text{otherwise.} \end{cases}$$

The complex integral in the denominator of (4.1) cancels in r, leaving us with the unnormalized density. Note that the proposal density is often chosen to be symmetric, so that we will sample under the "Metropolis" algorithm, where r is simply the ratio of the posterior densities,

$$r = \frac{\pi(\theta_*|\boldsymbol{x})}{\pi(\theta_{t-1}|\boldsymbol{x})}. \tag{B.1}$$

B.1.2 Gibbs Sampler

We defined the remaining subvectors at their current values $t-1$ when subvector j is sampled

$$\theta_{t-1}^{(-j)} = \left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)}\right),$$

where each $\theta_t^{(j)}$ is sampled from $\pi\left(\theta^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right)$.

Algorithm 3: 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 $p_{t,j}\left(\theta_*^{(j)}|\theta_{t-1}^{(j)}\right)$,
   (b) Compute the ratio

$$\begin{aligned} r =& \frac{\pi\left(\theta_*^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_{t-1}^{(j)}|\theta_*^{(j)}\right)}{\pi\left(\theta_{t-1}^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_*^{(j)}|\theta_{t-1}\right)} \\ =& \frac{\pi\left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_*^{(j)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)} \mid \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_{t-1}^{(j)}|\theta_*^{(j)}\right)}{\pi\left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_{t-1}^{(j)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)} \mid \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_*^{(j)}|\theta_{t-1}\right)}, \end{aligned} \tag{B.2}$$

   (c) Set

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

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, we can apply the 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 $r = 1$, by taking

$$p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)}) = \pi(\theta_*^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}).$$

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 simplify (B.2). It is usually taken to be the univariate normal distribution with mean $\theta_{t-1}^{(j)}$.

B.1.3 Hamiltonian Monte Carlo

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 $\boldsymbol{p} \sim N(\boldsymbol{0}, \boldsymbol{M})$. 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 $\boldsymbol{p}'\boldsymbol{M}^{-1}\boldsymbol{p} \cdot 2^{-1}$. Then the total energy of a closed system is the* ***Hamiltonian function***

$$\mathcal{H}(\theta, \boldsymbol{p}) = -\mathcal{L}(\theta) + \boldsymbol{p}'\boldsymbol{M}^{-1}\boldsymbol{p} \cdot 2^{-1}, \qquad \text{where} \qquad \mathcal{L}(\theta) = \log \pi(\theta). \tag{B.3}$$

△

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

$$f(\theta, \boldsymbol{p}) \propto \pi(\theta) \cdot \exp\{-\boldsymbol{p'M^{-1}p} \cdot 2^{-1}\} \propto \exp\{-\mathcal{H}(\theta, \boldsymbol{p})\}. \tag{B.4}$$

Following Hartmann and Ehlers [2016], the idea is to use the Hamiltonian dynamics equations (not shown here) to model the evolution of a particle that keep the total energy constant. Introducing the auxiliary variables $\boldsymbol{p}$ and using 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* which will introduce discretization. A MH acceptance step is then required to correct the error and ensure convergence. The new proposal $\mathcal{X}_* = (\theta_*, \boldsymbol{p}_*)$ will be accepted with probability

$$\alpha(\mathcal{X}, \mathcal{X}_*) = \min\left[\frac{f(\theta_*, \boldsymbol{p}_*)}{f(\theta, \boldsymbol{p})}, 1\right] = \min\Big[\exp\{\mathcal{H}(\theta, \boldsymbol{p}) - \mathcal{H}(\theta_*, \boldsymbol{p}_*)\}\ ,\ 1\Big]. \tag{B.5}$$

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

Algorithm 4: 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 $\alpha\Big[(\theta_I, \boldsymbol{p}_I, (\theta_L, \boldsymbol{p}_L)\Big] = \min\Big[\exp\{H^{(0)} - H^{(1)}\},\ 1\Big]$,
   (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).

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

B.2 Convergence Diagnostics

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

Discussed in Section **4.4**, it is important that the chains mix well. Having J chains $\theta_{i,j}$ of same size I each with different starting values, the goal is to estimate the Between and Within-sequences posterior variances

$$B = \frac{I}{J-1}\sum_{j=1}^{J}\left(\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot}\right)^2 \quad \text{and} \quad W = \frac{1}{J}\sum_{j=1}^{J} s_j^2 \quad \text{where } s_j^2 = \frac{1}{I-1}\sum_{i=1}^{I}\left(\theta_{ij} - \bar{\theta}_{\cdot j}\right)^2.$$

If all the chains have converged to the stationary distribution, they will share the same limiting posterior variance, with W and B being unbiased estimates of the variance. An unbiased estimate of the overall variance is given by

$$\hat{V}(\theta|\boldsymbol{x}) = \frac{I-1}{I}\cdot W + \frac{1}{I}\cdot B. \tag{B.6}$$

Hence, Gelman and Rubin [1992] suggested to compute

$$\hat{R} = \frac{\hat{V}(\theta|\boldsymbol{x})}{W}, \tag{B.7}$$

which is the factor by which one can expect to reduce $\hat{V}(\theta|\boldsymbol{x})$ if it were computed after convergence.

It is recommended to continue sampling until values below ≈ 1.1 are obtained. Having a $\hat{R}$ close to 1 is a necessary but not sufficient condition for convergence. The choice of the starting values for each chains is critical, and should be scattered over the parameter space. Informal plots displaying each chains can also be used to assess the mixing of the chains.

Geweke diagnostic

Proposed by Geweke [1992], this test compares the mean of the first portion (e.g. 10%) of a generated chain with the mean computed from the second half of the chain. The large buffer is taken between the two blocks with the hope that they can be assumed independent. Classical frequentist z-score test based on the effective sample sizes is used. The rejection of the hypothesis can be due for example to a too short chain, or a too short burnin.

Practically, the procedure starts with the first half of the chain partitioned in K consecutive blocks of iterations. The testing procedure is then applied repeatedly on the chain progressively truncated from its first $k = 0, \ldots, K$ blocks. The minimum number of iterations required to plead convergence corresponds to the value at which the hypothesis of mean equality starts not to be rejected.

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

The *within* or *between* dependence in the generated chains influence the accuracy of the posterior estimates. As dependence becomes stronger, we must increase the run-length N to achieve the same precision. There exists two problems of correlations in the output delivered by a MC.

- **Autocorrelation** which calculates the serial correlation within a single chain being monitored. High autocorrelations within chains indicate slow mixing and slow convergence, and perhaps a need for reparametrization.
- **Cross-correlation** is the correlation between the monitored variables for each chain. High correlation among parameters also slower convergence and may also indicate a need for reparametrization.

The dependence structure within a single chain can be approximated by an AR(1), i.e.

$$\theta_t = \mu + \rho \cdot (\theta_{t-1} - \mu) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

At stationarity, i.e. when $\rho \in (0, 1)$, the mean $\bar{\theta}$ of n elements of the corresponding chain has variance

$$V(\bar{\theta}) = \sigma^2 \cdot \left(N \cdot \frac{1-\rho}{1+\rho} \right)^{-1},$$

and hence, we can calculate the *effective sample size*

$$\begin{aligned} N_{\text{eff}} &= N \cdot \left(\frac{1-\gamma(1)}{1+\gamma(1)} \right) \\ &= N \cdot \left(1 + 2 \sum_k \gamma(k) \right)^{-1}, \end{aligned} \tag{B.8}$$

following Hartmann and Ehlers [2016] for the second equality, where $\gamma(k)$ is the observed monotone k-th order autocorrelation.

Raftery and Lewis's Diagnostics The diagnostic of Raftery and Lewis [1992] is a run length control diagnostic based on a criterion of accuracy of estimation of a certain quantile, in order to inform whether the number of iterations is too small. The criterion is based on the autocorrelation inside the generated samples, and informs us about the minimum sample size required for a chain with no correlation between consecutive samples.

Thinning It is a method that aims at reducing the autocorrelation of a chain by storing iteration k only if $k \mod t_h$ is zero (and if k greater than or equal to the burn-in B), where t_h is called the thinning interval. The iterations that have been stored are assumed to be sampled from the same target distribution of the original chain, but with reduced dependence within the chain. Following Link and Eaton [2012], thinning will typically reduces the precision of posterior estimates since it reduces the number of kept samples, but it may still represent a necessary computational saving.

B.3 Additional Figures and Tables

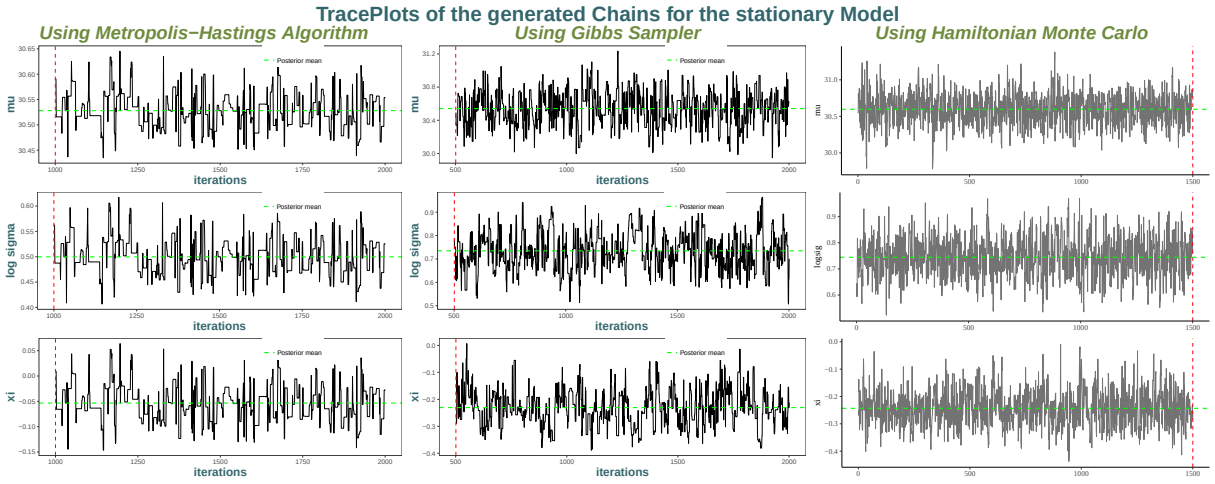


Figure B.1: *Traceplots of the single chains chains generated by the three usual Bayesian algorithms considered. Acceptance rate is* ≈ 0.21 *of the MH and individual acceptances rates all between* 42% *and* 50% *for the Gibbs sampler. In the HMC,* 49 *iterations are divergent and the acceptance rate statistic is of* 94%*. We ran one single chain for all algorithms with the same starting values.*

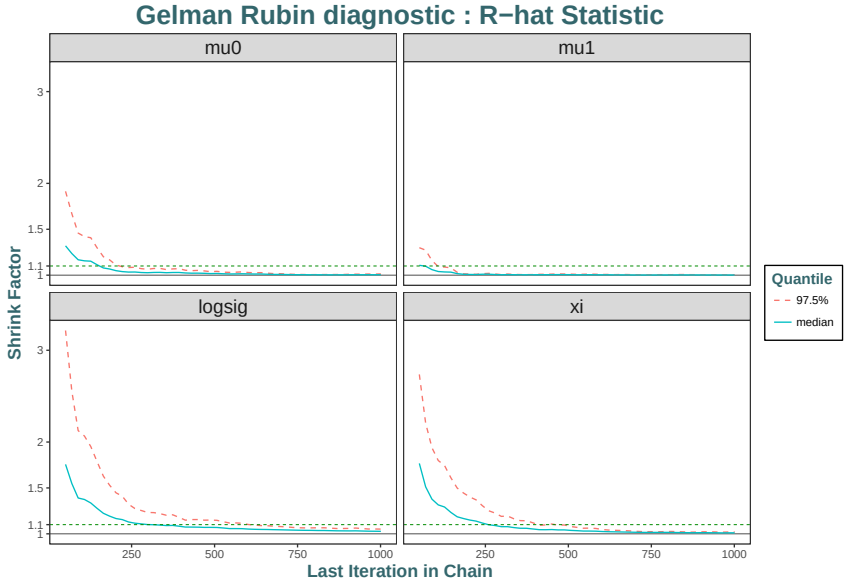


Figure B.2: *Gelman-Rubin diagnostic :* $\hat{R}$ *statistic computed by each repeating blocks of iterations, for each parameters. We refined the basic plot provided by* `coda` *in order to put the four graphs on the same* y*-scale, for comparison purposes.*

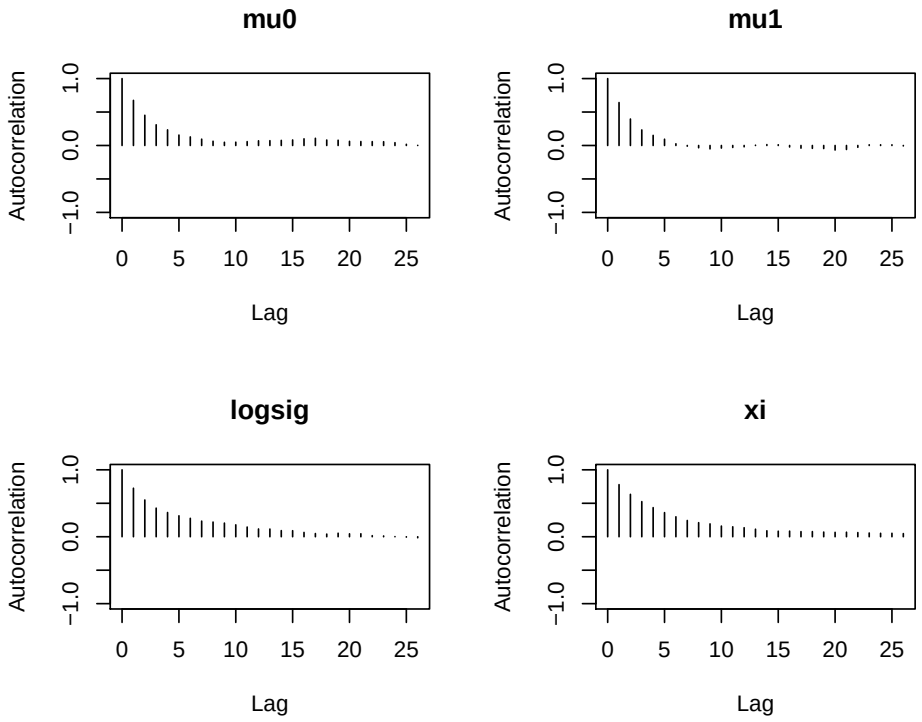


Figure B.3: *Autocorrelation functions for each of the parameters' Markov chains for a maximum lag of 25. Output provided by* coda.

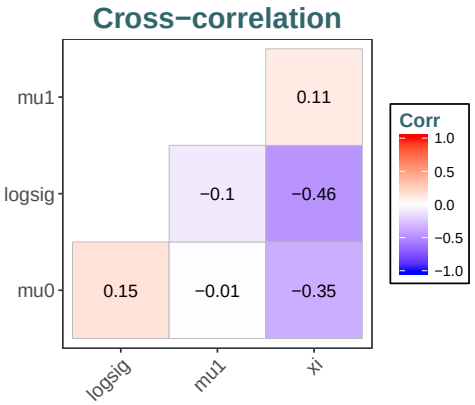


Figure B.4: *Cross-correlation between each of the parameters' Markov chains.*

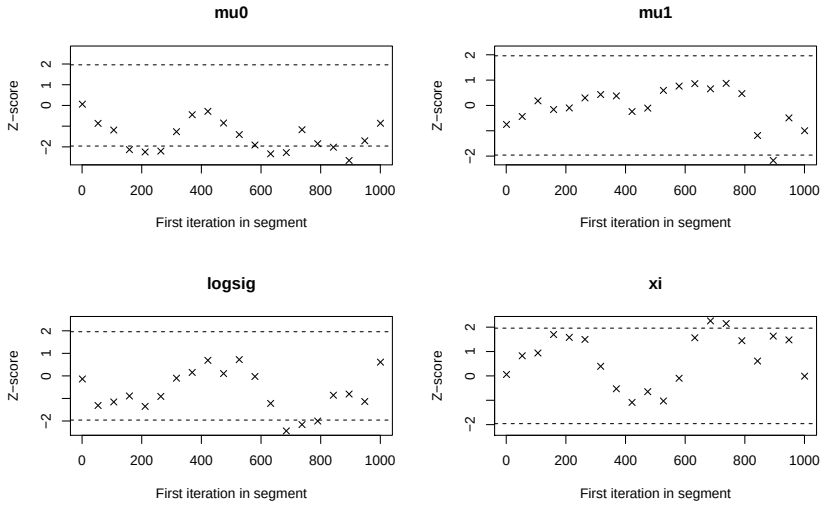


Figure B.5: *Geweke diagnostic : compute* 20 *z-scores that test the equality of the means between* 10% *and* 50% *of the chains. Dotted horizontal lines indicates the confidence regions at* 95%.

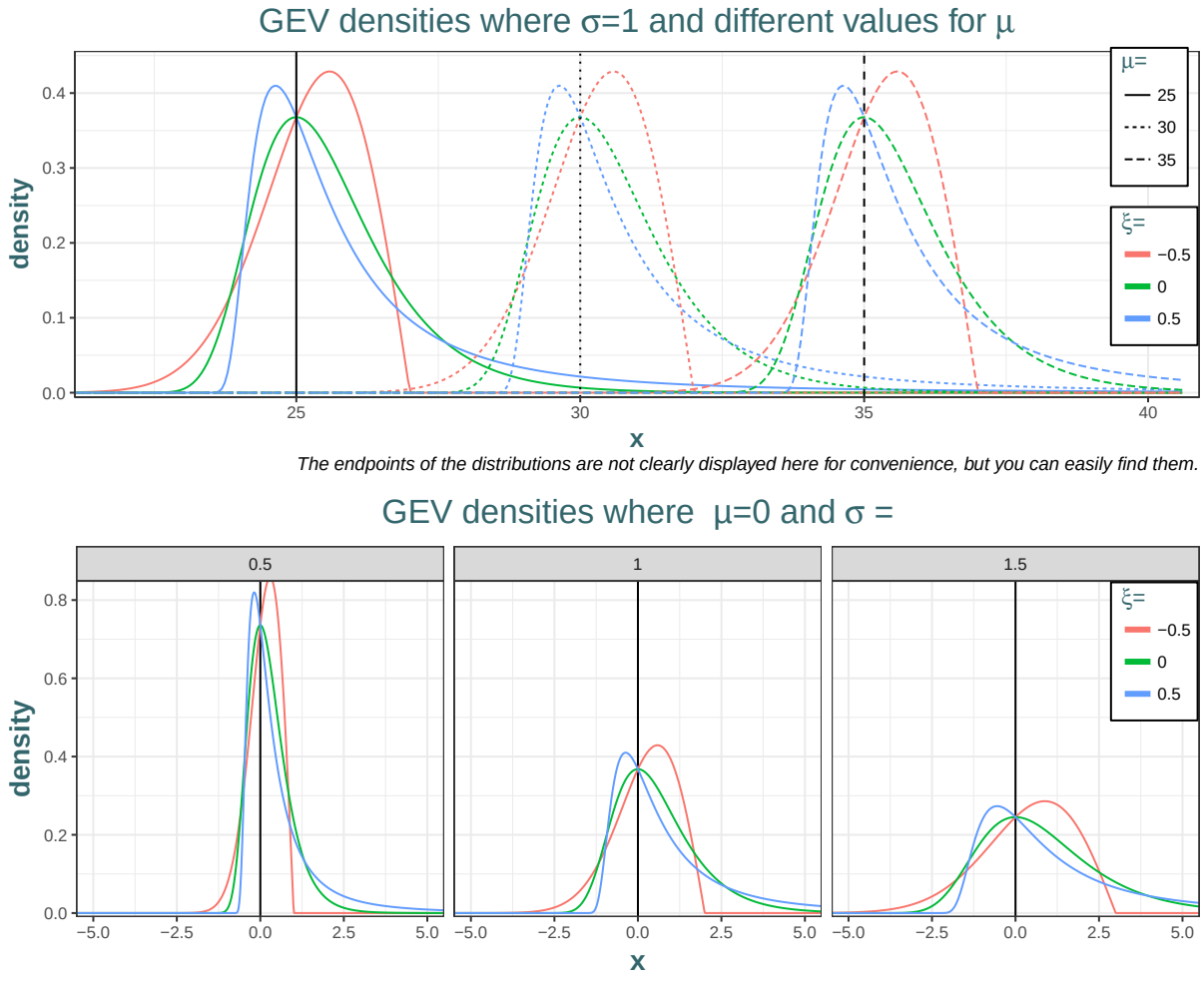
	B	N_{advised}	N_{min}	Dependence Factor
μ_0	17	2189	457	4.790
μ_1	13	1689	457	3.700
$\log \sigma$	12	1573	457	3.440
ξ	16	2110	457	4.620

Table B.1: *Raftery-Lewis diagnostic. "B" is the avdised number of iterations to be discarded at the beginning of each chain. "$N_{advised}$" is the advised number of iterations. "N_{min}" is the minimum sample size based on zero autocorrelation. The "dependence factor" informs to which extent the autocorrelation in the chains inflates the required sample size, with values above 5 indicating a strong autocorrelation.*

Appendix C

Other Figures and Tables

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



The endpoints of the distributions are not clearly displayed here for convenience, but you can easily find them.

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

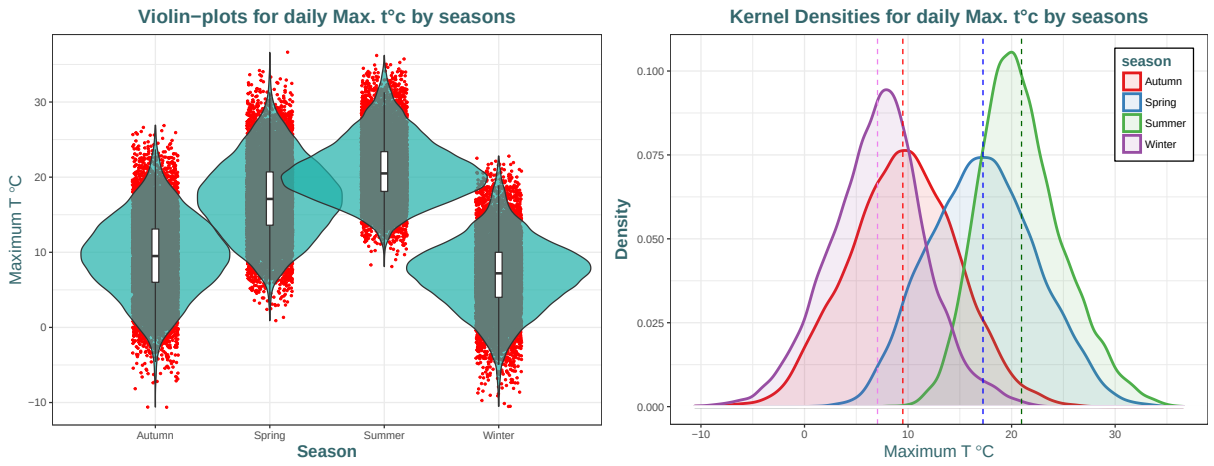


Figure C.2: *Violin-plot (left) and density plots (right) for each seasons. In the density plots, vertical dotted lines represent the mean of each distribution.*

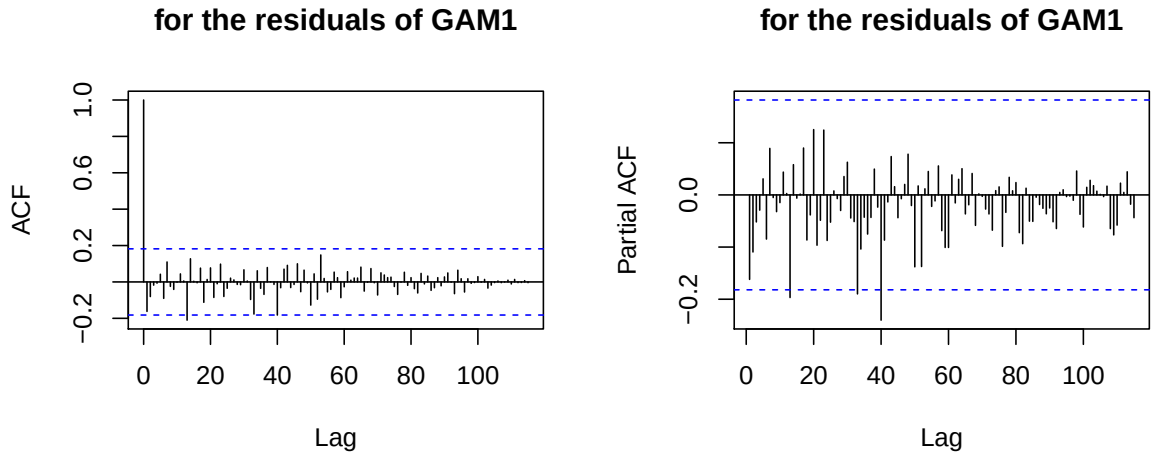


Figure C.3: *ACF and PACF for the residuals of the fitted GAM model with assumed independent errors*

Table C.1: *Models' comparisons for the residuals of the GAM model based on AIC and BIC criterion.*

	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

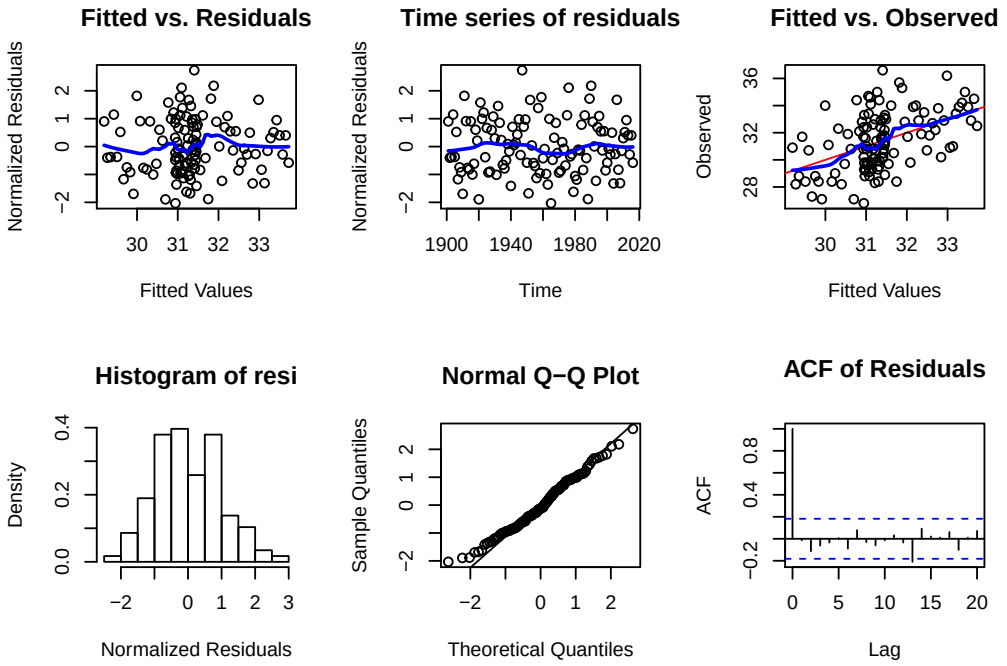


Figure C.4: *Diagnostics of the chosen GAM model with Whinte Noise process on the errors, based on the residuals.*

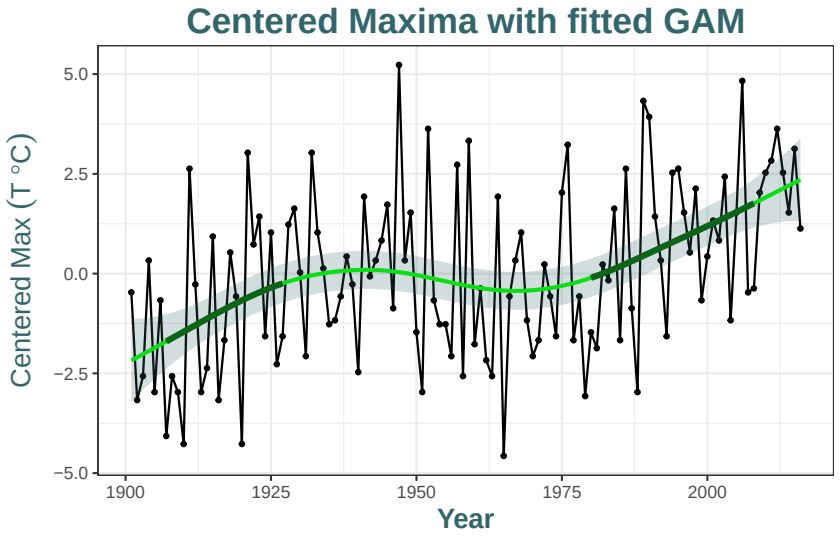


Figure C.5: *Series of annual maxima together with the fitted GAM model (in green)* ***with MA(1) model on the residuals****. Thicker lines indicate that the increase is significant for pointwise confidence interval. Shaded area represent a "95%" interval for the predicted values which looks quite narrow.*

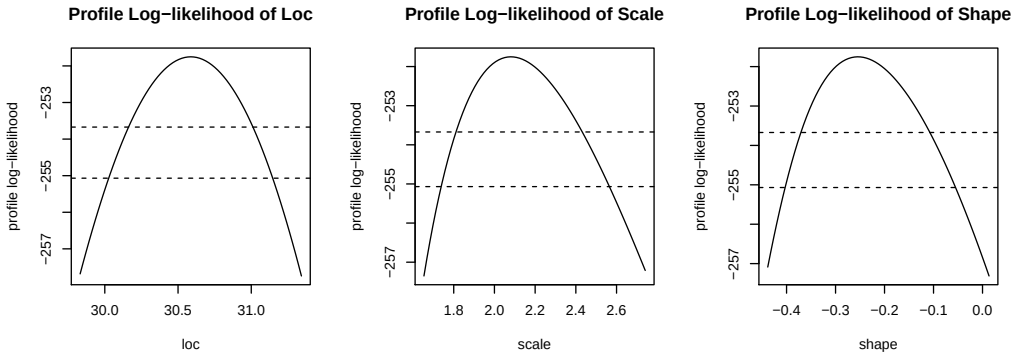


Figure C.6: *Profile likelihood intervals for the stationary GEV parameters. The two horizontal dotted lines represent the* 95% *(above) and* 99% *(below) confidence intervals by taking the intersection with the horizontal axis.*

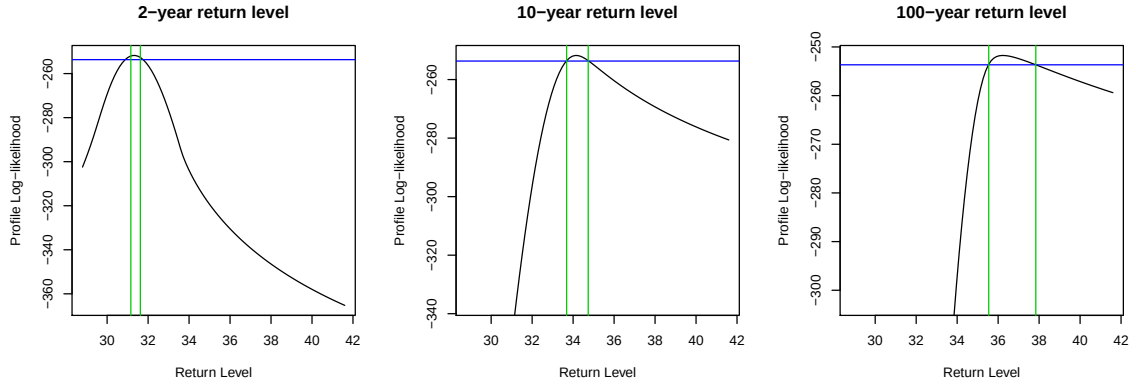


Figure C.7: 95% *Profile likelihood intervals for return levels with return periods of* 2, 10 *and* 100. *We kept the same* x*-scales for the three plots but not the* y*-scales. We used the* `ismev` *package but we modified the function to allow for more flexibility because the default y-scale in produced ugly visualizations for high return levels. Green lines represent the intervals from Table 6.3 computed with another package from E.Gilleland,* `extRemes`.

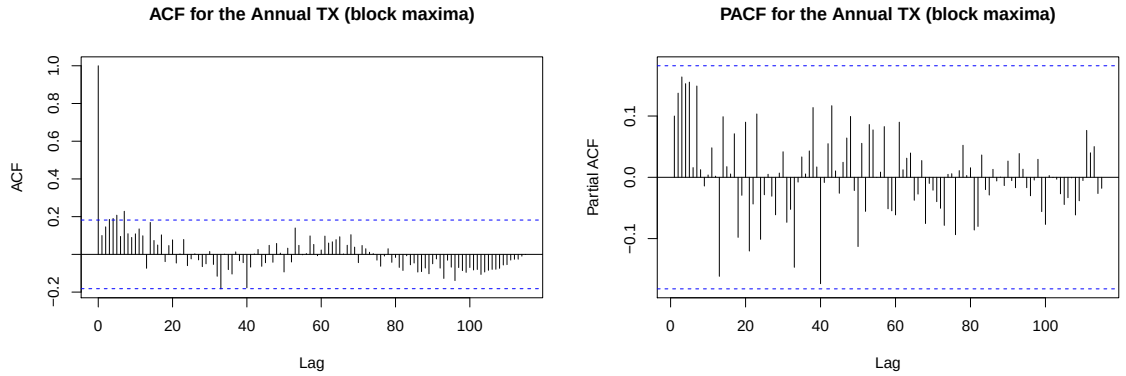


Figure C.8: *Autocorrelation functions for the series of annual maxima.*

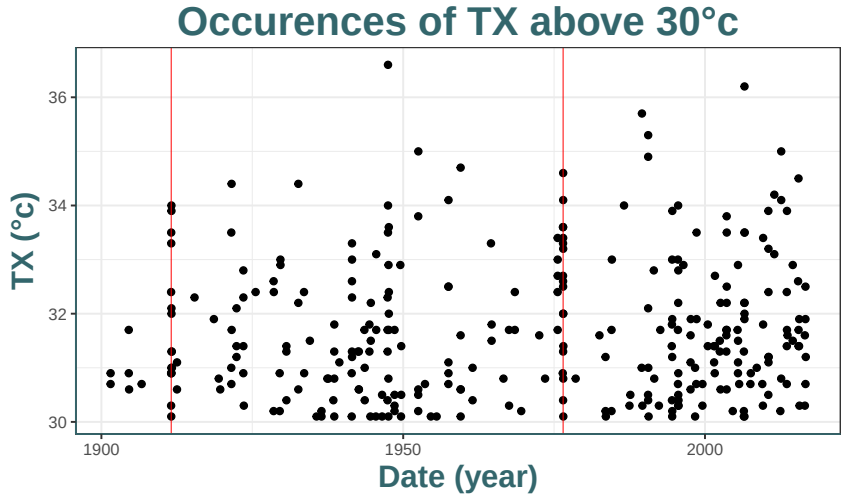


Figure C.9: *Plot of all daily TX that exceeded* $30°c$ *in the period [1901,2016] in Uccle. Red lines highlights two periods of heavy heat waves during summers 1911 and 1976.*

	Location β_0	Location β_1	Scale α_0	Scale α_1	Shape ξ
Estimates	29.19	0.0242	0.686	-0.0012	-0.199

Table C.2: *Estimation of the bootstrap aggregated GEV-CDN model with 2 hidden layers for* $\sigma(t)$ *and* $\mu(t)$*, and* $M = 500$ *resamples. In red are denoted rough estimates of the nonstationary parameters as if the model were parametric for both parameters* $\sigma(t) = \exp(\alpha_0 + \alpha_1 \cdot t)$ *and* $\mu(t) = \beta_0 + \beta 1 \cdot t$*. But this is actually not reliable since there are 2 hidden layers, and only the shape parameter can be reliably estimated.*

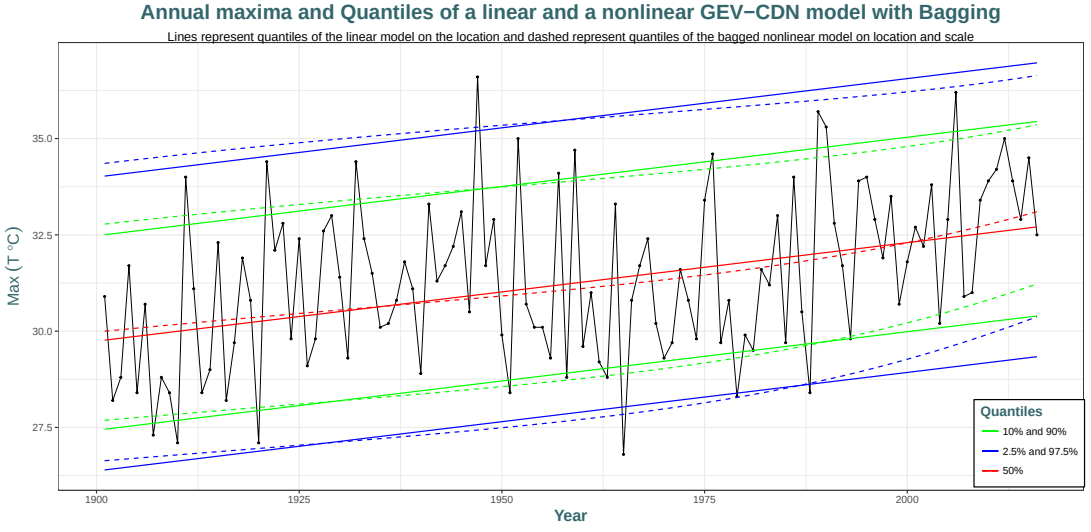


Figure C.10: *This graph gathers the two plots of Figure 6.5 which could yield to a better visualization.*

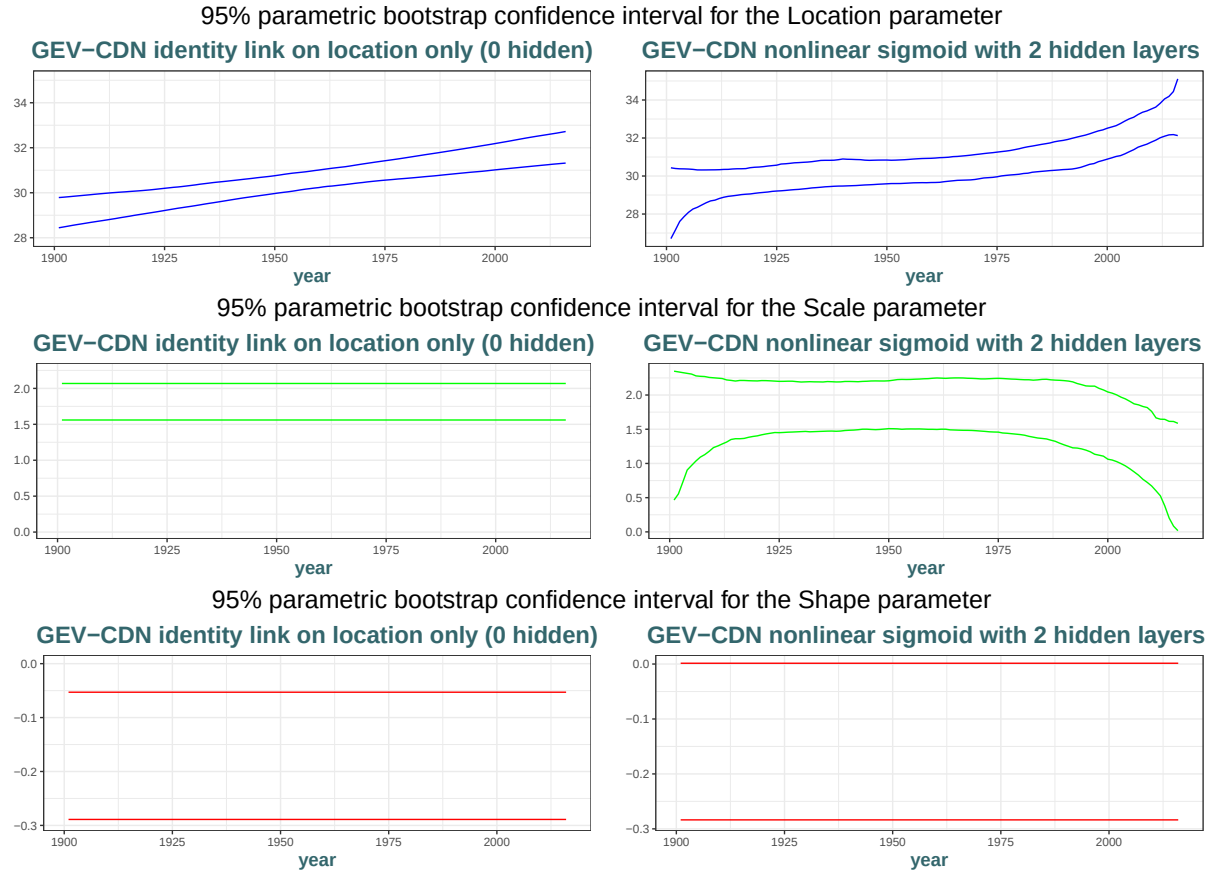


Figure C.11: *Left plots show the* ***parametric*** *bootstrap* 95% *intervals computed with the GEV-CDN model allow a linear nonstationary location parameter only, and Right plots show these interval for the nonlinear nonstationary model in location and scale parameters with 2 hidden layes.*

Appendix D

Github Repository : Structure

From the huge amount of code needed for the analysis, we decided to divide it so that it will be more conveniently used and read. It was also preferred to create a R package to gather all the functions created. 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. The README file contains valuable information to install the package and provide an overview of the Shiny applications. It has the following structure, with `files` divided in **/folders/** :

- **/R/** : Folder containing the scripts with all the functions that have been build for this thesis and are made available through the package.
  - ▷ `1UsedFunc.R` : functions created for the first part, including the stationary analysis of annual maxima in GEV, analysis in POT, nonstationary analysis in GEV and POT, etc.
  - ▷ `BayesFunc.R` : functions for the Bayesian Analysis, e.g. the Metropolis-Hastings and Gibbs Sampler, for both stationary and nonstationary). It includes information criteria, diagnostics with several plotting function, etc. Dey and Yan [2016, chap.13] gave most insights that helped us to build the functions.
  - ▷ `NeuralNetsFunc.R` : Refined functions from Cannon [2010] to compute GEV-CDN models, in order to yield more convenient outputs for the nonstationary analysis with NN than those provided by the GEV`cdn` package, especially useful for our plots.
  - ▷ `BootstrapFunc.R` : functions to compute (double) bootstrap confidence intervals **(not updated and not used in the text)**.
  - ▷ `runExample.R` : function allowing to run the Shiny applications directly through the package : by putting the name of the application in ' ' in the function, and it will run the application.

  The documentation of the functions are directly made through the `roxygen2` package infrastructure. You can then access the help of each functions as usual in R, by typing `?function's_name`. *We were not able to fully complete the documentations of every functions (e.g. for Bayesian functions). Furthermore, some functions have been left in the scripts.*
- **/Scripts-R/** : Folder containing scripts that allow to retrieve any analysis made during this text.

- ▷ `1GEV_plots_(chap1).R` and `1GEV_ggplot_(chap1).R` : code to compute the plots in Chapter **1**. Second script contains the code to construct the displayed plots, made with `ggplot2`.
- ▷ `1POT_ggplot.R` : script used to obtain Figure 2.1 from our data.
- ▷ `1intro_stationary.R` : code that provides the introduction to the data (preprocessing + descriptive analysis ; see beginning of Chapter **5**). It also yields the GEV stationary analysis (with a little part for POT) of Section **6.1**. It provides additional analysis by taking other block-lengths (months, 6months, etc) or by taking minima.
- ▷ `1intro_trends(splines).R` : code used to compute the trend analysis in Section **5.3.3**, i.e. the splines analysis with GAM model ; the comparisons of simultaneous and pointwise intervals, coverage analysis, etc.
- ▷ `2NeuralsNets.R` : code for the Neural Network's analysis by GEV-CDN made in next Chapter **6.3** ; including selecting of the best model relying on information criteria, bagging, computing confidence intervals with bootstrap, etc.
- ▷ `2Nonstationary.R` : code for the parametric nonstationary analysis made in Section **6.2**, but also with POT or point process.
- ▷ `Bayes_evdbayes.R` : code used for the Bayesian analysis with the `evdbayes` package ; stationary GEV and nonstationary GEV (did not converge), POT.
- ▷ `Bayes_own_gev.R` : code that computes all the Bayesian analysis made with our own functions (in `BayesFunc.R`). From the Gumbel model to more complex parametric nonstationary models, with model comparison, convergence analysis, inference and visual diagnostics, etc.
- ▷ `Bayes_stan.R` : code to execute the Bayesian GEV models written in **/stan/** with the `rstan` package, together with several visualizations and diagnostics.
- ▷ `Funs_introSplines.R` : code that contains all the functions used in the script `1intro_trends(splines).R`. These functions are thus not included in the package (yet?) and we just sourced this script.
- ▷ `Shiny_stan_test.R` : code to deploy the Shiny application provided by the `shinystan` package, from a Stan model executed through `rstan`

- **/Shiny_app_visu/** : one way to include Shiny applications inside the package **(not updated)**.
- **/data/** : Folder that only contains the annual maxima and minima data in Uccle in a .RData format. These data can be directly used inside the Shiny application, and hence these applications can be run in any local environment only having loaded the `PissoortThesis` package. We were not able to provide all the data used during this thesis since the IRM wanted keep it confidential.
- **/inst/** : Folder that contains the code that generates the Shiny applications. This code is divided in sub-folders for each applications. See Section **5.1** for more information on the applications.
- **/man/** : This folder contains automatically generated files by `roxygen2` that yields the documentation of each functions. These `.Rd` files are generated from the ' descriptor portion' that are above the functions contained in the files in **/R/**.

- **/stan/** : Folder that contains the STAN scripts, trying to have convergence to the stationary posterior distribution ....
- **/vignettes/** : Folder that contains two "vignettes" made for appointments with J.Segers during the academic year. It can be downloaded in `.html` from the compressed file.
  - ▷ `Summary1_intro.Rmd` : introduction to the analysis, descriptive statistics, stationary GEV model, first nonstationary GEV analysis and model comparison, first try with GEV-CDN, first attempt with Bayesian analysis with `evdbayes`.
  - ▷ `Summary_Bayesian` : Bayesian analysis with Metropolis-Hastings and Gibbs Sampler (for stationary GEV), and then complete Bayesian analysis for the nonstationary GEV model with a linear model on the location parameter ; all this computed with our "own" functions.

Bibliography

A. AghaKouchak, D. Easterling, K. Hsu, S. Schubert, and S. Sorooshian. *Extremes in a Changing Climate: Detection, Analysis and Uncertainty*. Springer Science & Business Media, Oct. 2012. ISBN 978-94-007-4478-3.

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.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002. ISSN 0016-6731. URL http://www.genetics.org/content/162/4/2025.

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. A Conceptual Introduction to Hamiltonian Monte Carlo. *ArXiv e-prints*, Jan. 2017.

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.

C. S. Bos. *A Comparison of Marginal Likelihood Computation Methods*, pages 111–116. Physica-Verlag HD, Heidelberg, 2002. ISBN 978-3-642-57489-4. doi: 10.1007/978-3-642-57489-4_11. URL https://doi.org/10.1007/978-3-642-57489-4_11.

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.

M. Bédard. Optimal acceptance rates for metropolis algorithms: Moving beyond 0.234. *Stochastic Processes and their Applications*, 118(12):2198 – 2222, 2008. ISSN 0304-4149. doi: http://dx.doi.org/10.1016/j.spa.2007.12.005. URL http://www.sciencedirect.com/science/article/pii/S0304414907002177.

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. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. doi: 10.1080/01621459.1995.10476635. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476635.

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.

S. G. Coles and E. A. Powell. Bayesian methods in extreme value modelling: A review and new developments. *Extremes*, Vol. 64, No. 1 (Apr., 1996)(1):119–136, 1996. URL

http://www.jstor.org/stable/1403426?origin=crossref&seq=1#fndtn-page_thumbnails_tab_contents.

S. G. Coles and J. A. Tawn. A bayesian analysis of extreme rainfall data. pages 153–178, 1996.

M. Crowder. Bayesian priors based on a parameter transformation using the distribution function. *Annals of the Institute of Statistical Mathematics*, 44(3):405–416, Sep 1992. ISSN 1572-9052. doi: 10.1007/BF00050695. URL https://doi.org/10.1007/BF00050695.

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.

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 and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7 (4):457–472, 11 1992. doi: 10.1214/ss/1177011136. URL http://dx.doi.org/10.1214/ss/1177011136.

A. Gelman, C. Robert, N. Chopin, and J. Rousseau. *Bayesian Data Analysis*. 1995.

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.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, Nov. 1984. ISSN 0162-8828. doi: 10.1109/TPAMI.1984.4767596. URL http://dx.doi.org/10.1109/TPAMI.1984.4767596.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. pages 169–193, 1992.

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.

H. Jeffreys. *Theory of probability*. Oxford University Press, Oxford, England, third edition, 1961.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572.

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.

V. V. Kharin, F. W. Zwiers, X. Zhang, and G. C. Hegerl. Changes in Temperature and Precipitation Extremes in the IPCC Ensemble of Global Coupled Model Simulations. *Journal of Climate*, 20(8):1419–1444, Apr. 2007. ISSN 0894-8755. doi: 10.1175/JCLI4066.1. URL http://journals.ametsoc.org/doi/full/10.1175/JCLI4066.1.

A. N. Kolmogorov, N. Morrison, and A. T. Bharucha-Reid. *Foundations of the theory of probability*. Chelsea Publishing Company, New York, 1956. OCLC: 751236060.

P. Lafaye de Micheaux and B. Liquet. Understanding convergence concepts: A visual-minded and graphical simulation-based approach. *The American Statistician*, 63(2):173–178, 2009. URL http://amstat.tandfonline.com/doi/abs/10.1198/tas.2009.0032.

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.

W. A. Link and M. J. Eaton. On thinning of chains in mcmc. *Methods in Ecology and Evolution*, 3(1):112–115, 2012. ISSN 2041-210X. doi: 10.1111/j.2041-210X.2011.00131.x. URL http://dx.doi.org/10.1111/j.2041-210X.2011.00131.x.

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.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114. URL http://dx.doi.org/10.1063/1.1699114.

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.

R. A. Naveau, Philippe, F. W. Zwiers, and J.-M. Azaïs. A new statistical approach to climate change detection and attribution. *Climate Dynamics*, 48(1):367–386, Jan 2017.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248.

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. Northrop. Revdbayes: Ratio-of-uniforms sampling for bayesian extreme value analysis. 2017. URL https://CRAN.R-project.org/package=revdbayes.

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.

A. E. Raftery and S. M. Lewis. [practical markov chain monte carlo]: Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statist. Sci.*, 7(4):493–497, 11 1992. doi: 10.1214/ss/1177011143. URL http://dx.doi.org/10.1214/ss/1177011143.

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. 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.

B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

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.

Stan Development Team. *Stan Modeling Language User's Guide and Reference Manual, Version 2.14.0*. 2017. URL http://mc-stan.org/.

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.