Universitá degli Studi di Padova

Dipartimento di Ingegneria Civile,

Edile ed Ambientale

Tesi di Laurea

On the use of Metastatistics in extreme rainfall analysis

Relatore: Prof. Gianluca Botter

Correlatore: Prof. Marco Marani

Laureando: Enrico Zorzetto

Anno Accademico 2014-2015

A Mario e Ivana,
perché questo é il frutto
non delle mie fatiche ma delle vostre.

This Master thesis was prepared at the Department of Civil and Environmental Engineering (DICEA) at the University of Padova in fulfillment of the requirements for acquairing a M.Sc. degree in Civil Engineering. The work that led to the realization of this dissertation was carried out from May 2014 to February 2015 under the supervision of Professor Gianluca Botter from department DICEA of University of Padova and with the support of Professor Marco Marani, from the department of Civil and Environmental Engineering, Duke University.

Padova, March 2015

Enrico Zorzetto

Contents

List of Figures

List of Tables

Introduction

The assessment of weather extremes is matter of paramount importance in engineering and a cornerstone in the field of hydrology. The design of any civil work meant to manage water resources is based on a proper understanding of the risk of the occurrence of events whose magnitude would cause the system's failure, possibly causing harm to human well being or the environment. The task is not easy, since the amount of available historical observations is always limited and often the design purposes require extrapolation outside the range of the data at hand. Standard methodology to tackle the problem consists in the adoption of asymptotic models to describe the distribution of the extreme values of the natual process at hand. Hence, a suitable model is fitted to the available data and forecast is then possible on the basis of the inferred model. Despite the vast number of studies that in the last century have contributed to define the Classical Extreme Value theory (EVT), the problem is somehow still open.Traditional methods fail with an embarassing frequency and lead, in several circumstances, to serious underestimation of the occurrence of extremes. The shortcoming of this methods is inborn in the nature of the problem itself: when we study the extremes of a given process, we do not deal with the bulk of its probability distribution, but with the tails, where only a small fraction of the observations lie. For this reason, the uncertainties involved play a key role and may lead to severe errors in high quantiles estimations. Some authors (e.g. Coles, 2002) argue that the most natural approach to takle these uncertainties is to adopt a Bayesian estimation of the parameters of the extreme value distributions. The advantage of this approach is that the parameters of the distribution are considered as random variables and the outcome of the analysis yields the entire posterior distribution of the parameter set. On the other hand, the Bayes' Law requires the knowledge of the prior distribution of the parameters, representing our belief priory to the information conveyed by the available data. The shortcoming of the method is that in the case of rainfall extremes such a prior information is missing and therefore there is no legitimate reason to rely on other than the available observations. The present work is based on a different approach: If we assume the random process to be ergodic (i.e. its dinamics represents the entire space of all the system's states, referred to as phase space) then we can estimate the distribution of the parameters by fitting a distribution to all the single years of a given window, obtaining a sample of estimated parameters which gives us some information about the distribution of the parameters themselves. This approach is referred to as 'Metastatistics of Extreme Values' (Marani and Ignaccolo, 2014) since the higher order randomness inborn in the estimated parameters is explicitely taken into account. Moreover, this new framework allows to remove the asymptotic hypothesis and to consider the cardinality of the process (i.e. the number of events per year) to be a random variable as well; In this way the probability distribution of the annual maxima can be described in a fully probabilistic way. The goal of the present work is to compare the performance of this new Metastatistic approach with the traditional models used to describe extreme rainfall events in a stationary framework. The dissertation is structured as follows: Chapter 1 consistes in a brief summary of the classical extreme value statistics, deriving the three limiting probability laws and describing the two main methods currently used to model the tail of rainfall distributions, namely the annual maxima approach (AM) and the point process methods, e.g. the Peak Over Threshold (POT) method. In chapter 2 an alternative derivation of the limiting distributions is derived and its application explored, evalutating the possibility to remove the asymptotic assumption and to use the so called *Penultimate Approximation* instead of the classical asymptotic probability laws. In chapter 3 the Metatstatistical distribution of Extreme Values (MEV) is introduced, removing some of the hypotheses on which the classical theory relies, and discuss its applications in the context of daily rainfall analysis. Chapter 4 describes the data sets used in the analysis and reports some statistical tools used to obtain a first characterization of the tail behaviour of the rainfall records. In chapter 5 the more common fitting methods for Weibull, Generalized Pareto and Generalized Extreme Value distributions are discussed and compared. In chapter

6 MEV and GEV performances are compared in the context of stationary extreme rainfall analysis. The chapter opens describing the method implemented to obtain a benchmark of the two models in a stationary framework and then the main results of the application are presented. In the last chapter a possible application of the MEV approach in a non-stationary framework is discussed, analyzing the interannual variability of rainfall extremes. The Metastatistical approach is speculated to be the most natural and robust way to incorporate nonstationarities or climatic covariates in extreme rainfall analysis.

Ringraziamenti

Giunto alla fine di questa tesi e di questi anni da studente ci sono parecchie persone, dentro e fuori l'universitá, che sento dover ringraziare:
I relatori Gianluca Botter e Marco Marani, che sono stati una guida fondamentale durante gli ultimi mesi; mi hanno trasmesso la passione per l'idrologia e mi hanno anche lasciato la libertá di fare 'un po di testa mia', che ho apprezzato molto.
I professori in gamba che ho avuto la fortuna di incontrare in questi anni, in particolare Luigi d'Alpaos, il quale ha trovato il tempo per insegnarci, oltre ai fatti dell'idrodinamica, anche a non scendere mai a compromessi ne come ingegneri e ne come uomini.
I compagni allievi ingegneri, in particolare Carol, Elisa, Daniele e Sara, perché oltre ad aver affrontato fianco a fianco le sfide che l'universitá ci ha messo davanti, avete sopportato le mie battute senza lamentarvi (troppo).
I compagni di cordata Frank, Filippo e Nicola con cui non sono ancora stato sul Civetta, ma con cui rimedieró.
I miei amici 'Padovani': Paola, perché la sua amicizia che é stata un punto fermo per tanti anni, Andrea e Gian per i tanti momenti che abbiamo condiviso a Padova, i Tiepolini Marco, Salvo e Lucia per avermi iniziato all' uso sconsiderato dei Post-it.
I miei amici Matteo, Pasquale, Nicholas, Checco, Puppo, Leo, Manuel, Marco, Francesca, Anna e Valentina.
I miei genitori che mi hanno sostenuto e sopportato per tutto questo tempo e non mi hanno tirato una 'spelazada sul da drio' neanche quando averbbero, in tutta fede, dovuto.
Le mie sorelle, in special modo Irene che ha corretto la grammatica di questa tesi (eventuali errori residui sono suoi, e suoi soltanto).
Memi e Isetta, che mi mancano e vorrei fossero qui oggi.

Chapter 1

Classical extreme value theory

1.1 A brief history of extreme values analysis

The statistical analysis of extreme values originated by the pioneering paper of Fischer and Tippett (1928) in which the extreme value theorem was first stated and the three fundamental limiting distributions were defined. Their findings were later strengthened by Von Mises, who in 1936 first proposed the expression of the Generalized Extreme Value distribution and identified the convergence conditions for the three limiting laws. The Extreme Value Theorem was then proved in the general case by Gnedenko in 1943. He also enunciated more formally the conditions for the weak convergence to the limiting laws, defining the corresponding domains of attraction. A lot of further contributions followed, refining the asymptotic theory and establishing the rate of convergence to the asymptotic laws for a wide range of distribution functions. Among others, the work of Gumbel (1958) is worth to be mentioned. Gumbel applied the extreme value methods to the statistical modelling of floods and spread the interest for this subject among hydrologists and practicioners engineers. In the seventies, the developement of the Peak Over Threshold method by Balkema and De Haan (1974) and then Pickands (1975) was a major leap in EV analysis which had deep impacts on hydrological applications. In the last two decades EV analysis has seen a widening range of applications in diverse fields such as insurance and finance, where the traditional modelling based on light tails was proved to fail in several circumstances (e.g. Mandelbrot and Taleb, 2005). Numerous contributions and theoretical progresses followed in the last decade. The main question remains which of the three limiting distributions should be used to model a given data set, yet some authors proposed different approaches to tackle the problem. Just to mention some of these newest contributions, recently Cook and Harris (2004) pointed out that relying on asymptotic models it is unnecessary and often may lead to significant errors in high quantile estimation; Instead, they proposed to use the penultimate approximation in order to reduce the error. Coles (2002) compared classic and Bayesian parametric estimators for the extreme value distributions and found that using a fully probabilistic approach, i.e. considering a second order randomness in the estimated parameters is the most natural approach and may lead to more accurate extreme rainfall appraisals. The debate in this fascinating field is still open and an actual 'battle of the extreme value distributions' is being fought (Papalexiou and Kudsoyiannis, 2013; Serinaldi and Kilsby, 2013) and the winner is still to be declared.

1.2 Extreme value theorem

Suppose $X_1, ..., X_n$ are random variables with a common cumulative distribution function (CDF) given by $F(x) = P(X_j \leq x) \quad \forall j = 1, ..., n, \quad \forall x \in \Omega$, where Ω is the common population of all the X_i. If X_i, X_j are independent for any $i \neq j$, then the distribution function of the maximum $M_n = max\left\{X_1, ..., X_n\right\}$ with cardinality n is given by the nth power of $F(x)$:

$$P(M_n \leq x) = F_n(x) = F(x)^n \tag{1.1}$$

The classical *extreme value theory* focuses on the asymptotic behaviour of this distribution, where the term 'asymptotic' refers to the weak convergence, or convergence in distribution. However the value of x sampled from the population of the random variable Ω, $0 < F(x) < 1$ holds for definition of CDF. Hence, when $n \longrightarrow \infty$, $F(x)^n \longrightarrow 0$. Thus, if we want to know the behaviour of such a distribution for large sample sizes we have to renormalize it by defining a family of series scaling costants a_n and b_n in order to obtain a nongenerate distribution (say,

$H(x)$) for large n:

$$\begin{aligned} P\left(\frac{M_n - b_n}{a_n} \leq x\right) &= P(M_n \leq a_n \cdot x + b_n) \\ &= F^n(M_n \leq a_n \cdot x + b_n) \\ &= \longrightarrow H(x) \quad as \quad n \longrightarrow \infty \end{aligned}$$

It turns out that there are only three *types* of limiting distributions H(x), the *Gumbel*, *Frechet* and *Weibull* distributions. This result is referred to as *'Extremal Types Theorem'* and was for the first time enunciated by Fischer and Tippett (1928) and later proved by Gnedenko (1943). Central to the understanding of this theorem are the two concepts of Type of a distribution and of Max-stable distribution.

Definition 1.1 Two distribution functions $H_1$and H_2 are said to be *of the same type* if one can be transformed into the other through a linear trasformation $H_1(x) = H_2(ax + b)$ where $a > 0$ and $b \in R$ are two suitable constants.

Definition 1.2 A non degenerate probability distribution function F is said to be a *Max-Stable distribution* if for a sequence of independent and identically distributed (i.i.d.) random variables $(X_i)_{i \in N}$ with common distribution F, and for each $n \in N$, there exist $a_n > 0$ and $b_n \in R$ such that $\frac{M_n - b_n}{a_n}$ also has distribution F. In other words, a CDF$F(x)$ is *Max-Stable* if, for each $n \in N$, there exist $a_n > 0$ and $b_n \in R$ such that

$$F^n\left(a_n x + b_n\right) = F(x) \quad for\ all\ x \in R \tag{1.2}$$

This property implies that taking any power of F results only in a change in location and scale parameters, not in a change in the *type* of the distribution. As a consequence, if we change the cardinality of the maximum we are considering, an extreme value model will be consistent i.e. there will be only a change in the parameters whereas the functional form of the distribution will be the same. For example, we might consider two models, one for the annual maxima and another for the N-years maxima of the same underlying process. The stability property guarantees that, since the second one will be the maximum of N annual maxima, the two models shall be mutually consistent. The same property holds if we consider a model for the exceedances over a given threshold: the model will remain consistent if we study exceedances over different thresholds. For this reason, the concept of Stability is of 'extreme' importance in modeling EV processes.

Theorem 1.3 A non degenerate probability distribution H is said to be Max-Stable if and only if there exist iid random variables $(X_i)_{i\in N}$ and two successions $a_n > 0$ and $b_n \in R$ such that the distribution of $\frac{M_n - b_n}{a_n}$ converges weakly to H.

Theorem 1.4 [Extremal types theorem] Every Max-Stable distribution H is of extreme value type i.e. it is of the same type as one of the three following distributions. Conversely, every distribution of extreme value type is Max-Stable.

- *Gumbel* :

$$H(x) = exp(-exp(-x)) \quad -\infty < x < \infty. \tag{1.3}$$

- *Frechet* :

$$H(x) = \begin{cases} exp(-x^{-\alpha}), & \alpha > 0, \quad \text{if } 0 < x < \infty, \\ 0, & \qquad\quad \text{if x <0.} \end{cases} \tag{1.4}$$

- *Weibull* :

$$H(x) = \begin{cases} exp(-(-x)^{-\alpha}), & \alpha > 0, \quad \text{if } -\infty < x < 0, \\ 1, & \qquad\quad \text{if x >0.} \end{cases} \tag{1.5}$$

Theorems 1.3 and 1.4 together imply that given $M_n = max_{1\leq i\leq n}\{X_i\}$ for a sequence of iid random variables $(X_i)_{i\in N}$, every non-degenerate limit of $\frac{M_n - b_n}{a_n}$ is of one of the three extreme value types. Conversely, every distribution of extreme value type is the weak limit of $\frac{M_n - b_n}{a_n}$ for two suitable successions $a_n > 0$ and $b_n \in R$ where we can choose $(X_i)_{i\in N}$ to have distribution H.

1.3 Criteria for attraction to the limiting types

The *Domain of Attraction* of a type H is defined as the set of distribution functions F such that, given a sequence of iid random variables $(X_i)_{i\in N}$ whose common cumulative distribution is F, the distribution of $\frac{M_n - b_n}{a_n}$ converges weakly to H for two suitable successions $a_n > 0$ and $b_n \in R$.

For sufficently smooth distributions ($F \in C^2(R)$) we can determine the limiting type H by defining the *Reciprocal Hazard Function* as:

$$r(x) = \frac{1 - F(x)}{f(x)} \tag{1.6}$$

and by defining:

$$b_n = F^{-1}\left(1 - \frac{1}{n}\right), \quad a_n = r(b_n) \tag{1.7}$$

Thus the limiting distribution of $\frac{M_n - b_n}{a_n}$ is

$$exp\{-(1 + \xi \cdot x)_+^{-1/\xi}\} \quad \text{if } \xi \neq 0 \tag{1.8}$$

$$exp(-exp(-x)) \quad \text{if } \xi = 0 \tag{1.9}$$

where the shape parameter can be determined as $\xi = lim_{x \to \infty} r'(x)$

1.3.1 Criteria for attraction to the Gumbel limiting type

Any survival distribution $1 - F(x)$ whose right tail dacays faster than any polinomial function for $x \longrightarrow \infty$ belong in the domain of attraction of the *Gumbel* distribution. This is the case for any exponential distribution of the form $1 - F(x) = e^{-h(x)}$ where $h(x)$ is any differentiable function positive and monotonically increasing faster than any power of $log(x)$. To be more precise, a distribution of the exponential family is in the domain of attraction of the Gumbel distribution if either of the following conditions are satisfied:

1. $h'(x) = x^{\alpha-1}L(x)$ for some $\alpha > 0$ and slowly varying (at ∞) function L, i.e. it satisfies $\lim_{x\to\infty} \frac{L(xy)}{L(x)} = 1 \quad \forall y > 0$

2. $h(x) = x^{\alpha}L(x)$ for some $\alpha > 0$ and slowly varying L, and h'(x) is monotone on (x_0, ∞) for some $x_0 > 0$.

This can be expressed by the *Von Mises condition*: Given a CDF F(x) and its derivative PDF f(x) *Gumbel* is the limiting distribution if and only if $\xi = lim_{x\to\infty} r'(x) = 0$, i.e. if:

$$\frac{d}{dx}\left\{\frac{1 - F(x)}{f(x)}\right\} \longrightarrow 0 \quad as \quad x \longrightarrow \infty \tag{1.10}$$

Only in this case there exist two constants a_n and b_n for which *Gumbel* is the limiting EV distribution. The easiest possible case, in which $h(x) = x$, that is the case of an exponential distribution with mean 1. Let $a_n = 1$ and $b_n = \log(n)$. Then

$$F^n(a_n x + b_n) = \left(1 - e^{-x-\log(n)}\right)^n = \left(1 - \frac{e^{-x}}{n}\right)^n \to \exp(-e^{-x}); \tag{1.11}$$

Yielding Gumbel as limiting distribution.

1.3.2 Criteria for attraction to the Frechet limiting type

The domain of attraction of the Frechet distribution consist of all the distribution function F(x) such that $r'(x) > 0$. Any F whose tail is of power law form $1-F(x) \sim x^{-\alpha}$, for $x \longrightarrow \infty$ for some costants $k > 0$ and $\alpha > 0$ is in the domain of attraction of the *Frechet* type with the same exponent α. This include, among others, the family of the *Pareto distributions.* Let consider the simple case of $1-F(x) = k \cdot x^{-\alpha}$ *(Pareto type I distribution).* Let $b_n = 0, \quad a_n = (n \cdot k)^{1/\alpha}$. Then for $x > 0$,

$$F^n(a_n x + b_n) = \left(1 - k(a_n x)^{-\alpha}\right)^n = \left(1 - \frac{x^{-\alpha}}{n}\right)^n \tag{1.12}$$

As $n \to \infty$ the right hand side of the equation converges to $exp(-x^{-\alpha})$, Frechet distribution with shape parameter α.

1.3.3 Criteria for attraction to the Weibull limiting type

Any F with a finite upper endpoint ω_F (i.e. if there exist ω_F such that $F(\omega_F) = 1$) and characterized by a power law behaviour as $x \longrightarrow \omega_F$ so that $1 - F(\omega_F - y) \sim ky^{\alpha}$ as $y \longrightarrow 0$ for some constants $k > 0$ and $\alpha > 0$ is in the domain of attraction of the Weibull type.

1.4 The GEV distribution

Von Mises (1936) proposed a single distribution which encompasses all three of the previous extreme value limit families:

$$H(x; \xi, \psi, \mu) = exp\left\{ -\left[1 + \frac{\xi}{\sigma} \cdot (x - \mu)\right]_+^{-1/\xi} \right\} \tag{1.13}$$

Defined on every x such that:

$$1 + \frac{\xi}{\sigma} \cdot (x - \mu) > 0 \tag{1.14}$$

Elsewhere the value of H is either 0 or 1. The parameter space is characterized by:

- $\mu \in R$ is the location parameter and indicates the value of x at which the pdf is centered (has a maximum).
- $\sigma > 0$ is the scale parameter and controls the spreading of the distribution around its location μ.
- ξ is a shape parameter determining the rate of tail decay. $\xi > 0$ gives the heavy tailed case (*Frechet*) in which the tail behaves like a power law with exponent $\alpha = 1/\xi$; For $\xi \to 0$ the GEV corresponds to the *Gumbel* distribution (light tailed case) and its decay is exponential, whereas negative values of ξ yield the short tailed case (*Inverse Weibull*): in this case the distribution is characterized by a finite upper endpoint and exponent $\alpha = -1/\xi$.

Let f(x) be the *GEV* probability density in the Gumbel Case ($\xi = 0$); For $x \to \infty$, $f(x) \sim e^{-x}$, so the right tail of the distribution decays with an exponential tail whereas the left tail decays as a double exponential function: $f(x) \sim exp(-exp(x))$ as $x \to -\infty$. In the Frechet case the GEV is left bounded at $x = 0$ and the right tail decays as fast as a pareto law with the same shape parameter: $f(x) \sim x^{-\alpha}$ as $x \to \infty$. In the Weibull case the EV distribution is right bounded whereas $f(x) \sim x^{-\alpha}$ as $x \to -\infty$.

The k-th moment of the GEV distribution exists if $\xi < 1/k$; e.g. the mean exists if $\xi < 1$ and the variance if $\xi < 1/2$. Mean and variance are respectively given by

$$E(X) = \mu + \frac{\sigma}{\xi} \left\{ \Gamma(1 - \xi) - 1 \right\}, \tag{1.15}$$

$$E\left\{ (X - E(x))^2 \right\} = \frac{\sigma^2}{\xi^2} \left\{ \Gamma(1 - 2\xi) - \Gamma^2(1 - \xi) \right\}. \tag{1.16}$$

In the limiting case $\xi \to 0$ these reduce to the mean and variance of the Gumbel distribution:

$$E(X) = \mu + \sigma * \gamma \tag{1.17}$$

$$E\left\{ (X - E(x))^2 \right\} = \frac{\sigma^2 \pi^2}{6} \tag{1.18}$$

Where $\gamma = 0.57720$ is Euler's constant. The first step required to fit the GEV distribution to a sample of rainfall 'maxima' is to decide what the maxima exactly are and to extract them from the available observations. The most commonly used approach is the so called *Block maxima method* in which a sequence of maximum values is extracted from blocks of equal length. In the case of daily rainfall records, usually the block length is one year and therefore this approach yields a data set of annual maxima. In this case the method is often referred to as *Annual Maxima method.* This method is commonly used both for its simplicity and for the fact that the annual maxima are without a shadow of a doubt independent variables; Its application was extensively studied in Gumbel's book (1958), which still is a milestone for engineering applications of the method.

1.5 Peak over threshold method

As we have seen in the previous section, the most immediate and widely used method to fit the GEV distribution to a series of observations in the *Block Maxima method.* Despite its simplicity, the shortcoming of this approach is that only one value from each block (year) is used. This may cause loss of some important information owing to the small sample size. Involved in particular, the sample of annual maxima obtained may not be representative of the actual tail of the underlying distribution, since some intense events that are neglected only because they are not annual maxima,even though they might easily be more intense that the annual maximum of some other year. To overcome these limitations the main alternative approach used in hydrology is the *Peak Over Threshold method* (Balkema and De Haan, 1974; Pickands, 1975). They showed that if a distribution exists for appropriately linearly rescaled excesses $Y_i = X_i - q$ of a sequance of iid observations $X_i, i = 1, N$ above a threshold q, than their limiting distribution will be a *generalized Pareto distribution* (GPD). In the applications, the model is defined by picking a 'high enough' threshold q and by studying all the exceedances of q. It must be taken into account that the number of exceedances over a given period of time and the excess values are themselves random variables. Usually the latter is described using the *generalized Pareto distribution* (GPD) whereas the arrivals of the threshold exceedances are commonly assumed to follow a Poisson distribution.

Let X be a random variable whose CDF is F and let define the *excesses* over a fixed threshold q as $Y = X - q$ conditioned on $X > u$. Then

$$P(Y \leq y) = P(X \leq x + q \mid X > q) = F_q(y) = \frac{F(q + y) - F(q)}{1 - F(q)} \tag{1.19}$$

Pickand (1975) showed that when the threshold approaches infinity (or the finite upper endpoint of the parent distribution F(x), if this is the case) there will exist a set of parameters ξ and σ_q (the latter depending on the threshold) such that the GPD is a good approximation for the distribution of the excesses:

$$F_q(y) \sim F(y; \xi, \sigma_q) = 1 - \left(1 + \frac{\xi}{\sigma_q} y\right)^{-1/\xi} \tag{1.20}$$

It must be remarked that this result holds for $q \longrightarrow \infty$ so this is still an asymptotic method. With the GPD, like the GEV there are three different cases depending on the sign of ξ:

1. If $\xi > 0$ then (1.20) is defined on $0 < y < \infty$ and the tail of the distribution decays as a power law or *'Pareto tail'*. This case corresponds to the *Frechet* type in the EVT.

2. If $\xi < 0$ then F(y) has an upper endpoint $\omega_F = \sigma_q / \mid \xi \mid$ similar to the Weibull type of the EVT.

3. If $\xi \longrightarrow 0$, recalling the definition of the number e, the GPD become the exponential distribution with mean σ_q in a similar fashion to the Gumbel type in the EVT.

$$F(y; \sigma_q, 0) = 1 - exp\left(-\frac{y}{\sigma_q}\right) \tag{1.21}$$

As with the GEV distribution, the mean exist if $\xi < 1$ and the variance if $\xi < 1/2$ and their expressions are, respectively:

$$E(Y) = \frac{\sigma}{1 - \xi}, \qquad Var(Y) = \frac{\sigma^2}{(1 - \xi)^2 (1 - 2\xi)} \tag{1.22}$$

Fixed a threshold q, the number n of exceedances in one year is assumed to be a random variable itself. The standard choice is to model the random variable N

with a Poisson distribution with $E(N) = Var(N) = \lambda$. thus the probability of occurrence of n exceedances in one year will be given by

$$P(N = n) = f(n; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1.23}$$

The adoption of a Poisson model implies that the number of arrivals in two non overlapping time windows are independent. Therefore, if we assume the number of yearly exceedances to have a Poisson distribution with mean λ and all the exceedances to be independent realizations and GPD distributed, the probability of the annual maximum of the process just described being less than a value x will be

$$\begin{aligned} P(max_{1\leq i \leq N} \leq x) &= P(N = 0) + \sum_{k=1}^{\infty} P(N = n, Y_1 \leq x, ..., Y_n \leq x) \\ &= e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \cdot \left[1 - \left(1 + \xi \frac{x - q}{\sigma}\right)^{-1/\xi}\right]^n \\ &= \exp\left[-\lambda \left(1 + \xi \frac{x - q}{\sigma}\right)^{-1/\xi}\right]. \end{aligned}$$

This result shows that the probability distribution of the annual maxima of a GPD-Poisson process is the same as a GEV with parameters ξ, ψ, μ. Hence the two models are consistent if and only if

$$\xi = \xi \tag{1.24}$$

$$\sigma = \psi + \xi \cdot (q - \mu) \tag{1.25}$$

$$\lambda = \left(1 + \xi \frac{q - \mu}{\psi}\right)^{-1/\xi} \tag{1.26}$$

Thus the shape parameter is the same for GPD and GEV; furthermore we can find scale and location parameter of the GEV as

$$\psi = \sigma \cdot \lambda^{\xi} \tag{1.27}$$

$$\mu = q - \frac{\sigma(1 - \lambda^{\xi})}{\xi} \tag{1.28}$$

The GPD satisfies a threshold stability property i.e. if one fixes an higher threshold $u_2 > u_1$ the subsequent excesses will also follow a GPD with the same shape parameter but shifted scale $\sigma_{u_2} = \sigma_{u_1} + \xi \cdot (u_2 - u_1)$. Hence, dependence of the scale parameter on the threshold disappears only in the limiting case $\xi = 0$, in which GPD reverts to the exponential distribution. The Peak over threshold method allows an estimation of the GEV parameters based on the real tail behaviour of the daily rainfall ditribution whereas the traditional fitting methods, considering only the annual maxima, may determine a distorsion in the tail modeling. The price for this achievement is that the POT is not as straightforward as the Annual Maxima method to apply, but the threshold selection requires particular care. The theoretical assumption on which POT is based would require a 'high' threshold; In practice, the chosen threshold must be high enough for the two hypotesis (Poisson arrivals of exceedances and GPD distribution of the excesses) to be satisfied. The optimal threshold requires a trade off between standard deviation (which increases with higher thresholds due to the smaller number of excesses) and bias, which arise when the threshold is too low and the GPD-Poisson model is too rough a description of the data at hand.

Chapter 2

The Penultimate approximation

2.1 Cramer's Method

We are interested in studying the probability distribution of maxima of some naturally variable quantity across blocks with a fixed length of T years. Let n be the number of realizations of the relative variate in the considered interval; We are thus looking for the distribution of the maximum of n independent samples, each drawn from the same parent distribution $F(x) = P(X \leq x)$. Classical probability theory indicates that, under the above mentioned hypothesis, the cumulative probability function of the maximum will be

$$F_n(x) = [F(x)]^n \tag{2.1}$$

It is useful to introduce here the so-called *characteristic largest value of* X, X_T which is defined as:

$$Q(X_T) = P(X \geq X_T) = 1/n \tag{2.2}$$

So the characteristic largest value of X is the value that it is exceeded with probability $1/n$ (i.e. the mean largest value of a sample of n values extracted from a variate whose cumulative distribution function is $F(x)$). Thus if equation (2.1) is expressed in terms of the survival probability function $P(X \geq x) = Q(x) = 1 - F(x)$ the following expression is obtained:

$$F_n(x) = [1 - Q(x)]^n = \left[1 - \frac{Q(x)}{n \cdot Q(X_T)}\right]^n \tag{2.3}$$

If we consider a value $x > X_T$ for the variate X, then $\frac{Q(x)}{Q(X_T)} \leq 1$ (This being the case whenever we look at return times greter than 1 year) and being in any case $n > 1$ the expression can be simplified using the *Cauchy Approximation*, as many authors refer to the first order expansion of the function (2.3) as a Taylor series:

$$F_n(x) = \left[1 - \frac{Q(x)}{n \cdot Q(X_T)}\right]^n \simeq 1 - n\frac{Q(x)}{n \cdot Q(X_T)} \simeq exp\left[-\frac{Q(x)}{Q(X_T)}\right] \tag{2.4}$$

The error in neglecting the higher order terms depends grows with n and decreases with the distance between x and X_T. In particular if we fix a value of the variable $z = \frac{Q(x)}{Q(X_T)}$, i.e. if we fix the distance from the point at which the series is centered, the error will depend solely on n. This may be demonstrated by evalutating the error associated with the *characteristic largest value* X_T. In this case the result from using the *Cauchy Approximation* is $F_n(X_T) = e^{-1}$ while the exact result depends on the n. The error is given by their difference:

$$err(n) = e^{-1} - \left[1 - \frac{1}{n}\right]^n \tag{2.5}$$

For $x > X_T$ we have $Q(x) < Q(X_T)$ and therefore equation (2.5) overestimates the error that ensue from the *Cauchy approximation* for every $x > X_T$. The error tends to zero rather quickly as $n \to \infty$; for example for $n = 50$ the corresponding relative error is $err(X_T, n) = 0.01$. The*Cauchy approximation* is the only asymptotic step necessary in the *Cramer's method* because in the case $n \longrightarrow \infty$ the error is zero whatever the value of z may be. The linearized expression in eq. (2.4) it is often referred to as '*Penultimate approximation*'.

2.2 The asymptotic distribution

Cramer derived the *Gumbel* or EVT-TYPE I distribution using the *Penultimate approximation* in the case of exponential type parent. A parent distribution is said to be of the exponential type if it can be written in the form

$$F(x) = 1 - e^{-g(x)} \tag{2.6}$$

where $g(x)$ is positive and increases monotonically faster than $\log(x)$. From the latter form of the *Cauchy approximation* in the case of a parent distribution of the exponential type we obtain

$$F(x) \simeq e^{-\frac{e^{-g(x)}}{e^{-g(X_T)}}} \tag{2.7}$$

and

$$y_T = -ln(-ln(F_n(x)) = g(x) - g(X_T) \tag{2.8}$$

This is the *Gumbel reduced variate.* If we expand equation (2.8) as a Taylor series around the *characteristic largest value* X_T we obtain

$$y_T = g'(X_T) \cdot (x - X_T) + \frac{g''(X_T)}{2!} \cdot (x - X_T)^2 + \frac{g'''(X_T)}{3!} \cdot (x - X_T)^3 + ... \tag{2.9}$$

This is still the penultimate distribution for a parent variable with an exponential distribution; In fact we do need to know the function $g(x)$ in order to evalutate it and the only error in it is the one that ensue from the application of the *Cauchy Approximation.* The ultimate asymptotic form of the *Gumbel* or EVT-TYPE I is obtained from equation (2.9) by dropping the second and higher terms of the Taylor expansion:

$$y_T = g'(X_T) \cdot (x - X_T) \tag{2.10}$$

Thus, if we rename $\mu = X_T$ and $\alpha = g'(X_T)$ then we note that what we have just obtained is nothing else but the usual expression of the Gumbel cumulative distribution function:

$$F(x) = e^{-e^{-\alpha \cdot (x-\mu)}} \tag{2.11}$$

We can see that, being this an asymptotic distribution, there is no more dependence on n; This is due to the fact that it applies in the limit as $n \longrightarrow \infty$ and neither it depends on the underlying distribution i.e. on the particular function $g(x)$. The expression is obtained as a linearization of the more general penultimate distribution and the errors introduced by discarding the second and higher terms are in general very much more significant than that from the earlier *Cauchy Approximation.*

2.3 Penultimate approximation in the case of a Weibull variate

The penultimate approximation and the corresponding error are now derived in the important case of a Weibull parent distribution, in which:

$$F(x) = P(X \leq x) = 1 - e^{-\left(\frac{x}{C}\right)^w} \tag{2.12}$$

This is a distribution of the exponential type with $h(x) = (x/C)^w$ so that $h'(x) = wx^{w-1}/C^w$ and $h''(x) = w(w-1)x^{w-2}/C^w$. The error implied by the use of the asymptotic distribution depends on the degree of convergence to the ultimate asymptote i.e. on the value of n. This error arises from the terms of order higher than one that are neglected in eq. (2.9) in order to obtain the EV1 distribution. In the case of integer values of the Weibull shape parameters w, all the terms after the w-th are zero, whereas in the more general case of non integer values of w the series has infinite terms. In both cases we can obtain a first order estimate of the error from the value of the first term neglected:

$$\frac{\epsilon(y_T)}{y_T} = \frac{h''(x_T)(x - x_T)^2}{2h'(x_T)(x - x_T)} \tag{2.13}$$

And in the case of a Weibull variate

$$\frac{\epsilon(y_T)}{y_T} = \frac{w-1}{2}\left(\frac{x}{x_T} - 1\right) \tag{2.14}$$

For $w = 2$, that is the case of the Rayleigh distribution, the error given by (2.14) is exact, since the value of all the neglected terms is zero. In the case of $w = 1$, the exponential distribution, the error is zero. This result points out that in the special case of exponential parent, the ultimate asymptotic form EV1 is the exact penultimate distribution. Hence, in this case the only error is the small error associated with the *Cauchy Approximation*. This means that the convergence to the Gumbel asymptote for the cumulative distribution of maxima extracted from a parent exponential variate is very fast. On the other hand, in the general case $w \neq 1$ the asymptotic distribution will be affected both from the error due to the *Cauchy Approximation* and from the one associated with the neglected terms in eq. (2.9). In this case the convergence speed depends on the rate with which the nonlinear terms in the taylor expansion tend to zero as $n \to \infty$. Hence, for certain values of the shape parameter w the second term of the error might be much greater than the *Cauchy*'s one and thus the convergence might be very slow.

2.4 Preconditioning for a faster convergence

As we have seen in the previous section, despite the fact that a Weibull parent belongs to the domain of attraction of the Gumbel-FT1 limiting distribution, its convergence to the asymptote can be very slow depending on the value of the shape parameter. In general for any finite n and for any fixed value $w \neq 1$ the Gumbel distribution is not the correct cumulative density function of the annual maximum. Some authors (e.g. Kudsoyiannis, 2012) proposed that in this case rainfall extremes should be fitted to the Frechet distribution with a fixed positive shape parameter rather than to the Gumbel. In general the adoption of the GEV will guarantee a better fit thanks to the additional degree of freedom. However, in the general case of Weibull variate a variable trasformation can be applied in order to obtain faster convergence to the Gumbel asymptote. We have remarked that in the unique case of exponential distribution only the first component of the error arise, producing an extremely fast convergence. We note that any variate X with a Weibull distribution can be trasformed into a new exponentially distributed variate Z by performing a simple change of variable

$$Z = X^w \tag{2.15}$$

Hence, for the new variate Z the general penultimate distribution corresponds to the asymptotic Gumbel distribution; its penultimate form can be written as follows

$$F_n(z) \simeq exp\left(-n \cdot exp\left(-\frac{z}{C^w}\right)\right) = exp\left(-exp\left(-\frac{z}{C^w} + \log n\right)\right) \tag{2.16}$$

This *Penultimate approximation* is the exact cumulative distribution function of the annual maxima, if we neglect the sole error due to the Cauchy's approximation. We point out that, because of the variable change, this expression corresponds to the Gumbel, whose scale and location parameter can be determined as follows

$$\alpha\,(x - \mu) = \frac{1}{C^w}\,(z - C^w \cdot \log n) \tag{2.17}$$

And therefore

$$\alpha = \frac{1}{C^w}$$
$$\mu = C^w \cdot \log n$$

This approach allows the Gumbel parameters to be estimated directly from the parameters of the Weibull parent and from the knowledge of the yearly number of independent realizations. This result holds also for variates whose distribution function is right-tail equivalent to a Weibull distribution. Some authors (e.g. Cook and Harris, 2004) refer to the practice of transorming the initial variable to get complete convergence to the asymptotic form as '*Preconditioning*'.

Chapter 3

The MEV distribution

3.1 Daily rainfall distributions and tails

The Cramer method and the attraction theorems suggest that the parent distribution chosen for the daily rainfall values plays a key role in determining the shape of the extreme rainfall distribution. Effectively, the right tail of the distribution governs the survival probability of events with a 'sufficiently high' magnitude. There is no current knowledge of a general universal law for modeling the distribution of daily rainfall values. The probability discribution of rainfall, at daily or even smaller timescales, belong to the family of mixed type distributions, with a discrete part describing the probability of rainfall events arrivals, and a continuous part expressing the magnitude of the rain events; The latter represents the distribution of rainfall amounts during wet days. The traditional and easiest way to treat the problem is to assume the rainy days arrivals to have a Poisson distribution and to describe the rainfall intensities with some distribution which belongs the exponential family, such as e.g. the Gamma or the Exponential distribution. This simple approach has been proven to fail in severalcases: often the tail of the distribution is subexpontial (heavier than an exponential one) and the adoption of a light tail distribution such as the Gamma would lead to underestimation of the probability of high magnitude events. A second deficiency lies in the hypotized Poissonian nature of the wet days arrivals: Rainfall events tends to occur to cluster, and therefore the Poisson distribution should only be used to model the arrivals of exceedances over a relatively 'high' thresholds (Leadbetter, 1983). The higher the threshold, the closer the Poisson model to the empirical data will be. Moreover, when we consider the whole range of daily rainfall values, only a small fraction of them belong to the tail and in general is the tail where the fitting error have the greatest relevance: Effectively the fitting procedure estimates the set of parameters which best describe the largest portion of the data. Such a procesdure may lead to serious errors if the aim of the study is extreme event modeling. Papalexiou et al (2013) compared the upper part of the empirical distributions from several stations with four common theoretical tails: those of Generalized Pareto, Lognormal Weibull and Gamma, considering the 'tail' composed by the N largest values in a series of N years. They found Generalized Pareto (with a shape parameter mode of 0.134) and Lognormal to better fit the tails, followed by the Weibull distribution. From this analysis it is clear that heavier tailed distribution in general are to be adopted in right tail modeling (in 72.6% of the records subexponential tails performed better than exponential or hyperexponential).

3.2 A physical justification for the adoption of Weibull parent distribution

Several statistical distributions are commonly used in the practice to approximate daily precipitation totals, such as for example the Exponential, Gamma and Generalized Pareto distributions. In most of the cases the choice of the distribution is unclear and lacks of a physical justification. Wilson and Tuomi (2005), interpreting the water balance equation, were able to find an expression for the daily precipitation probability distribution as a product of mass flux, specific humidity and precipitation efficency. Precipitation can be expressed as the moisture flux integrated over the air column and, using a two-layers model to describe the atmosphere, the of the precipitation rate can be expressed as follows:

$$R \simeq -\int_0^{z_m} \overline{\nabla \cdot (q\rho\vec{v})} dz = \int_0^{z_m} \overline{\frac{\partial q\rho\vec{w}}{\partial z}} dz = \overline{(q\rho\vec{w})}_{z_m} \tag{3.1}$$

Where R is the precipitation rate, ρ the air density and q the specific humidity or mass mixing ratio. The horizontal velocity and upward vertical velocity are, respectively, $\vec{v}$ and $\vec{w}$; $\overline{\rho\vec{w}}$ is the mean upward mass flux, the over-bar representing a temporal average. The moisture flux is integrated between $z = 0$, ground level, and $z = z_m$, that represents the moist level. In a more general expression, if we include in the model upper level divergence and increases in moisture storage, we can write the actual precipitation rate as

$$R = \overline{k\left(q\rho\vec{w}\right)}_{z_m} \tag{3.2}$$

In this expression k is the istantaneous precipitation efficency and represents the fraction of the vertical moisture flux at z_m which is precipitated out. Previous works have shown that mass flux, precipitation efficency and specific humidity can be assumed with good approximation to be three independent variables. Therefore the accumulated precipitation total R_{acc} for a given storm can be modeled with the triple product $R_{acc} = \bar{k}\bar{q}m$, where m is the mass of air into the column that is advected and pushed through the moist level. The empirical characterization of the three aforementioned distributions is not feasible in practice, since available observations are not available that allow to carry out the time average. A different approach can be pursued if one assumes sufficently light tails and a sufficient averaging period. In this case the distributions of the three variables can be assumed by the Central Limit Theorem to be Gaussian; Effectively, the longer the period in which the temporal average is performed, the closer the three distributions will be to Gaussian bells. By rescaling the three variables, we can express precipitation as proportional to the product of three unit normal variables. Frisch and Sornette (1997) have shown that the probability density function of the product of a finite number of independent random variables is approximately of the stretched exponential form in the upper right tail of the distribution. Wilson and Tuomi (2005) showed that for large enough values, the three terms in the product are of the same order. Therefore the probability of precipitation is the joint probability of three random variables all having common value $R^{1/3}$; Moreover the joint pdf can be written as the product of the three marginal pdf because of their independence, such that

$$P(R) \sim \left[P(n = R^{1/3})\right] \tag{3.3}$$

where $P(n = R^{1/3}) \propto \exp(-\left(R^{1/3}\right)^2)$. Integrating the resultant probability density we can obtain the cumulative distribution function of heavy precipitation

$$F(x) = P(X \leq x) = 1 - e^{-\left(\frac{x}{C}\right)^w} \tag{3.4}$$

This is the expression of the Weibull cumulative distribution with shape parameter $w = 2/3$ and scale parameter $C \in R$. For $w < 1$ this equation is known as stretched exponential distribution and it is slightly heavy tailed. Wilson and Tuomi also explored the global variation in the stretched exponential shape parameter by fitting the Weibull distribution to the daily precipitation values greater than 10 mm from several station from the NCDC data center. They found the annual global mean shape parameter to be 0.66 with a standard deviation of 0.16 (a value consistent with the theoretical framework adopted) and to be slightly dependent on climate change effects. The byprocuct of the last finding is that changes in the rainfall generating process are more likely to produce changes in the scale parameter of the distribution rather than in the shape of the tail.

3.3 A fully probabilistic approach

If we remove the asymptotic assumption, the exact expression of the cumulative probability for the n-sample (yearly) maximum M_n is

$$P(M_n \leq x) = H_n(x; \vec{\theta}, n) = F^n(x; \vec{\theta}) \tag{3.5}$$

where n is the number of rainy days per year i.e. the number of realization of the rv X. Thus the cumulative distribution of the annual maximum will be depending on n and on the vector $\vec{\theta}$ of parameters of the parent distribution. Since the parameter estimators of the parent distribution depends on the data set, the parameters obtained as an outcome of the process are to be considered random variables themselves. At the same way the number n of yearly rainy days is the realization of a discrete random variable N. Therefore a general definition of the yearly maximum cumulative distribution should take this higher-order randomness into account. This can be done considering the expected value of $H_n(x; \vec{\theta}, n)$ computed over all the possible realizations of n and $\vec{\theta}$:

$$\zeta(x) = \sum_{n=1}^{\infty} \int_{\vec{\theta}} g(n, \vec{\theta}) \cdot H_n(x; n, \vec{\theta}) \cdot d\vec{\theta} \tag{3.6}$$

where $g(n, \vec{\theta})$ is the joint probability distribution function of the random variables $\{N, \Theta_1, ..., \Theta_n\}$ and $d\vec{\theta}$ denotes the differential $d\theta_1 \cdot d\theta_2 \cdot ... \cdot d\theta_n$ This expression has recently been proposed as '*Metastatistic Extreme Value distribution*' (MEV) (Marani and Ignaccolo, 2014). The general formulation (3.6) is appealing, but in practice it might be arduous to identify an analitical expression for the distributions $g(n, C, w)$ or $h(C, w)$. In the absence of such analytical distributions one can use the empirical distributions of the parameters. This procedure is known as *Monte Carlo integration* and is based on the approximation of the probability weighted integral of a given function of a random variable $f(\vec{\theta})$ based on a given target pdf $p(\vec{\theta})$:

$$\int p(\vec{\theta})f(\vec{\theta})d\vec{\theta} \simeq \frac{1}{T}\sum_{j=1}^{M} f(\vec{\theta}_j) \tag{3.7}$$

The integral can be approximate through an algebric summation, provided that the $\vec{\theta}_j$ for $j = 1, 2, ...M$ are randomly sampled from their original target distribution $p(\vec{\theta})$. The precision of such an approximation increases with the number of elements sampled M. In the case of independent samples equation (3.7) is a consequence of the strong law of large numbers, but the result holds in general. To proof the result it should be recognized that the number of times that a value $f(\vec{\theta}_*)$ appears in the summation is nearly $Mp(\vec{\theta}_*)d\vec{\theta}_*$, and the higher M, the closer the real value to this quantity. This holds for any value of $\vec{\theta}_* \in \Omega_{\vec{\theta}}$ and therefore if we assume (without any loss of generality) that the set of extracted values of $\vec{\theta}$ can be treated as countable, we can express the right hand side of (3.7) as:

$$\frac{1}{T}\sum_{j=1}^{M} f(\vec{\theta}_j) \simeq \sum_{\vec{\theta}_*=inf\left(\Omega_{\vec{\theta}}\right)}^{sup\left(\Omega_{\vec{\theta}}\right)} f(\vec{\theta}_*)p(\vec{\theta}_*)d\vec{\theta}_* \tag{3.8}$$

Which is indeed a discrete approximation of the original integral given in eq. (3.7). Therefore, in the case in which $g(n, \vec{\theta})$ is not known a priori it is still possible an estimation of the (3.6). For example, given a window of daily rainfall record of T years, we can write $\zeta(x)$ as an average of $H_n(x)$ over all the yearly realizations of n and $\vec{\theta}$:

$$\zeta(x) = \frac{1}{T}\sum_{j=1}^{T} H_{n_j}(x; n_j, \vec{\theta}_j) \tag{3.9}$$

This expression thereafter will be referred to as *MEV Complete* and may be also formally obtained from eq. (3.6) by considering the joint probability density of n

and $\vec{\theta}$ as a bivariate Dirac Delta centered in n_j and $\vec{\theta_j}$, i.e. in the actual realizations of the random variables at the j-th year:

$$g(n,\vec{\theta}) = \frac{1}{T}\sum_{j=1}^{T}\delta(n-n_j,\vec{\theta}-\vec{\theta_j}). \tag{3.10}$$

In this way it is possible, after the selection of a proper distribution for describing the daily rainfall amounts, obtain the expression of the annual maxima CDF. The *Monte Carlo Integration* allows the empirical distribution to be used instead of a parametric model for $\vec{\theta}$ and N. For some distribution (e.g. GPD) a single year is in general too small a sample to obtain a good estimated parameter set; In particular the mean and variance of estimated shape parameter depends on the sample length (Serinaldi and Kilsby, 2014). Therefore the sample size for fitting the parent distribution shoud be carefully selected depending on the particular adopted parent analytical function.

3.4 MEV in the case of a Weibull variate

Eq. (3.6) can be made explicit by choosing a parent distribution to describe daily rainfall. As we have showed in the previous chapter, the stretched exponential distribution is the fundamental heavy daily rainfall distribution. We now present the formulation of the Metastatistic extreme value distribution in the important case in which the daily rainfall is Weibull distributed. The general expression of the MEV in this case will be

$$\zeta(x) = \sum_{n}\int_{C}\int_{w} g(n,C,w)\cdot\left[1-e^{-\left(\frac{x}{C}\right)^{w}}\right]^{n} dCdw \tag{3.11}$$

If n is independent on the scale and shape parameters of the daily rainfall pdf, their joint distribution can be expressed as $g(n,C,w) = f(n)\cdot h(C,w)$. Furthermore, if we apply the Cauchy approximation to eq. (3.11) the above integral expression become

$$\zeta(x) = 1-\sum_{n} n\cdot f(n)\cdot\int_{C}\int_{w} h(C,w)\cdot\left[e^{-\left(\frac{x}{C}\right)^{w}}\right] dCdw \tag{3.12}$$

In this expression the higher order randomness in the variable N is still taken into account, but since the *Cauchy Approximation* has been applied, only its mean value appears in the MEV, and therefore there is no need to know the exact distribution of the yearly number of events. A simpler epression can be obtained if one considers C and w fixed constants rather than random variables (which is equivalent to estimate these parameters for the entire period of record). Under those assumptions the formulation of the MEV is the same as the *Penultimate approximation*:

$$\zeta(x) = 1 - \bar{n} \cdot \ e^{\left(-\frac{x}{C}\right)^{w}} \tag{3.13}$$

Therefore it can be applied to fit a sample of daily data either using this expression or by preconditioning, as previously showed. Prior information on the distributions of the parameters of the daily rainfall pdf is missing and it is not convenient to use analitical expressions for the parent distribution. This shortcoming can be avoided by when the expression of the *Mev complete* derived in the previous section, suitably particularized to the case of Weibull variate, is used:

$$\zeta(x) = \frac{1}{T}\sum_{j=1}^{T}\left[1 - e^{\left(-\frac{x}{C_j}\right)^{w_j}}\right]^{n_j} \tag{3.14}$$

When N and $\vec{\theta} = [C, w]$ are independent, the *Cauchy Approximation* can be applied and the average number of wet days is taken out of the summation leading to the following simplified expression:

$$\zeta(x) = \frac{1}{T}\sum_{j=1}^{T}\left[1 - n_j \cdot e^{\left(-\frac{x}{C_j}\right)^{w_j}}\right] = \tag{3.15}$$

$$= 1 - \bar{n}\frac{1}{T}\sum_{j=1}^{T}\left[e^{\left(-\frac{x}{C_j}\right)^{w_j}}\right] \tag{3.16}$$

Despite eq. (3.16) being simpler (3.14), it still needs to be solved numerically. Therefore in general its use is not recommended beacuse of the error introduced by the application of the Cauchy approximation.

3.5 MEV in the case of a Generalized Pareto variate

The Peak Over Threshold approach described in chapter 1 can be thought of as a particular case of Metastatistic Extreme Value distribution;In fact, in order to obtain the POT estimation of the GEV parameters, the expression of the annual maximum cumulative distribution has to be integrated over all the possible values of the cardinality N (where N is assumed to follow Poisson distrbution). On the other hand, the randomness of the parameters of the Generalized Pareto distribution is not taken into account by the traditional POT method. In this section a more general expression of the POT method is presented, which does not rely on the Poisson hypothesis for the distribution of the random variable N and does include the inter-annual variability of the GPD parameters. As a first step a threshold u is set and the excesses of u are defined as the difference $Y = X - u$ where X identifies the daily rain dephts, which are assumed to be i.i.d. random variables. Hence, we can write the probability of a daily realization being smaller than a given value x as

$$P(X < x) = P(Y \leq y|X > u) + P(X \leq u) \tag{3.17}$$

The distribution of the excesses over the threshold is modeled with a GPD distribution: $F(y) = 1 - \left(1 + \frac{\xi}{\psi} \cdot y\right)^{-1/\xi}$. Moreover the daily rainfall distribution below the threshold is modeled with the non exceedance frequencies of the observed data, such that $F(u) = \frac{n-k}{n}$ where k is the number of yearly exceedances of the thresholds and n is the yearly number of events. With these assumptions, the cumulative distribution of the daily amounts become

$$P(X < x) = \left\{1 - \left(1 + \frac{\xi}{\psi} \cdot (x - u)\right)^{-1/\xi}\right\} \cdot \frac{k}{n} + \frac{n-k}{n} \tag{3.18}$$

Furthermore, we can derive the cumulative distribution of the annual maximum for a given value of the threshold u and of the parameter set $\{n, k, \xi, \psi\}$:

$$F_n(x) = \left\{1 - \frac{k}{n}\left(1 + \frac{\xi}{\psi} \cdot (x - u)\right)^{-1/\xi}\right\}^n \tag{3.19}$$

The general formulation of the MEV distribution for equation (3.19) can be obtained as the expected value of the annual maximum cumulative distribution over all the possible values of the parameters $\{n, k, \xi, \psi\}$ considered as random variables.

$$\zeta(x) = \sum_n \sum_k \int_\xi \int_\psi g(n, k, \xi, \psi) \cdot \left[1 - \frac{k}{n}\left(1 + \frac{\xi}{\psi} \cdot (x - u)\right)^{-1/\xi}\right]^n d\xi d\psi \quad (3.20)$$

In analogy with the Weibull case, if n is independent of the scale and shape parameters we can apply the Cauchy approximation in order to pull the variable k out of the integral operator, so that the MEV depends only on the expected value of K and not on its actual distribution.

$$\zeta(x) = 1 - \sum_k k \cdot p(k) \cdot \int_\xi \int_\psi h(\xi, \psi) \cdot \left(1 + \frac{\xi}{\psi} \cdot x\right)^{-1/\xi} d\xi d\psi \quad (3.21)$$

A simpler epression can be obtained by considering scale and shape parameters fixed constants rather than random variables, which leads to the following expression:

$$\zeta(x) = 1 - \bar{k} \cdot \left(1 + \frac{\xi}{\psi} \cdot (x - u)\right)^{-1/\xi} \quad (3.22)$$

Where $\bar{k}$ is the mean value of the yearly number of exceedances over the threshold, averaged over all the years of observation. As a consequence of the lack of a general analytical expression for the parameter distribution, in general the best way to proceed is to obtain a discrete approximation of eq. (3.21) via Monte Carlo integration. The general expression of the MEV is thus approximated using the empirical distributions of the parameters obtained by fitting a Generalized Pareto to the yearly samples of a T-years period of record:

$$\zeta(x) = \frac{1}{T} \sum_{j=1}^{T} \left[1 - \frac{k_j}{n_j}\left(1 + \frac{\xi_j}{\psi_j} \cdot (x_j - u)\right)^{-1/\xi_j}\right]^{n_j} \quad (3.23)$$

Which is used for practical applications.

3.6 Previous Montecarlo experiments

Marani and Ignaccolo (2014) performed Montecarlo experiments based on the Padova data set in order to show that the MEV distribution estimates the actual probability of extreme events. They generated synthetic data sets drawing daily rainfall values from a parent Weibull distribution with given values of the shape and scale parameters and then compared the MEV approach, eq. (3.6) particolarized for a Weibull parent distribution, with the traditional GEV and Gumbel distributions. The artificial data sets were constructed for different values of the cardinality N and of the Weibull parameters set $\vec{\theta} = [C, w]$ in order to obtain time series with the characteristic of the observed data and for which the distribution of the annual maxima is known by construction. They found that the GEV estimates consistently underestimate the reference distribution, whereas the adoption of the Gumbel asymptote leads to an overestimation of the true value. As a consequence, for a given return time (i.e. of non exceedance probability) the adoption of GEV and Gumbel models lead to respectively to an overestimation and underestimation of the associated precipitation value. They found good agreement between the MEV estimated of the recurrence of extreme events and their 'true' probabiliy of occurrence, known by construction of the synthetic data sets. They analyzed also the performances of GEV and MEV in the case of Non-stationarity (arificially represented by generating data sets with different frequencies of wet days occurrence) and showed that in this case GEV and Gumbel produce a bias in quantile estimation, owing to the fact that they are more likely to fail when applied to non stationary time series. They argue that instead eq. (3.6) should be used in the case of statistically inhomogeneous periods.

Chapter 4

Case studies and data analysis

4.1 The datasets

The majority of the daily rainfall records used in this study were obtained from the NOAA's National Climatic Data Center (NCDC), located in Asheville, North Carolina, which maintains the world's largest climate data archive. The data used are part of the GHCN (Global Historical Climatology Network). Its daily documentation includes daily rainfall total, snowfall, snow depth, maximum and minimum daily temperature, evaporation and more. The only exception is the Padova time series, the older records of which are still conserved at the Museo dell'Osservatorio Astronomico di Padova. The GHCN archive includes thousands of record worlwide but many datasets are affected from non negligible percentages of missing values. The dataset selection for this study was based on two main aspects: Completeness and record length. The first characteristic is essential because the MEV approach requires a good description of the daily rainfall distribution function; Therefore the data were selected among the GHCN station whose records have a percentage of missing values per time series less than 2% in the analyzed timespans; In some of the longest series (Milano and Padova) a few years had to be removed from the series becuse they are characterized by a non negligible number of missing values. The second fundamental aspect we focused our selection on was the record length; In particular, since the stationary analysis requires an estimate of the empirical magnitude of the event characterized by a given return time, we needed records as long as possible in order to bestow significance to the statistical analysis. Fur-

Dataset	Years of obs.	Elev. (m)	Coverage	Missing y.
Asheville (NC)	1903-2006	627.9	100%	-
Heerde (NL)	1893-2013	6	100%	-
Hoofdoorp (NL)	1867-2014	-3	100%	-
Kingston (RI)	1897-2013	75	100%	-
Livermore (CA)	1903-2014	146.3	99%	-
Milano (IT)	1858-2007	150	100%	2
Padova (IT)	1725-2013	275	100%	14
Philadelphia (PA)	1901-2006	21	100%	-
Putten (NL)	1868-2013	14	100%	-
Roosvelt (AZ)	1906-2014	672.1	98%	-
San Bernardo (FR)	1901-2006	2472	98%	-
Zurich (CH)	1901-2014	556	100%	-
Bologna (IT)	1814-2003	53	100%	-
Worcester (SA)	1880-1998	270	91%	-
Albany (GA)	1901-2014	54.9	99%	-

Table 4.1: Stations selected for the study

thermore in the perspective of assessing the effects of climate change on rainfall extremes, only a long enough time series allows trend and non stationarity detections. For this reason most of the selected datasets have more than 100 years of observations and in some cases the number is closer to 200 years. The Padova dataset in particular is, to our current knoledge, the longest existing time series of daily rainfall observations for a total of 275 complete years of daily records. It spans from 1725 to 2013, of which only a few years were removed from the datased owing to missing data.

4.2 Autocorrelation of the daily rainfall

In all the extreme value models described in chapter 1,2 and 3 the basic assumption is that the daily values are iid random variables. Even if one uses a block maxima approach (annual maxima can be reasonably assumed to be independent), the GEV distribution is obtained as limiting distribution for the maximum of n i.i.d. random variables. Often, when we are considering daily rainfall amounts, this hipothesis may not be thereby complied. On the contrary, we observe that wet days tends to occur in clusters, determining a often a non negligible correlation between the precipitated amounts of consecutive days. To evalutate the intensity of this correlation for a given temporal lag, it is useful to plot the autocorrelation function. Any time series of daily precipitated amounts of rainfall can be thought of as the realization of a stochastic process; In this framework it is useful to define the autocovariance of the process, that is the covariance between the random variable $X(t)$ and $X(t+\tau)$ for a given lag τ:

$$\gamma(\tau) = E\left[(X(t)-\mu)\cdot((X(t+\tau)-\mu)\right] \tag{4.1}$$

We point out that for $\tau = 0$ the above expression represents the variance of the random variable $X(t)$. We computed the empirical autocorrelation of lag r days as follows

$$r_k = \frac{\sum_{i=1}^{n-k}(x_i-\bar{x_i})\cdot(x_{i+k}-\bar{x}_{i+k})}{\sum_{i=1}^{n}(x_i-\bar{x_i})^2}\cdot\frac{n}{n-k} \tag{4.2}$$

Note that in the summations only the non-zero rainfall values need be considered, since the zeros are perfectly correlated with each others.

4.3 Heavy-tailed and light-tailed distributions

A non negative random variable X (or its distribution) is said to be heavy-tailed if

$$\lim_{x\to\infty}\frac{P(X\geq x)}{e^{-\psi x}} = \infty \qquad \text{for all} \quad \psi > 0 \tag{4.3}$$

The above condition states that the tail distribution function of X is asymptotically heavier than that of any exponential distribution i.e. such distribution decays to zero slower than any exponential distribution as $x \to \infty$. An equivalent definition can be expressed using the moment generating function: in this case, we say that a non negative random variable X (or its distribution) is heavy tailed if

$$E[e^{sx}] = \infty \qquad \text{for all} \quad s > 0 \tag{4.4}$$

Intuitively, heavy-tailed distributions take extremely large values with a non negligible probability. Some of the distributions we use to model daily rainfall are heavy tailed. The Pareto distributions for example belong to this family. The Weibull distribution is also heavy tailed if the shape parameter $w \in (0, 1)$, and in this case we refers to it as stretched exponential distribution. The Weibull distribution is also defined for $w \geq 1$ but it is light tailed over this range. We now define three important subclasses of heavy tailed distributions.

1. A non-negative random variable X (or its distribution function) is said to be *long-tailed* if

$$\lim_{x\to\infty} \frac{P(X \geq x+y)}{P(X \geq x)} = 1 \qquad \text{for all} \quad x > 0 \tag{4.5}$$

   We recall that $P(X \geq x+y|X > x) \cdot P(X \geq x) = P(X \geq x+y)$. Therefore the above definition states that for any fixed $y > 0$ and large x, if a long tailed random variable X exceeds x then it also exceeds $x + y$ with high probability.

2. A non-negative random variable X (or its distribution function) is said to be *subexponential* if

$$\lim_{x\to\infty} \frac{P(max\,\{X_1, X_2\} > x)}{P(X_1 + X_2 > x)} = 1 \qquad \text{for all} \quad x > 0 \tag{4.6}$$

   where X_1 and X_2 are independent random variables distributed as X. This definition may be interpreted by noting that the quantity in the limit equals $P(max\,\{X_1, X_2\} > x|P(X_1 + X_2 > x)$. Therefore, informally, the above definition states that the sum of X_1 and X_2 is large most likely because one of the X_is is large. This is the most interesting case since the distributions we use to model daily rainfall totals such as Generalized Pareto and heavy tailed Weibull belong to this class.

3. The third class of heavy tailed distributions is the class of *regularly varying* distributions. A non negative random variable (or its distribution function) is said to be regularly varying with index $\alpha > 0$ if

$$P(X > x) = x^{-\alpha} \cdot L(x) \tag{4.7}$$

where $L(x)$ is a slowly varying function i.e. it satisfies $\lim_{x\to\infty} \frac{L(xy)}{L(x)} = 1 \quad \forall y > 0$. This condition implies that the tail distribution function of a regularly varying distribution decays asymptotically as a power law; the smaller the index α, the heavier the tail of the distribution. This class is a generalization of the Pareto family and it is strictly contained in the class of subexponential distributions.

Similarly, a non-negative random variable X is said to be light-tailed if it is not heavy-tailed, i.e. if there exist $\psi > 0$ such that

$$P(X \geq x) < e^{-\psi x} \qquad \text{for large enough} \quad x \tag{4.8}$$

The above condition states that the tail distribution function of X is asymptotically bounded above by that of an exponential distribution. In other words, the tail distribution function decays to zero exponentially or faster. Equivalently, a non-negative random variable X is light-tailed if there exists $s > 0$ such that $E[e^{sx}] < \infty$. Before proceeding with extreme value modeling, it is important to analyze the selected datasets in order to evalutate the tail behaviour of the empirical daily rainfall distributions. In the following sections we will define and apply statistical tools (Mean Excess Function and Hill plot) to explore the tail behaviour of the empirical data.

4.4 Mean Excess Function

The *Mean Excess function* is a graphical tool that allows a first analysis of the tail behaviour of a sample (or of a distribution). Let X be a random variable with right endpoint x_F; then the mean excess function of X over the threshold u is defined as

$$e(u) = E(X - u | X > u), \qquad 0 \leq u \leq x_F \tag{4.9}$$

If X is $\exp(\lambda)$ distributed, then $e(u) = 1/\lambda$ does not depend on the threshold. Let assume that X is a random variable with support unbounded to the right and distribution function F. If for all $y \in R$

$$\lim_{x\to\infty} \frac{\bar{F}(x-y)}{\bar{F}(x)} = e^{\gamma y} \tag{4.10}$$

for some $\gamma \in [0, \infty]$, then $\lim_{u \to \infty} e(u) = 1/\gamma$. For the class of the subxponential distributions eq. (4.10) is satisfied with $\gamma = 0$. So that for this class of heavy tailed distributions, encompassing both heavy tailed Weibull and GPD, the mean excess function diverges: $e(u) \to \infty$ as $u \to \infty$. On the contrary for superexponential functions of the type $F(x) \sim exp(-x^{-\alpha})$ with $\alpha > 1$ satisfy eq. (4.10) with $\gamma = \infty$ so that the mean excess function tends to 0 as $u \to \infty$.

In the case of the Generalized Pareto an interesting result holds under the condition $\xi < 1$ i.e. in the case in which the expected value of the distribution is defined: $E[X] < \infty$. Under this hypotesis the mean excess function is linear in u:

$$e(u) = \frac{\psi}{1-\xi} + \frac{\xi}{1-\xi} \cdot u \tag{4.11}$$

where $0 \leq u < \infty$ if $0 \leq x < 1$ and $0 \leq u \leq -\psi/\xi$ if $\xi < 0$. So the GPD distribution is characterized by a linear mean excess function with slope $\xi/(1-\xi)$. Moreover the Pickands-Balkema-de Haan Theorem provides the justification for the peak over threshold method by showing that for a large class of distributions the mean excess function is asymptotically equivalent to a GPD law as the threshold u approaches the right end point of the distribution.

In the case of a Weibull distribution Cook and Harris (2004) showed that the slope of the mean excess function depends on u/C, the ratio of the threshold to the scale parameter of the Weibull distribution through the expression

$$\frac{\xi}{1-\xi} = -\frac{\partial e}{\partial u} = \frac{(1-1/w)}{(u/C)^w} \int_0^1 \left[1 - \frac{log(\xi)}{(u/C)^w}\right]^{\frac{1}{w}-2} d\xi \tag{4.12}$$

In the case of the exponential distribution ($w = 1$), $\frac{\partial e}{\partial u} = 0$ for all threshold and $\xi = 0$: also the GPD reverts to the exponential case. In the case ($w \neq 1$), $\frac{\partial e}{\partial u}$ is a function of the threshold u such that the estimate of the shape parameter of the GPD decreases to asymptotically to zero as the threshold goes to ∞.

Given a independent and identically distributed random sample $x_1, ...x_n$, it is possible to compute the empirical mean excess function $\hat{e}(u)$ to estimate the natural one. It is defined as

$$\hat{e}(u) = \frac{\sum_{i=1}^{n}(x_i - u) \cdot I_{x_i \geq u}}{\sum_{i=1}^{n} I_{x_i \geq u}} \tag{4.13}$$

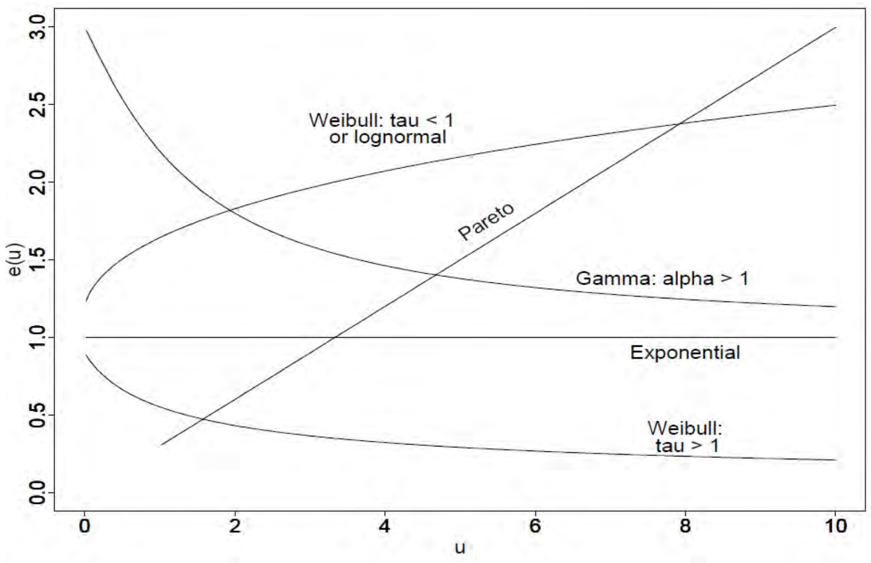


Figure 4.1: Mean excess function for some common distribution [Figure from Embrechts et al., 1997]

Yang suggested the use of the empirical mean excess function and established the uniform strong consistency of $\hat{e}(u)$ over compact u sets, that is, for every$b > 0$:

$$P\left[\lim_{n\to\infty} \sup_{0\leq u\leq b} |\hat{e}(u) - e(u)| = 0\right] = 1 \tag{4.14}$$

4.5 Hill estimator for the shape parameter

Let restrict our analysis to the intersting case in which $X_1, ...X_n$ are iid random variables with common distribution F which belong to the domain of attraction of the Frechet distribution. This is the case whenever $F(x) = x^{-\alpha}L(x)$ for a slowly varying function L. For this class of heavy tailed distributions, the knowledge of the exponent α is of major importance. The Hill estimator of the shape parameter $\xi = 1\alpha > 0$ can be defined and takes on the following form:

$$\hat{\alpha}_{Hill} = \left(1/\hat{\xi}\right)_{Hill} = \left(\frac{1}{k}\sum_{j=1}^{k} \log X_{j,n} - \log X_{k,n}\right)^{-1} \tag{4.15}$$

where k is the number of upper order statistic considered (i.e. the number of exceedances over a given threshold), n the sample size. A Hill plot can be constructed such that the estimated tail index ξ is plotted as a function of either the threshold or the k upper order statistics. Hill derived eq. (4.15) using an approach based on the maximum likelihood estimator. De Haan obtained the same result through a regular variation approach, and the same result can be obtained through a mean excess function approach, all these methods yielding equivalent versions of the Hill estimator. Under certain conditions in general satisfied if X_i are iid random variables and F a regular function, the following properties of the Hill estimator hold:

1. *Consistency* If $k \to \infty$, $k/n \to 0$ for $n \to \infty$, then the Hill estimator of the tail index converges in probability to the real value
$$\hat{\alpha}_{Hill} \to^P \alpha \tag{4.16}$$

2. *Asymptotic normality* Under hypotesis of refularity of F, then
$$\sqrt{k}\,(\hat{\alpha}_{Hill} - \alpha) \to^D N(0, \alpha^2) \tag{4.17}$$

The Hill plot can be instrumental in finding the optimal threshold u or, equivalently, the optimal number of upper statistics to use either to estimate of the shape parameter ξ or in the Peak Over Threshold method. Moreover, eq. (4.17) holds for $k = k(n) \to \infty$ at an appropriate rate and for $1 - F(x) = x^{-\alpha} \cdot L(x)$, $\alpha > 0$. However, for a given value of k and slowly varying function L, there is a trade of between variance and bias. For incresing k, the asymptotic variance α^2/k of the estimator $\hat{\alpha}_{Hill}$ decreases, so that one would take k as large as possible. Unfortunately, depending on the second order behaviour of the function L, when doing so a bias in the estimator may arise. In appendix A the Hill plot and Mean excess function for all the datasets are reported.

4.6 Considerations on the threshold selection for POT models

The standard procedure for calibrating a POT model consists in a so called *fixed threshold approach*: the threshold is selected performing a graphical inspection of the data using the aforementioned methods. This approach can be rather subjective and time consuming, since every data set requires an *ad hoc* threshold selection. In this situation it can be convenient to assume a constant quantile level across all the series (or, as an alternative, to fix a given number of exceedance over the threshold per year). In this study this latter method was avoided, since the validity of the GPD and Poisson hypothesis are not ensure just by selecting the same fixed number of yearly exceedances for every dataset. Instead, for every station a graphical inspection of Hill and MEF plot was carried out. The major drawback of the *fixed threshold approach* is that once the threshold has been selected it is considered as fixed and not as a calibration parameter; as a consequence, the associated subjectivity and uncertainty are ignored in the subsequent inferences. A possible alternative is the adoption of mixture models (e.g. Scarrott and Mac Donald, 2012) in which two different distribution are fitted, one to the bulk of the distribution and one to the tail. In this framework the threshold become a parameter of the model to be automatically estimated in the fitting procedure and the uncertainty involved with the threshold choice is naturally accounted for. These models are not used in the practice and suffer some drawbacks; Their application is not easy and a major problem is ensuring that the bulk and the tail fits are bobust to each other. Moreover often the behaviour at the threshold may be problematic (e.g. the fitted density may not be continuous). For this reason the application of POT model rely on a *fixed threshold approach.*

4.7 On the effects of record length and threshold selection on GPD parameter estimation

In its current formulation, the Peak Over Threshold model is derived by fixing a 'high enough' threshold and modeling the excesses arrivals with the Posisson distribution and their magnitude with the GP distribution. Particular care is to be assigned to the study of the Generalized Pareto shape parameter, which controls the shape of the tail, discriminating between bounded, exponential and power law decays. Analyzing the Mean Excess Function and the Hill plots (cfr. chapter 4), the range of thresholds over whom the distribution of excesses is approxiamtely a GPD has been assessed. Similarly, with e.g. the chi squared test the goodness of the Poisson hypothesis for the exceedances has been evalutated. Therefore, the main

Station	$q[mm]$	ξ	ψ $[mm]$	λ
Asheville (NC)	25	0.0356	13.095	7.553
Heerde (NL)	10	0.1079	5.633	20.266
Hoofdoorp (NL)	10	0.0579	5.823	21.360
Kingston (RI)	10	0.0789	14.129	39.316
Livermore (CA)	10	0.1149	7.372	11.712
Milano (IT)	25	0.0957	11.934	10.026
Padova (IT)	25	0.0351	11.101	7.398
Philadelphia (PA)	10	0.0765	12.015	34.914
Putten (NL)	10	0.0817	5.731	20.269
Roosvelt (AZ)	10	0.0376	10.369	13.268
San Bernardo (FR)	20	0.1181	12.501	31.257
Zurich (CH)	10	0.076	7.947	35.495
Bologna (IT)	15	0.106	11.385	11.778
Worcester (SA)	10	0.0092	8.8526	8.737
Albany (GA)	10	0.0806	14.822	39.690

Table 4.2: Calibration of POT model: Thresholds selected (q), fitted shape (ξ) and scale (ψ) parameters of the GPD and mean rate of excesses i.e. Poisson scale parameter λ.

problem with the use of the POT approach is the threshold selection. Serinaldy and Kilsby (2014) showed a twofold effect on the application of POT models is played by threshold selection and record length. First, as the threshold decreases, non extremal values are progressively incorporated in the POT sample. As the sample become larger, the standard deviation of the estimated shape parameter $\hat{\xi}$ decreases; Its mean value, on the other hand, increases, unraveling the heavy tailed nature of the distribution of the excesses. There is also a relationship between the generalized Pareto shape parameter ξ and the record length. Papalexiou and Kudsoyiannis (2013) and Serinaldi and Kilsby (2014) showed, respectively for the AM and POT approaches, that as the sample length L increases, the standard deviation of the estimated $\hat{\xi}$ decreases, as one would expect being the sample size larger. Furthermore, as the number of years of observation L increases, the sample mean of $\hat{\xi}$ tends to converge to a stable and positive value. This findings would suggest the exsistence of an asymptotic value for the shape parameter, as $L \to \infty$. As a consequence, the exponential decay observed in short time series is only an apparent behaviour due to an underestimation of the mean value of the shape parameter. The MEV-GPD approach proposed in the first section is subject to this limitation in the estimation of the shape parameter. In the case of GPD, 1-year is too small a sample, since both bias and standard deviation affect the goodness of fit. Therefore, in the GPD case the Penultimate approximation is to be preferred over the complete MEV espression. The bigger sample size allows a smaller bias and uncertainty in the estimated GPD parameters. We remark that in the case in which instead of the empirical distribution of the random N Poisson is used, the MEV-Penultimate approach correpond with the classical POT method, which can in this case be considered a particular case of MEV distribution.

4.8 Stationary quantile estimation

It is worth recalling that the word stationary refers to the weak or second order stationarity. This assumption implies that only the first order and second order moments of the time series are required to be time invariant and therefore only mean, variance and autocorrelation of lag τ are required to be independent on time. Most of the times, the practical aim of an extreme value analysis is the estimation of high quantiles or, in other words, the determination of the magnitude of the event correspondent to a fixed, and generally 'high', non exceedance probability. The term 'high' points out that the extreme value methodology is often applied in extrapolation outside the range of the available data. The common way to accomplish this task in the engineering practice is based on the key concept of *return time* T_r, that is the average time interval between two consecutive exceedances of the magnitude of the considered event. In general the intensity of a rainfall event depends both on the duration and on the return time of the event we are considering, so that we can write $h(\tau, T_r)$. In the present study we will consider only the latter dependence, considering a given duration of $\tau = 1$ day. The dependance on the duration can be explored as well, and in general the task is carried out by fixing the return time and thus obtaining the so-called intensity-duration curves. However, the case of daily rainfall analysis is of remarkable importance, since this is the most common sampling frequency used in many historical records. The definition of return time requires two hypotheses: stationarity of the process and independence of consecutive realizations. In the case of annual maxima the latter hypothesis holds without the shadow of a doubt, whereas the first should be tested. Under these assumptions, the probability to observe two exceedances of x separated by T years is given as follows:

$$P(T = t) = F(x)^{t-1} \cdot (1 - F(x)) \tag{4.18}$$

Therefore we can write the return time as the expected value of the random variable T:

$$T_r = E[T] = \sum_{t=1}^{\infty} F(x)^{t-1} \cdot (1 - F(x)) \cdot t = \frac{1}{P(X \geq x)} \tag{4.19}$$

The estimation of the intensity of the event associated with a given return time requires as first step the fitting of an extreme value distribution to the available data, that typically is one of the three limiting distributions resultimg from the extreme value theorem. The data to which the distribution is fitted can be either the annual maxima or the k upper order statistics in the peak over threshold approach. In a similar fashion, the concept of hydrological risk can be defined as well for the stationary case, to asses the probability of an event of given duration and intensity to accur in a T-years span

$$R = 1 - \left(1 - \frac{1}{T_r}\right)^T \tag{4.20}$$

4.9 Non-stationary quantile estimation

The usual definitions of hydrologic risk and return time are obtained under two main assumptions: Stationarity and Independence of the annual maxima. Whereas the latter is true in general without the shadow of a doubt, this is not the case for the Stationary hypotesis. Thus, for the aim of studying of extreme events under a non stationary framework, new definitions of T_r and Risk are to be introduced. Let consider the cumulative probability distribution of annual rainfall maxima (e.g. GEV or MEV distribution):

$$P(M_{n,t} \leq x) = H_t(x; \vec{\theta_t}) \tag{4.21}$$

This formulation takes into account a possible variation of the maximum rainfall distribution through the years; In fact now both the set of parameters $\vec{\theta_t}$ and the analytical expression of the distribution H_t might be variable in accordance with time. Let consider a given height of rainfall, say x_0, and a fixed starting year t_0. We are interested in the probability distribution of waiting time for the first annual rainfall maximum to exceed the given value x_0. If we name this new discrete random variable'*wating time between two consecutive exceedances*' M, its probability mass function will be:

$$p(M = m) = f(m) = [1 - H_m(x_0)] \prod_{t=1}^{m-1} H_t(x_0) \qquad \forall m \in \Omega_M \tag{4.22}$$

This is a generalization of the geometric distribution obtained in the stationary case often referred to as *nonhomogeneous geometric distribution*; Hence, when the CDFs of the annual maxima are the same (stationary conditions), this epression yields us the usual result. The CDF of the random variable M can be obtained as well as a sum of all the values of the PMF smaller than a certain value m:

$$\begin{aligned} p(M \leq m) &= F_M(m) = \sum_{i=1}^{m} f(i) \\ &= \sum_{i=1}^{m} [1 - H_i(x_0)] \prod_{t=1}^{i-1} H_t(x_0) \\ &= 1 - \prod_{t=1}^{m} H_t(x_0) \qquad \forall m \in \Omega_M \end{aligned}$$

Therefore the *Return Time* T_r can be defined in a non-stationary framework as the expected value of the aforementioned distribution of waiting times between two consecutive exceedances:

$$\begin{aligned} T_r = E[M] &= \sum_{m=1}^{m_{max}} m \cdot f(m) \\ &= \sum_{m=1}^{m_{max}} m \cdot [1 - H_m(x_0)] \prod_{t=1}^{m-1} H_t(x_0). \end{aligned}$$

This expression can be conveniently simplified (Cooley,2013) as follow:

$$T_r = E[M] = 1 + \sum_{m=1}^{m_{max}} \prod_{t=1}^{m} H_t(x_0). \tag{4.23}$$

This definition is consistent with the usual one of Tr in a stationary framework. However now in the non stationary case Tr is not only a function of the exceedance probability pr (a constant value). On the contrary, Tr will be a function of the time varying exceedance probabilities p_t. The variance of M can be obtained from $var(M) = E[M^2] - T_r^2$ where the second order moment of M will be:

$$E[M^2] = \sum_{m=1}^{m_{max}} x_0^2 [1 - H_m(x_0)] \prod_{t=1}^{m-1} H_t(x_0). \tag{4.24}$$

Let consider an hydraulic structure with a design life of n years which failure will occur as a consequence of the realization of the event $F : X \geq x_0$. In analogy with the stationary case. The reliability of the structure is defined as the probability that no rainfall event exceeding the design rainfall height x_0 will occur in the lifetime of the structure:

$$Re = \prod_{t=1}^{n} H_t(x_0) \tag{4.25}$$

Its completion to one:

$$R = 1 - \prod_{t=1}^{n} H_t(x_0) \tag{4.26}$$

This is defined '*Risk of failure*', in complete analogy with the stationary case.

Chapter 5

Optimal choice of parameters

In this chapter the methods are described that are used in the following analyses to estimate the parameters of Generalized Extreme Value (GEV), Weibull (WEI) and Generalized Pareto (GPD) distributions. The GEV distribution was fitted with Maximum Likelihood (ML), L-Moments (LMOM) and Mixed Methods (MM); the performance of the three method are then compared in a Montecarlo analysis using artificially generated data. For Weibull distribution Least squares (LS) and Maximum Likelihood were used. In the case of GPD Maximum Likelihood, Least Squares and L-Moments were implemented. The performances of the different methods were compared considering the dependence on the sample size.

5.1 Fit of the GEV distribution

5.1.1 Maximum Likelihood

The most commonly used method to obtain a parameter estimation for the GEV distribution is the maximum likelihood estimator. We define the likelihood function as the joint pdf of the available observations $p(\vec{x} \mid \vec{\theta})$, thus considering the sample $\vec{x}$ as a constant vector and $\vec{\theta}$ variable. If we assume the sample $\vec{x}$ to be a vector of realizations of i.i.d. random variables with common pdf $p(x_i \mid \vec{\theta})$, we obtain the following expression for the likelihood function:

$$\nu(\vec{\theta}) = p(\vec{x} \mid \vec{\theta}) = p(x_1 \mid \vec{\theta}) \cdot p(x_2 \mid \vec{\theta}) ... p(x_n \mid \vec{\theta}) \tag{5.1}$$

The optimal choice of the parameters set $\vec{\theta}$ is obtained by maximizing the objective function $\nu(\vec{\theta})$. Usually it is easier to work with the log likelihood function:

$$L(\vec{\theta}) = \log \nu(\vec{\theta}) = \log \prod_{i=1}^{n} p(x_i \mid \vec{\theta}) = \sum_{i=1}^{n} \log p(x_i \mid \vec{\theta}) \tag{5.2}$$

Thus the *Maximum likelihood estimator* $\hat{\theta}$ can be obtained by solving the *likelihood equations* that form a system of p equations, where p is the size of $\vec{\theta}$.

$$\frac{\partial L(\vec{\theta})}{\partial \theta_j} = 0 \quad j = 1, ..., p \tag{5.3}$$

In the case of the GEV density distribution

$$p(x; \xi, \psi, \mu) = \frac{1}{\psi} \cdot \left(1 + \frac{\xi}{\psi}(x - \mu)\right)^{\frac{\xi-1}{\xi}} \cdot e^{-1\left(1+\frac{\xi}{\psi}(x-\mu)\right)^{-1/\xi}} \tag{5.4}$$

and so the log likelihood is given by

$$L(\xi, \psi, \mu) = -n \log \psi - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{n} \log \left(1 + \frac{\xi}{\psi}(x_i - \mu)\right) - \sum_{i=1}^{n} \left(1 + \frac{\xi}{\psi}(x_i - \mu)\right)^{-1/\xi} \tag{5.5}$$

The optimization has been performed using the *Nelder-Mead simplex algorithm*, coupled with the imposition of two constraints necessary for the likelihood function to be defined: $\psi > 0$ and $1 + \frac{\xi}{\psi}(x_i - \mu) > 0$ for each x_i. The tolerance used in the function evalutations was 10^{-6}. A first guess for the unknown parameter set $\vec{\theta_0}$ was necessary as starting point for the numerical optimization. As starting values for the location and scale parameters have been used those determined using the method of moments for the Gumbel distribution, whereas the starting shape parameter has been set equal to $\xi = 0.1$

$$\mu = m - 0.57722 \cdot \psi \tag{5.6}$$

$$\psi = \frac{1}{\pi} \cdot \sqrt{6 \cdot s^2} \tag{5.7}$$

Where $m = \frac{1}{n}\sum_{i=1}^{n}$ and $s^2 = \sum_{i=1}^{n}(x_i - m)^2/(n-1)$ are respectively the sample mean and standard deviation. The assessment of the reliability of the ML estimation of the parameters can be performed exploiting the asymptotic properties of the ML estimator.In fact the standard errors of the estimated parameters are strictly related to the curvature of the likelihood function in the hyper-space of the parameters. The hessian matrix of the likelihood function, evalutated at $\hat{\theta}$, is called *observed information matrix*:

$$I = \left[-\frac{\partial^2 L(\hat{\theta})}{\partial\theta_i \partial\theta_j} \right] \quad i, j = 1, ..., p \tag{5.8}$$

The square roots of the diagonal entries of the information matrix are approximately the standard errors of the estimated parameters $\hat{\theta}_1, ..., \hat{\theta}_p$.

L-Moments

The currently favored method of estimation of the parameters of the GEV distribution is the aforementioned maximum likelihood method. Despite that, its justification and its asymptotic properties are based on large sample theory and there is little assessment of its performance when applied to small samples. In particular in the case of small sample size the hypotesis of normal asymptotic distribution of the estimator $\hat{\theta_{ML}}$ may not hold and moreover the optimization algoritm may be unable to find a global maximum. The conventional method of moment is not suited to estimate GEV parameters, since it can be easily subject to bias in estimation; furthermore the GEV convential k-th moments is not defined if $\xi \geq 1/k$. Hosking (1990) proposed to use L-moments instead of the convential moments; L-moments are defined as expectations of certain linear combinations of order statistics. Given a distribution function $F(x)$ and its inverse quantile function $x(F)$, we define first the probability weighted moments (PWMs) of the r-th order as:

$$\alpha_r = \int_0^1 x(F)(1 - F(x))^r dF, \qquad \beta_r = \int_0^1 x(F)F(x)^r dF \quad r = 1, 2.. \tag{5.9}$$

Thus for a random variable X the L-moments λ_r are defined as linear combination of the PWMs:

$$\lambda_{r+1} = (-1)^r \sum_{k=0}^{r} p^*_{r,k} \beta_k \tag{5.10}$$

where

$$p^*_{r,k} = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)^2} \tag{5.11}$$

In particolar to estimate the GEV parameters the first three L-moments are to be used

$$\lambda_1 = \beta_0$$
$$\lambda_2 = 2\beta_1 - \beta_0$$
$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0$$

The first L-moment λ_1 is the expectation whereas the second λ_2 serves as scale of the distribution, that is the analogous of the traditional standard deviation. Their ratio $\tau = \lambda_2/\lambda_1$ is therefore the coefficent of variation. The higher order L-moment are usually normalized with respect to the scale of the distribution λ_2 thus obtaining the so-called L-moment ratios

$$\tau_k = \frac{\lambda_k}{\lambda_2}, \qquad k = 3, 4.. \tag{5.12}$$

The L-moment ratios determine the shape of the distribution irrespective of the scale of the variables being measured. In particular $\tau_3 = \lambda_3/\lambda_2$ represent the L-skewness of the distribution and $\tau_4 = \lambda_4/\lambda_2$ the L-kurtosis coefficent. The fit of a distribution to a sample using the LMOM method is performed setting the L-moments of the distribution equals to the L-moments of the sample. Let consider a data sample of n observation arranged in ascending order $x_1, x_2, ...x_n$. The first step consists on computing the probability weighted moments for the sample as

$$b_r = \frac{1}{n} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)...(j-r)}{(n-1)(n-2)...(n-r)} \cdot x_j \tag{5.13}$$

For example:

$$b_1 = \frac{1}{n} \sum_{j=2}^{n} \frac{(j-1)}{(n-1)} \cdot x_j$$
$$b_2 = \frac{1}{n} \sum_{j=3}^{n} \frac{(j-1)(j-2)}{(n-1)(n-1)} \cdot x_j$$

By analogy with the L-moments of the distribution, we define with the same linear combination of probability weighted moments the sample L-moments

$$\begin{aligned} l_1 &= b_0 \\ l_2 &= 2b_1 - b_0 \\ l_3 &= 6b_2 - 6b_1 + b_0 \end{aligned}$$

Their general expression being

$$l_{r+1} = \sum_{k=0}^{r} p^*_{r,k} b_k, \qquad k = 0, 1, ...n-1 \tag{5.14}$$

where coefficients $p^*_{r,k}$ are those defined by eq. (5.11). We can then obtain the L-moment ratios

$$\begin{aligned} t &= l_2/l_1 \\ t_r &= l_r/l_2, \qquad r = 3, 4.. \end{aligned}$$

We remark that being the sample L-moments linear functions of unbiased estimates of the probability weighted moments, they are also unbiased estimates of λ_r. In the case of the GEV distribution, the LMOM estimators for location and scale parameter are respectively

$$\hat{\mu} = \hat{\lambda_1} + \frac{\psi}{\xi}\left[1 - \Gamma(1-\xi)\right] \tag{5.15}$$

$$\hat{\psi} = \frac{\hat{\lambda_2}\xi}{(2^{\xi} - 1)\Gamma(1-\xi)} \tag{5.16}$$

The LMOM estimator for the shape parameter $\hat{\xi}$ is the solution of the following equation

$$\frac{1 - 3^{\hat{\xi}}}{1 - 2^{\hat{\xi}}} = \frac{\hat{\tau}_3 + 3}{2} \tag{5.17}$$

Hence the shape parameter can be obtained either by numerical solution of eq. (5.17) or by using the approximate solution:

$$\begin{aligned} \xi &\simeq -7.8590c - 2.9554c^2 \\ c &= \frac{2}{\hat{\tau}_3 + 3} - \frac{\log 2}{\log 3} \end{aligned}$$

Mixed Methods

It has been shown that ML parameter estimator is unbiased but tends to have a large variance for positive values of the shape parameter ξ and this may lead to large errors in quantile estimation. On the other hand, the LMOM method produces a biased estimates but it can be preferable because of smaller variance in its quantile estimates. In particular LMOM estimation of the shape parameter produces a bias increasing with the value of the shape parameter. Morrison and Smith (2002) proposed a combination of the ML and LMOM in order to have an estimator with reduced variance compared to the ML estimator and reduced bias compared to the LMOM estimator. The idea behind the method is to improve the ML estimate of the shape parameter by imposing additional constraints to the optimization problem. In the method used in this work, we maximize the likelihood function as a function of the shape parameter ξ taking both ψ and μ from the LMOM. Thus the optimization problem becomes:

$$\begin{array}{rl} \text{maximize} & \log \nu(\vec{\theta} \mid x) \\ \text{subject to} & \psi = \dfrac{\hat{\lambda_2}\xi}{(2^{\xi}-1)\Gamma(1-\xi)} \\ \text{and} & \mu = \hat{\lambda_1} + \dfrac{\psi}{\xi}\left[1-\Gamma(1-\xi)\right] \\ & \xi(x_i - \mu) \geq \psi \quad i = 1, ...n \end{array}$$

An alternative approach Smith and Morrison proposed consists in maximizing the likelihood function as a function of scale and shape parameter, sobstituting the sole location parameter from the LMOM equation (5.15). The estimator of the parameters of the GEV is then the solution to the following optimization problem:

$$\begin{array}{rl} \text{maximize} & \log \nu(\vec{\theta} \mid x) \\ \text{subject to} & \mu = \hat{\lambda_1} + \dfrac{\psi}{\xi}\left[1-\Gamma(1-\xi)\right] \\ & \xi(x_i - \mu) \geq \psi \quad i = 1, ...n \end{array}$$

In this case the optimization problem involves only the variable ξ. The initial point ξ_0 was taken to be the LMOM estimate of the shape parameter. In this way we use only the first two L-moments, avoiding the estimator for τ_3 which might have large bias for small sample size and small values of ξ.

Monte Carlo simulations for GEV

In order to asses the performances of the different estimators of the GEV parameters, a Monte carlo simulations was performed. We generated 10000 samples from a GEV distribution with given parameter $\xi = 0.1, \psi = 12, \mu = 50$. These are typical value we found fitting the GEV to datasets with various length and from different and climatic conditions. The length of the generated samples is 100 years, a common value among the analyzed datasets. For every simulated series, the annual maxima were fitted to the GEV distribution using L Moments, Maximum likelihood and Mixed methods. We report the finding of the analysis in table (5.1). We point out that the Maximum likelihood allows the best estimation of the shape parameter whereas both L-Moments and Mixed Methods procude a more relevant error: this is particularly important parameter since it is responsible of the tail behaviour of the distribution. On the other hand the two latter method produce slightly better estimators of the scale and location parameters; the difference between the performances of L-Moments and Mixed Methods is negligible and therefore there is no convenience in the application of the latter method. Moreover when applied to small samples, the latter method's algorithm may fail in finding the exact minimum, whereas the L-Moment method does not have this limitation and it is more robust. Therefore we argue that the better choice is maximum likelihood (for large enough sample sizes) or L-Moment (in the case of smaller samples).

5.2 Fit of the Weibull distribution

Two methods have been applied to estimate the parameters of the Weibull distribution: Maximum likelihood and least squares. The first one is to be preferred because of the higher performances and the usefulness of the asymptotic properties of the ML estimator. Though, in the case of very small sample sizes the variance of the estimated parameters increases and the optimization algorithm used in the ML approach might not succeed in finding the global maximum.

Method	Parameter	mean est.	stdv. est.	mean quadratic error
Maximum likelihood	ξ	0.096	0.081	0.0386
	ψ	9.903	0.857	0.0097
	μ	50.058	1.145	0.0012
L Moments	ξ	0.092	0.083	0.0781
	ψ	9.998	0.929	$1.271 \cdot 10^{-4}$
	μ	50.030	1.147	$6.031 \cdot 10^{-4}$
Mixed methods	ξ	0.093	0.082	0.0623
	ψ	9.998	0.929	$1.271 \cdot 10^{-4}$
	μ	50.030	1.147	$6.031 \cdot 10^{-4}$

Table 5.1: Results of the Montecarlo simulation

Maximum Likelihood

The probability density function of the Weibull distribution reads:

$$f(x; C, w) = \frac{w}{C}\left(\frac{x}{C}\right)^{w-1} e^{-\left(\frac{x}{C}\right)^{w}} \tag{5.18}$$

Consider then a given sample $x_1, ..., x_n$ as a constant vector and the scale and shape parameters of the distribution function C, w as random variables. If we assume the sample $\vec{x}$ to be a vector of realizations of i.i.d. random variables with common pdf $f(x_i \mid C, w)$, we obtain the following expression for the log-likelihood function:

$$L(C, w|\vec{x}) = N \log \frac{w}{C} + (w-1) \cdot \sum_{i=1}^{n} \log \frac{x_i}{C} - \sum_{i=1}^{n} \left(\frac{x_i}{C}\right)^{w} \tag{5.19}$$

The maximum $(\hat{C}, \hat{w})$ can be determined by setting the two partial derivatives of eq. (5.19) equal to zero and by solving numerically the two resulting non linear equation:

$$\frac{\partial L}{\partial w} = -n + \sum_{i=1}^{n} \left(\frac{x_i}{C}\right)^{w} = 0$$

$$\frac{\partial L}{\partial w} = \frac{n}{w} - \sum_{i=1}^{n} \log\left(\frac{x_i}{C}\right) - \sum_{i=1}^{n} \left(\frac{x_i}{C}\right)^{w} \cdot \log\left(\frac{x_i}{C}\right) = 0$$

As for the GEV case, an estimation of the standard error in the parameter estimates can be obtained by evalutating the hessian matrix of the log-likelihood function. Moreover, since the Weibull probability density function satisfies the regularity conditions, the asymptotic proprieties of the maximum likelihood estimator hold and the confidence interval can be obtained by assuming a normal distribution for the estimated parameters.

Least sqaures method

For small sample sizes, maximum likelihood might be unable to find a global maximum. In these cases an alternative method is the least squares fitting method. To apply a least square regression to the Weibull distribution, it is useful to define first the reduced variate as follows

$$Y_r = \log(-\log(1 - F_i)) = w \cdot (\log y - \log C) \tag{5.20}$$

We point out that this is the equation of a straight line in the $(Y_r, \log y)$ plane. Therefore it is possible to perform a linear regression in order to obtain the slope w of the line and the Y_r-axis intercept $-w \cdot \log C$. To do so it is first necessary to approximate the non exceedance probability with a plotting position formula (such as for example the Weibull plotting position $F_i = i/(N+1)$ where i represents the position of the $i - th$ element in the sample sorted in ascending order. The linear regression has been carried out minimizing the sum of the normal distances to the line, squared. The procedure yields us the two unknown parameters of the Weibull distribution function

$$w = \frac{S_h}{S_{y_R}}, \qquad C = e^{\left\{\bar{h} - \frac{S_h}{S_{y_R}} \cdot \bar{y_R}\right\}} \tag{5.21}$$

where

$$\bar{h} = \frac{1}{n} \cdot \sum_{i=1}^{n} h_i$$

$$\bar{Y_R} = \frac{1}{n} \cdot \sum_{i=1}^{n} Y_{R_i}$$

$$S_h = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(h_i - \bar{h}\right)^2}$$

$$S_{Y_R} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(Y_{R_i} - \bar{Y_R}\right)^2}$$

5.3 Fit of the Generalized pareto distribution

Maximum Likelihood

In the case of the GPD the two parameters density function has the expression

$$f(\vec{y}; \xi, \psi) = \frac{1}{\psi} \cdot \left(1 + \frac{\xi}{\psi} \cdot y\right)^{-1/\xi} \tag{5.22}$$

log likelihood function expression is the following

$$L(\xi, \psi|\vec{y}) = -n \log(\psi) - \left(\frac{\xi - 1}{\xi}\right) \sum_{i=1}^{n} \log\left(1 + \frac{\xi}{\psi} \cdot y_i\right) \tag{5.23}$$

We recall that the numerical optimization has to satisfy the domain restriction for the two parameters $(\hat{\xi}, \hat{\psi}) \in [0, +\infty)$. This method has been proved to work fine if $\xi > -1/2$. In this case it has been showed that the usual properties of the MLE like consistency and asymptotic efficency hold:

$$\sqrt{n} \cdot \left(\hat{\xi} - \xi, \frac{\hat{\psi}}{\psi} - 1\right) \rightarrow^{d} N\left(0, M^{-1}\right) \tag{5.24}$$

where $N\left(0, M^{-1}\right)$ is a bivariate normal distribution with mean vector 0 and covariance matrix proportional to the hessian matrix of the likelihood function:

$$M^{-1} = (1 + \xi) \cdot \begin{pmatrix} 1 + \xi & 1 \\ 1 & 2 \end{pmatrix} \tag{5.25}$$

L-Moments

Similarly to our discussion in the previous section for the GEV, the L-moment method can be used as well in the case of the GPD. The theoretical L-moments of the distribution are defined as follows:

$$\xi = \frac{3\tau_3 - 1}{1 + \tau_3}$$
$$\psi = (1 - \xi)(2 - \xi) \cdot \lambda_2$$
$$\mu = \lambda_1 - (2 - \xi) \cdot \lambda_2$$

By replacing the the L-moments with the empirical estimators we obtain a system of three equations whose solution yields us the parameter estimators. Hosking and Wallis (1990) also give formulae to evalutate the standard errors of these estimators. When GPD is assumed to be the distribution of the excesses over a given threshold, the location parameter is fixed and equal to the threshold; as a consequence the previous problem shrinks to a bidimensional system of equations.

Least sqaures method

A least squares method has been used to compare its performances with the traditional methods of MLE and LMOM. Given a sample $x_i, i = 1, n$ we can write

$$-\xi \cdot \log(1 - F(x_i)) = \log\left(1 + \frac{\xi}{\psi} x_i\right) \tag{5.26}$$

Hence we define the empirical non exceedance frequencies of the elements in the sample using the Weibull lotting position formula $Fi = i/n + 1$ and choose to minimize the sum of the squared distances defined as follows

$$\sum_{i=1}^{n} \left[\log(1 - Fi) + \frac{1}{\xi} \log\left(1 + \frac{\xi}{\psi} x_i\right)\right]^2 = min \tag{5.27}$$

In this case the regression problem is not linear, but the solution of the previous problem can be easily obtained using a numerical optimization algorithm.

Station	*wetdays*	Scale parameter	Shape parameter
Asheville (NC)	127	6.638	0.8025
Heerde (NL)	188	3.788	0.826
Hoofdoorp (NL)	194	3.624	0.795
Kingston (RI)	115	9.323	0.789
Livermore (CA)	57	5.634	0.849
Milano (IT)	109	7.734	0.773
Padova (IT)	105	6.727	0.751
Philadelphia (PA)	118	7.632	0.773
Putten (NL)	174	4.273	0.871
Roosvelt (AZ)	49	7.209	0.839
San Bernardo (FR)	117	10.793	0.878
Zurich (CH)	188	4.783	0.736
Bologna (IT)	93	6.175	0.785
Worcester (SA)	38	6.855	0.954
Albany (GA)	105	10.396	0.753

Table 5.2: Weibull scale (C) and shape(w) parameters and mean value of the yearly number of wet days for all the considered stations.

Chapter 6

Results: Stationary Series

Despite the increasing attention to the non-stationary analysis of hydrologic extremes, nowadays the engineering practice still focus on the traditional stationary analyses of extremes. Therefore, a validation of the MEV model requires a preliminary assessment in a stationary framework. In the following sections we describe in detail the method used to produce synthetic stationary data sets and the results obtained by estimating high quantiles with both the traditional GEV and the MEV approach. We first compare the performances of the various methods used to estimate the parameters of the GEV. Then, the performances of the MEV are explored, perticularized for the case of a Weibull parent distribution.

6.1 Method of analysis

In order to evalutate the performances of the different models in an actual stationary case, a method has been developed in this study which allows the use of observed rainfall records. The following procedure has been applied to all the selected time series of daily rainfall records: synthetic time series were generated using the actual daily precipitated amounts of the original series, but randomly scrambling their respective positions in the time series in order to obtain new data sets charcterized by a lack of serial correlation and temporal trends, but sharing the same pdf of the observed rainfall. In this way data sets are obtained that are stationary by definition, to which apply the MEV and GEV models and evalutate their respective performances in a stationary framework. This procedure preserves the original distribution of daily rainfall values and the distributions of the parameters of the distribution used to describe them. On the other hand, a simple scrambling of the single daily values over all the time series would change the values of the yearly numbers of rainy days. In the original series the yearly numbers of wet days n_i are considered to be realizations of a discrete random variable N, whose distribution function is not known *a priori*. After the implementation of the scrambling procedure the distribution of N will not be the original one, but only its expected value will remain the same. In fact, after the scrambling the distribution of the rainy days arrivals will be a Poisson distribution (the number of realizations in any temporal window will be independent on the number of realizations in any other disjoint window) but in the original case this may not be (and in general is not) the case. To eliminate this shortcoming the following procedure has been used in order to generate randomly scrambled time series without changing the distribution of the random variable N:

1. The m realizations $n_1, .., n_m$ of the random variable N (yearly number of wet days) in the m years of the original series have been sorted in a random way obtaining a new sample n_i^j where the superscript j refers to the new random order of the sample. The new time series has been generated by allocating for every year i a number n_i^j of rainy days.

2. All the non-zero daily rainfall totals of the original time series have been randomly scrambled. For every year i of the new synthetic series, n_i^j daily rainfall realizations randomly selected among the whole time series have been allocated in the n_i^j empty slots.

To perform a benchmark of GEV and MEV performances in the stationary case, 1000 randomly scrambled time series have been generated from every data set. For every synthetic series, a first window of t years was used to fit the extreme value distributions to the data and, for some fixed return time (i.e. non exceedance probabilities), the pertaining quantiles were estimated. To assess the performances of the different models, the estimated quantiles x_{est} for a fixed T_r were compared with the observed quantiles x_{obs}, evalutated using the whole synthetic time series. In particular the annual maxima of the whole m years record have been ranked in ascending order and for every value the empirical non exceedance frequency have been estimated using the Weibull plotting position formula. The reliability of such non exceedance frequency F_i and of the corresponding return time $T_r = 1 - 1/F_i$ depends on the ratio of length of the series over the considered return time. As the return time decreases with respect to the length of the series, for the weak law of large numbers the exceedance frequency get closer to the real survival probability and the interarrival time of a given height of rainfall tends to its return time. This is the reason why particular importance was given to the length of series in the data sets selection process (All the selected datasets cover more than a century, spanning from 106 to 275 years of continuous observations). This allowed to evalutated the effects of the extrapolation outside the range of the data used to compute the theoretical quantiles (30 or 50 years). For every considered window length T and return time T_r, the distribution of the error in the predicted magnitude of the event was estimated over all the random generations, using as indicator the root mean squared error (RMSE) defined as

$$RMSE = \sqrt{\sum_{i=1}^{ngen} \left(\frac{x_{est_i} - x_{obs_i}}{x_{obs_i}} \right)^2} \tag{6.1}$$

This measure of the error was used beacuse it accounts both for the bias in the reliability of estimates and the standard deviation of the estimated values; In particular the latter gives an indication of the stability of the method with respect to the sample available. We also evalutated the standard deviation in the result obtined with MEV and GEV. Histograms of the error were used to characterized the error distribution for all the station over all the 1000 random generations of the series.

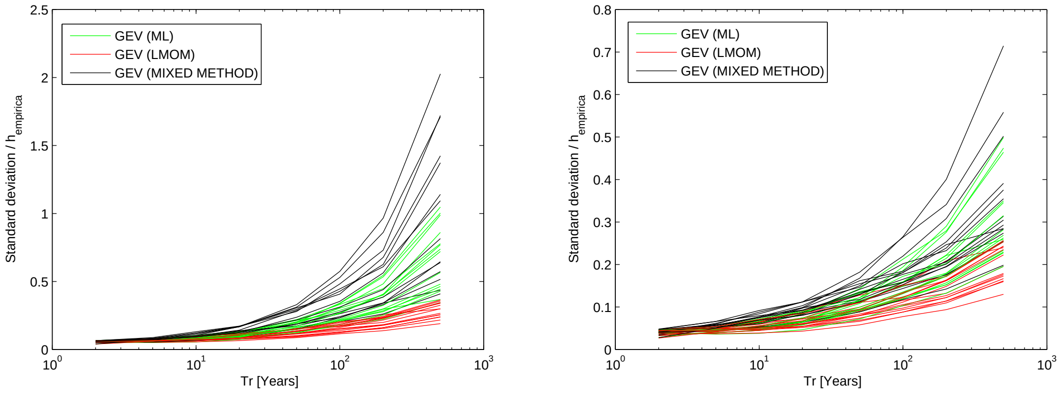
6.2 Benchmark of the estimators for the GEV parameters

Three different fitting methods have been used to estimate the GEV parameters: Maximum likelihood, L-Moments and Mixed Methods. The performances of these three approaches were then compared by applying all the methods in the stationary analysis for all the selected datasets. For every scrambled series, periods of 30 and 50 years were randomly selected. Respectively, samples of 30 and 50 annual maxima were obtained to whom the GEV pdf was fitted by means of the three different estimation methods. The results are shown in the figures 6.1 and 6.2, in terms of non-dimensional Standard Deviation and RMSE respectively. The global performances over all the datasets suggest that the method that outperforms the others is LMOM. In particular, the smaller the sample size, the larger is the increment of performances allowed by the L-Moments. For a sample size of 50 annual maxima the difference is still noticeable in terms of stardand deviation whereas if we look at the RMSE performances are similar, with a few exceptions corresponding to stations where the ML error is still remarkably bigger than LMOM. For larger sample sizes, instead, the ML is expected to refine its performances more than LMOM and therefore in such cases it might be competitive as well. The worst method, in terms of both standard deviation and root mean square error, is the mixed method. This may be due to the fact that in this case the optimization algorithm involves only the shape parameter. It is speculated that the combination of parameters obtained with the mixed method does not correspond to a maximum of the likelihood function and neither features the properties of the LMOM estimator. These findings are coherent with the results obtained performing the Montecarlo analysis whose outcomes are reported in chapter 5. Therefore, when analyzing sample of this size 50 years or smaller, using LMOM ensures the best performances.

6.3 MEV-Penultimate and MEV-Complete

The performances of MEV-Complete with MEV-Penultimate have been compared for the noteworthy case of a Weibull parent distribution. In the former case the Weibull distribution was fitted to the daily data in every single year of the period of record. Hence, the parameters C_i, w_i and the number of rainy days n_i were computed for every year. Thereafter, the quantile associated with a given non exceedance probability (or T_r) could be computed by solving numerically the MEV-Complete expression given by:

$$\zeta(x) = \frac{1}{T} \sum_{j=1}^{T} \left[1 - e^{\left(-\frac{x}{C_j} \right)^{w_j}} \right]^{n_j} \tag{6.2}$$



$T = 30$ years

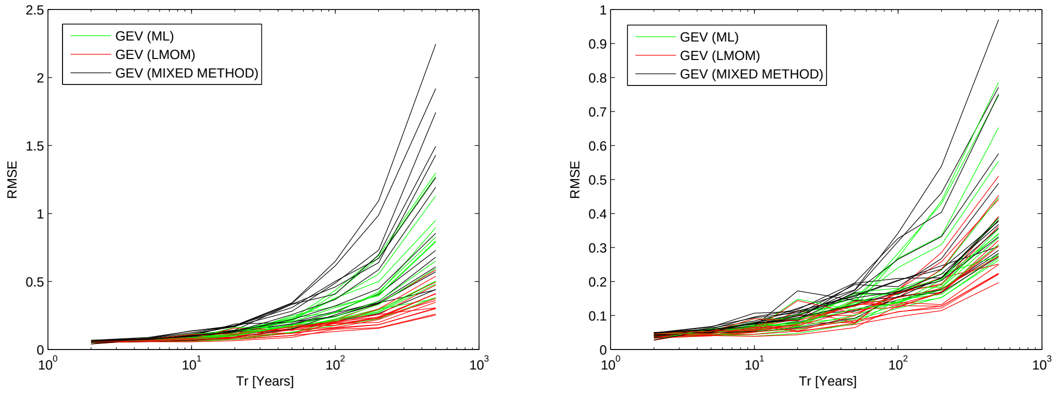
$T = 50$ years

Figure 6.1: Standard deviation of the quantiles estimated using GEV fitted with ML, LMOM and MM for different return times and record lengths. Every line corresponds to a single station, for which the standard deviation was computed over 1000 random scrambling of the original dataset

For what concerns the MEV-Penultimate, only the second order randomness in the cardinality N is accounted for, and the MEV expression corresponds to the penultimate approximation. Hence, for a fixed window of T years the Weibull distribution has been fitted to the whole time inteval and the mean value of the yearly number of rainy days has been computed. Then, the parameters of the Gumbel distribution could be obtained as:

$$\alpha = \frac{1}{C^w}$$
$$\mu = C^w \cdot \log \bar{n}$$

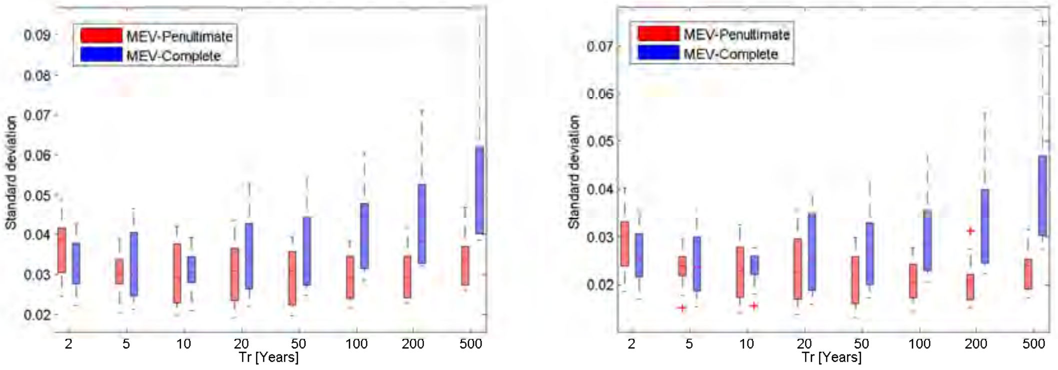
Afterwards, inverting the Gumbel formula the quantiles corresponding to any given value of the T_r were computed. It is worth to be remarked that this is not an asymptotic method and the actual rate of convergence (i.e. the error, for the given value of $\bar{n}$) is only the one of the Cauchy approximation. In Fig. 6.3 and 6.4 boxplots are presented showing respectively non-dimensional Standard Deviation and RMSE of MEV Complete and MEV Penultimate for all the 15 analyzed datasets. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles respectively, the whiskers extend to the most extreme data points not



$T = 30$ years $T = 50$ years

Figure 6.2: RMSE of the quantiles estimated using GEV fitted with ML, LMOM and MM for different return times and record lengths. Every line corresponds to a single station, for which the RMSE was computed over 1000 random scrambling of the original dataset

considered outliers, while outliers are plotted individually. The standard deviation is smaller in the case of the Penultimate distribution: this is due to the fact that, for a record of T years, in the MEV Penultimate approch the yearly number of wet days and the Weibull scale and shape parameters were computed within the whole period of record whereas the same quantities were computed in every single year in the MEV-Complete formulation. On the other hand, MEV-Complete outperforms the MEV-Penultimate if one considers the RMSE as a measure of the error. Fig 6.4 shows that the error of the MEV Complete is general smaller both in the case of $T = 30$ and $T = 50$ years. Moreover, increasing the record length T it is observed that in the advantage of MEV-Complete over the Penultimate formulation increases. This can be easily observed by applying the two methods to the original records, thus analyzing records with lengths over 100 years (results are shown in appendix C for all the stations). In this case the error of the Penultimate is, in general, greater than the one of the MEV-Complete in the majority of the analized stations.



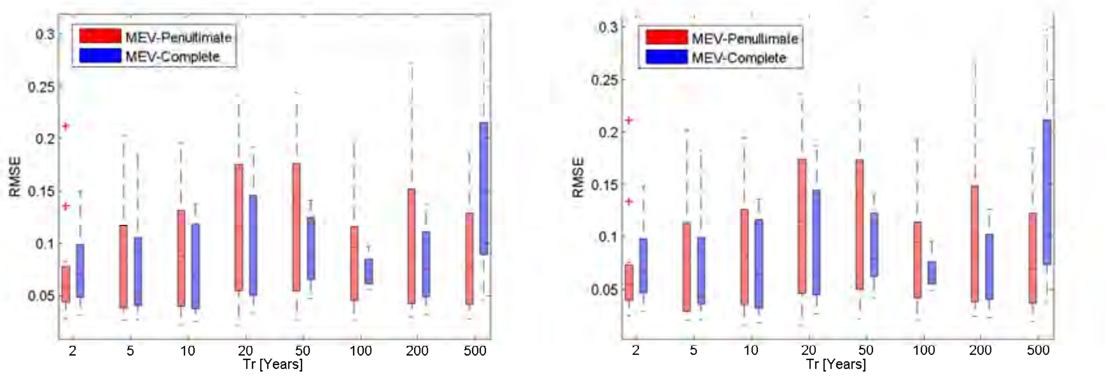
$T = 30$ years

$T = 50$ years

Figure 6.3: Standard deviation of the quantiles estimated using MEV-Complete and MEV-Penultimate for different return times and record lengths. For every station a single value of the standard deviation was computed over 1000 random scrambling of the original dataset; The boxplots represent the distribution of the standard deviation over all the stations.

6.4 Benchmark of GEV and MEV performances

In this section the results obtained with annual-maxima fitted GEV-AM (henceforth referred to simply as GEV), POT and MEV-Complete are compared for all the stations. The thresholds used in the Peak Over Threshold have been determined using a *fixed threshold approach* as discussed in Chapter 4, and vary from $10mm$ to $25mm$ for all the stations. As in the previous case, we here report nondimensional standard deviation and RMSE over 1000 random generation for all the series. For every one of the 1000 random generation, the three distribution have been calibrated using windows of 30 and 50 years; The fitted distribution then have been used to estimated the quantiles associated with the most common return times used for practical purposes (spanning from 2 to 500 years, corresponding to exceedance probability ranging from 0.5 to 0.005). GEV distribution was fitted to the sample of annual maxima using LMOM, as suggested from the analysis carried out in section 6.2. POT have been applied after a proper threshold selection for every station. The GPD was then fitted to the excesses over the threshold using ML estimator. MEV distribution was calibrated by fitting a Weibull distribution



$T = 30$ years

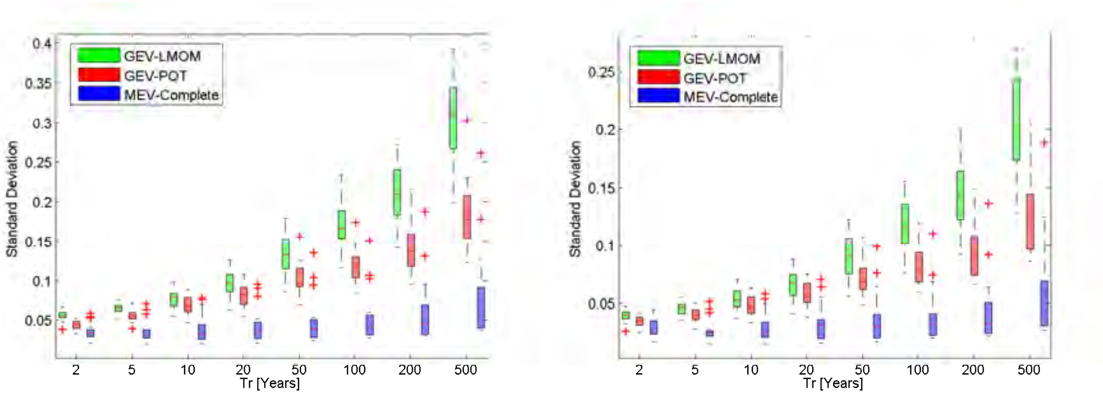
$T = 50$ years

Figure 6.4: RMSE of the quantiles estimated using MEV-Complete and MEV-Penultimate for different return times and record lengths. For every station a single value of the RMSE was computed over 1000 random scrambling of the original dataset; The boxplots represent the distribution of the RMSE over all the stations.

to the daily values of every year of the considered period of record using ML estimation for scale and shape weibull parameters. To assess the goodness of fit implied by the various approaches in quantile estimation, the Root Mean Square Error (RMSE), as defined in eq. 6.1 and the non-dimensional Standard Deviation (STDV), obtained dividing the standard deviation by the corresponding observed quantile, were computed for every station. The distribution of RMSE and STDV over all the stations was then explored using boxplots, as showed in Fig. 6.5 and Fig. 6.6 for STDV and RMSE respectively. GEV has the largest STDV. POT allows to reduce STDV using more data than GEV, but if the threshold is too low a bias may enter when poisson end/or GPD hypothess are no longer met by threshold excesses. The STDV of the MEV estimated quantiles is remarkably smaller if compared with GEV and POT.

The boxplots representing the RMSE show that, for the smaller return times (from 2 to 50 years) the RMSE of GEV and POT is small and the diffence among the stations is not relevant, whereas in the case of MEV the spreding of the distribution is more relevant, suggesting a difference in the behaviour of the various stations. This is speculated to be linked to the goodness of fit of the Weibull distribution, which is not the same in all the datasets. In some cases the adoption of the MEV approach leads to a limited bias in quantile estimation. When considering higher return times ($T_r = 100, 200$) the role played by uncertainty grows. In this case we are extrapolating outside the range of the data used to fit the distribution (30 or 50 years) but the observed quantiles are still reliable, because evalutated using all the years of record available (from 106 to 275 years, depending on the particular station). The boxplots in this case show that MEV ouperforms both GEV and POT approaches, leading to a more accurate quantile estimation. The effect of the length of the record can be evalutated as well; In the case $T = 30$ years the performances of POT and GEV decrease with respect to the case of $T = 50$ years window. This suggests that the fit of GEV or GPD to a sample of smaller size leads to an increase in the uncertainty when a fixed 'high' quantile is to be estimated. On the other hand, the MEV approach uses more data from the bulk of the distribution and does mantain similar performances in both the cases. The POT in turn outperforms GEV, using more data than just the annual maxima. The reduction in the uncertainty inherent in the MEV approach can be explored by analizing the pdfs of the error. Histograms of the non-dimensional error, defined as $\epsilon = (h_{est} - h_{obs})/h_{obs}$ are reported in the figures in panel 6.8 for POT and MEV. The histogram pertaining to the GEV is not considered here, being outperformed both by POT and MEV in the error standard deviation. For $T_r = 20$ the variance of the error distribution is very similar for POT and MEV; The histogram of MEV error exhibits a bimodal pdf with a peak centered in $\epsilon = -0.15$, showing that for some station a poor fit of the Weibull distribution may lead to a bias in the estimated annual maximum cdf. Incresing the T_r the superiority of MEV over POT become clearer, as suggested by the reduced variance of the error pdf.

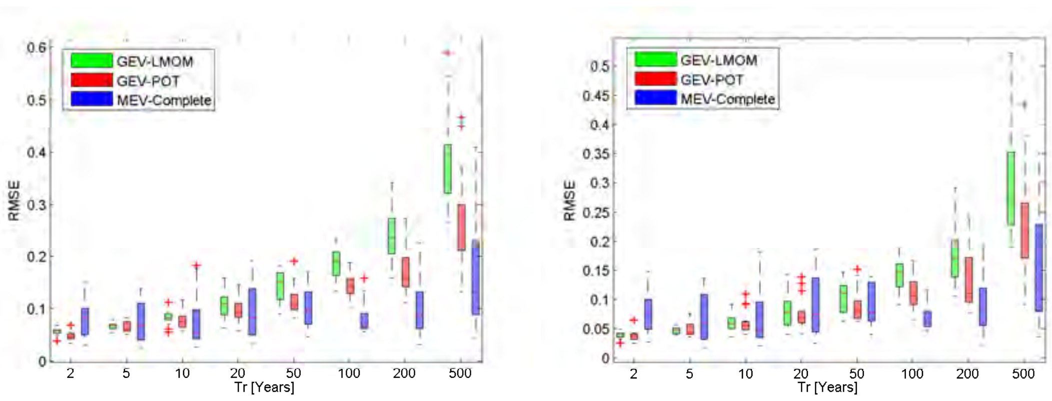
The dependance of the various methods on the sample size was explored by setting a value of return time and considering periods of record of different lengths, spannig from 2 to 100 years. Only the stations with 140 years of records or more were used in this analysis. For every dataset, 1000 random reshuffling of the original series have been carried out, using the procedure described in Section 6.1. For every synthetic series, samples were extracted of different lengths. The quantiles associated with a given return times ($T_r = 100$ years) were then computed with each method. The RMSE was computed over all the random generations and its distribution over all the stations is reported in the boxplots in Fig. 6.7.



$T = 30$ years $T = 50$ years

Figure 6.5: Standard deviation of the quantiles estimated using MEV-Complete, GEV-LMOM and POT approaches for different return times and record lengths. For every station a single value of the standard deviation was computed over 1000 random scrambling of the original dataset using each method; The boxplots represent the distribution of the standard deviation over all the stations.

For small sample sizes, MEV outperforms both GEV and POT, whose estimated values have a considerable variance. Increasing the sample size, GEV and POT estimates converges to smaller values of RMSE whereas the distribution of the RMSE of MEV distribution does not appear tobe dependent on the sample size. This behaviour underline an interesting property of the MEV approach: MEV-estimated quantiles are not overly sensible neither to the specific sample nor to the sample size. This is due to the fact that the MEV approach uses information from the bulk of the daily rainfall distribution and therefore is able to infer the general characters of the underlying stochastic process. The GEV and POT approaches are good models for the sample of annual maxima on which they have been calibrated (as Fig. 6.7 shows for high sample sizes, where the period of record used to compute theoretical qualtiles tends to the whole record from which observed quantiles are drawn). For small sample sizes, instead, GEV ensures a good modeling of the ”local”’ statistical properties of the sample at hand, but when the observed quantiles are drawn from the longer record it fails to ”‘generalize”’ and the high RMSE observed indicates that GEV, when fitted to small samples, does not manage to capture the underlying statistical properties of the population.



$T = 30$ years

$T = 50$ years

Figure 6.6: RMSE of the quantiles estimated using MEV-Complete, GEV-LMOM and POT approaches for different return times and record lengths. For every station a single value of the RMSE was computed over 1000 random scrambling of the original dataset using each method; The boxplots represent the distribution of the RMSE over all the stations.

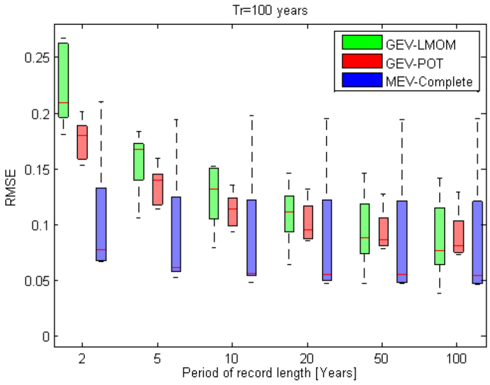
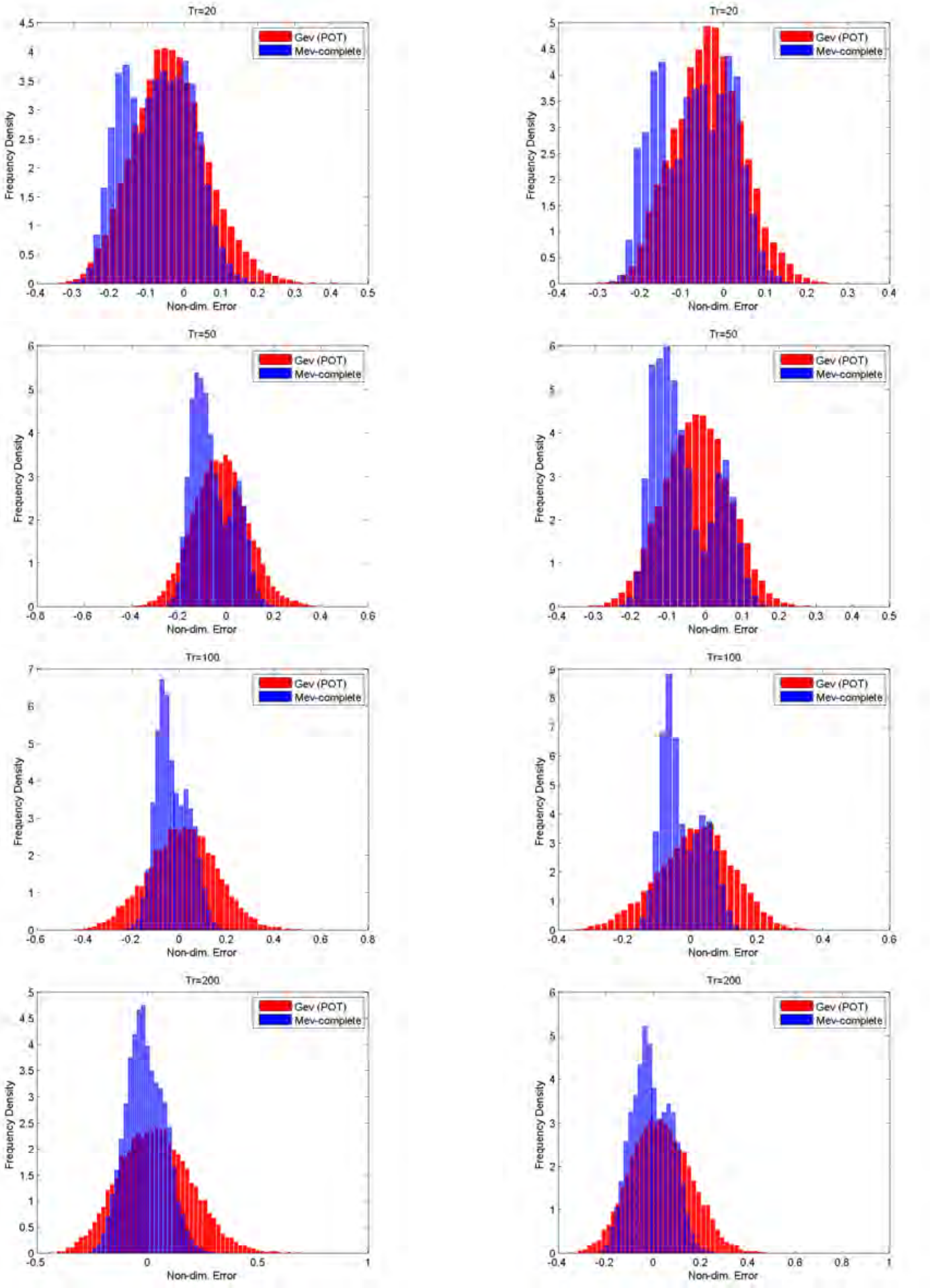


Figure 6.7: Boxplot representing RMSE obtained with GEV, POT and MEV for a fixed $T_r = 100$ years for various windows sizes. Only the record longer than 140 years were selected in this case.



$T = 30$ years

$T = 50$ years

Figure 6.8: Histograms representing the distribution of the error over all the random generations for all the stations analyzed.

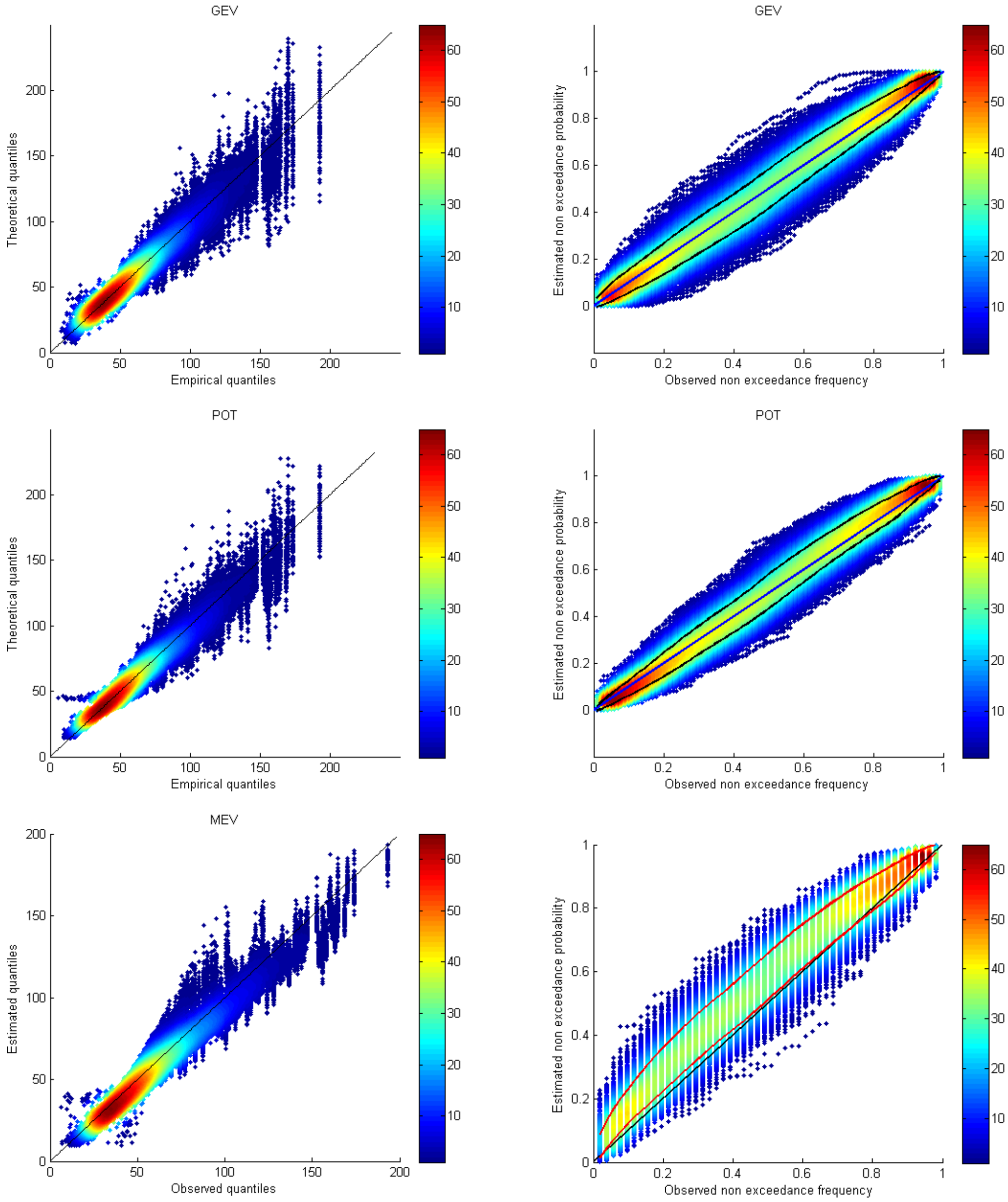


Figure 6.9: QQ and PP plots for all the stations. Estimated quantiles have been computed fitting GEV, POT and MEV distributions to samples of 30 years and then have been compared with the quantiles observed in a 100 years sample. Both samples have been obtained for 100 consecutive random scrambling of every station, so that independece was ensured between the two samples used to evalutate observed and estimated quntiles.

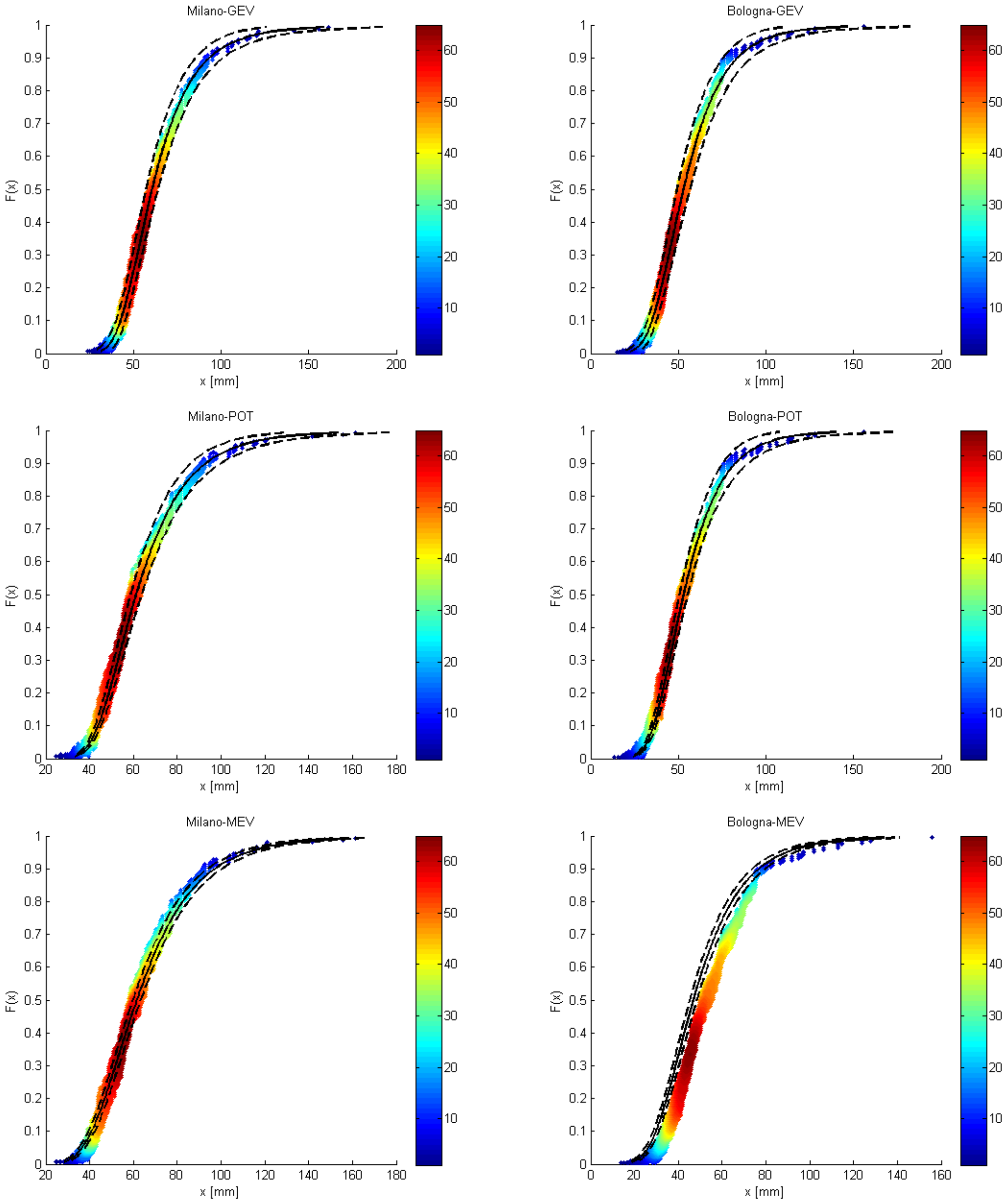
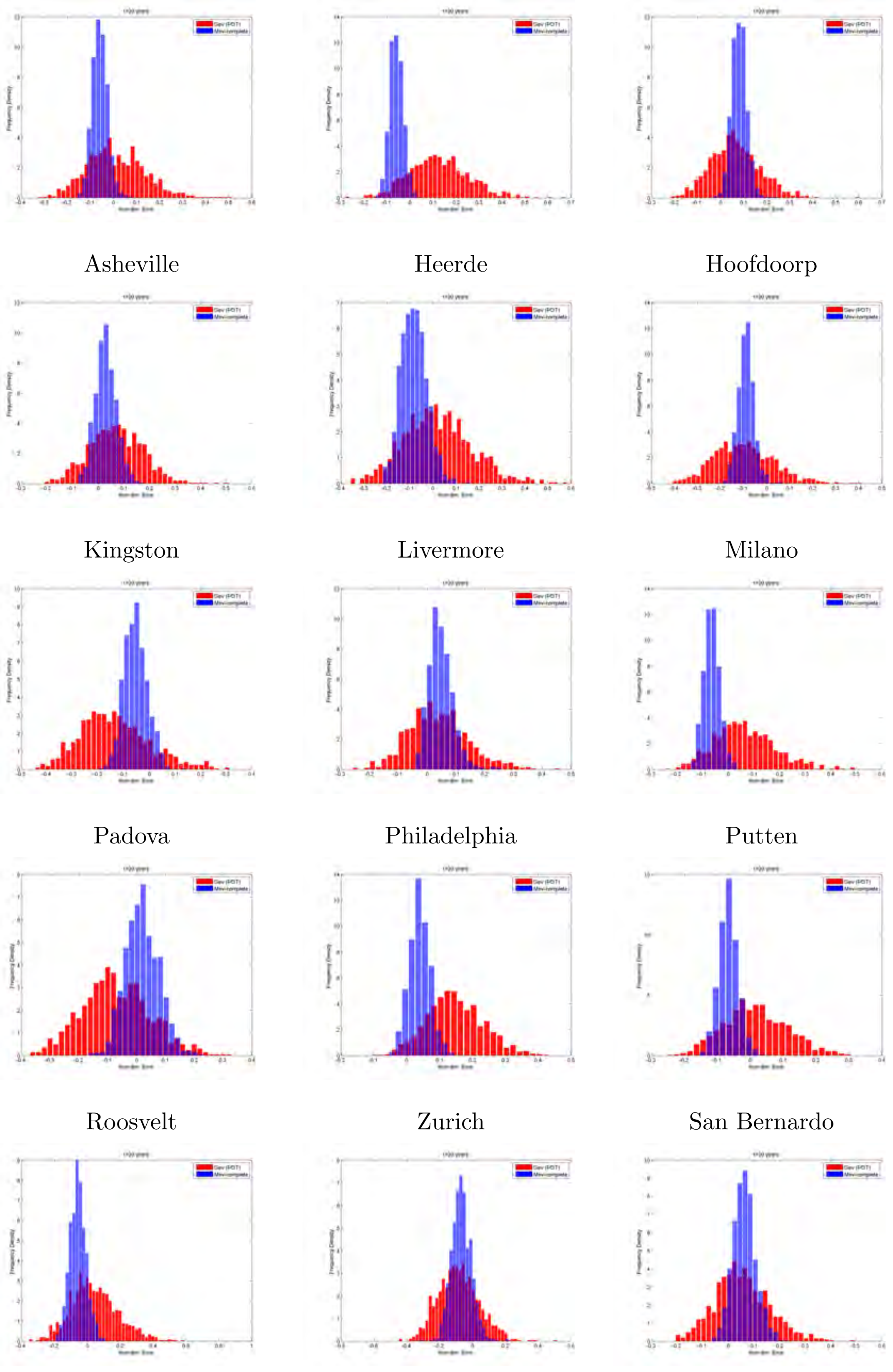


Figure 6.10: Empirical cdf of the annual maximum and theoretical cdf estimated using POT, GEV and MEV distributions obtained performing 100 random reshuffling of the two long records of Bologna and Milano. Solid line represents mean value over all the random generations while dashed lines represent one standard deviation distances from the mean. Color of the scatter points is indicates their local density

.



Asheville

Heerde

Hoofdoorp

Kingston

Livermore

Milano

Padova

Philadelphia

Putten

Roosvelt

Zurich

San Bernardo

Bologna

Worcester

Albany

Figure 6.11: Distribution of the error for the single stations for $T_r = 100$ years.

Chapter 7

Results: non-stationary series

7.1 Is stationarity really dead?

Traditionally, water resources analysis has always been based on the stationarity assumption. Under this hypothesis, the fluctuations of the natural process at hand are supposed to vary in time within a same fixed envelope. As a consequence any measured value (e.g. annual discharge or annual maximum daily rainfall) is the realization of a random variable featured by a given unchanging distribution. When the aim of analyis is extreme weather risk assessment, this time invariance is projected into the future. Some recent works (e.g. Milly et al., 2008) argue that the hypothesis of stationarity has long been compromise by several factors. The first source of non-stationarity is the human intervention on the natural environment. Water infrastuctures and land use change, for example, have a deep impact on water quality and on the risk of flooding events occurrece. The same has occurred for rainfall: Anthropogenic global warming has led to an increasing in air water holding capacity, which is speculated to be one of the reasons of more frequent and intense rainfall events. Effectively, circulatory and thermodynamic responses to human activities have been linked to changes in means and extremes of precipitation, evapotranspiration rate and distribution of river discharges. The natural climate change is another source of non-stationarities: climatic dinamycs are characterized by internal low frequency oscillations which might have an impact on daily rainfall pdf. For example the North Atlantic Oscillation (NAO) impact on extremes has been studied (Marani and Zanetti, 2014) and correlations with the number of yearly rainy events and with the occurrence of extreme events have been found. Milly et al. (2008) state that water resource engineers are required to adopt non stationary models when assessing hydrologic risk or optimizing water systems management, *de facto* depracating most of the currently adopted models. On the other hand, the complexity of nonstationary models may lead to an increase in the uncertainty: in fact non-stationary models are fitted by inductive inference and the structure of the model may an additional source of uncertainty (Serinaldi and Kilsby, 2014). Therefore non-stationary models adoption does not guarantee any practical enhancement of the accuracy of extreme rainfall analysis, whereas possible misspecification of the model would lead to seroius under /overestimation of the predicted quantiles. The choice of the model have to be carefully evalutated, performing a preliminary analysis of the records at hand.

7.2 The Padova time series

The Padova dataset of daily precipitation and temperature is the longest record of its kind. It covers the time span from 1725 to 2013 with only a few year of incomplete data. The data have been recorded across the years at three different stations, all of them located within 1 Km. Camuffo (1984) identified five different periods:

1. 1725-1768 Giovanni Poleni collected the data on the roof of his own house (10m above ground), using a raingauge constructed according to the indications of the Royal Society;

2. 1768-1813 Giuseppe Toaldo and later Vincenzo Chiminello kept measuring daily rainfall using a raingauge located on top of the Specola tower, Padova astronomical observatory. The height of the instrument is approximately 25m above ground.

3. 1814-1877 Giovanni Santini is the new director of the astronomical observatory. In this period the observation are not always systematic; The data recorded in the years from 1815 to 1823 were exclude for their sparsity and uncertainty. Starting in 1838 the entire roof of the Specola ($27.5m^2$) is used

as a funnel; this may have lead to underestimation of the smaller rainfall events. Three years (1838-1840) are missing from this period.

4. 1878-1934 Giuseppe Lorenzoni installed a new raingauge with an area of $0.4m^2$ located at 21m above the ground. The subsequent directors of the Specole kept measuring with the same device to end in 1934

5. 1878-1934 Measurement are run by the Venice Water Authority, who arranged a new station located about 800m away from the Specola, with a increased sampling frequency (measuraments are now hourly).

6. From 1994-present the Veneto Region Environmental Agency (ARPAV) is appointed to carry on daily measurements. The reference station is now located at Padova Botanical Gardens, $1Km$ away from La Specola.

The Padova series of rainfall records consists of 275 years of complete daily observations; only 14 years are missing from the record. The data recorded in the different epochs can be considered to have spatial homogeneity since all the observation were collected inside a $1km$ radius, a distance significanlty smaller than the characteristic spatial scale of precipitation evets. On the contrary, some concern may regard inhomogeneities due to changes in the measuring instrument across the years. In particular the major change occurred in the years from 1838 to 1877 when the whole roof of the Specola tower was used as a funnel to collect the precipitated volumes.This may have caused underestimation of the smaller events and even of the number of rainy days, its sensibility being certainly less than the one of the other raingauges. Moreover, it is likely that a fraction of the precipitated water was retained by the roof surface anc lost for evaporation. The more intense events likely are only slightly affected from this instrumental inhomogeneity, whereas the smaller ones (and in particular the yearly recorded number of rainy events) might not be properly described. A recent analysis of the properties of the Padova time series (Marani and Zanetti, 2014) explored the fluctations observed in the GEV-estimated rainfall extremes and in the number of wet days. They found high amplitude cycles in the GEV-estimated quantiles for return times of $T_r = 100$ and $T_r = 200$ years. In panel (7.3) scatter plots are reported for the variable w and C (Weibull shape and scale parameters) and N (yearly number of rainy days) for three different historical periods (1725-1814, 1823-1877, 1878-2013) corresponding to different measuring instrument. Shape and scale Weibull parameter show a positive correlation between them and a negative correlation with the yearly number of rainy days. The temporal variations can be explined both with climatic variability and instrumental change. For exaple, in the interval 1823-1877 a drop can be observed in the number of rainy days and an increment in the mean value of the scale and shape parameter. This could be explained by considering that a different measuring device was used, which led to a possible underestimation of the yearly number of rainy days and, as a consequence, to an overestimated mean scale parameter. This would also explain the inverse relation between scale parameter and number of rainy days: If the smaller rainfall amounts are not reported, when the number N of rainy days is smaller, then the scale parameter (which corresponds with mean observed value) grows.

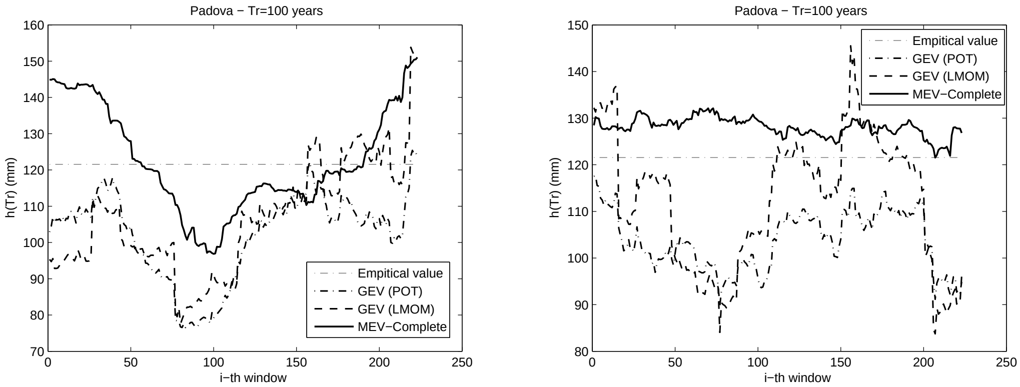


Figure 7.1: 50-years sliding windows analysis for the original (right) and reshuffled (left) Padova dataset, for a given return time $T_r = 100$ years.

7.3 Sliding windows analysis

In this section an original dataset is analyzed using sliding and overlapping windows. The length (e.g. T=30 and T=50 years) is fixed and for every period the magnitude of the event for a given return time is computed using GEV, POT and MEV approaches. The estimated quantiles are then compared with the empirical non exceedance frequencies, computed over all the years of observation. This way, it is possible to evalutate the meaningfulness of hypotesis of stationarity in EV analysis. Fig. 7.1 shows the 50-years sliding windows result for the Padova series; the quantiles associated with the $T_r = 100$ years were compute both for the original Padova series and for a datset obtained by randomly reshuffling the original daily records using the methodology explined in Chapter 6. The analysis of the original series shows remarkable variations in the estimated quantiles across the last 3 centuries, suggesting that the adoption of stationary EV models is not justified. On the other hand, the same analysis performed on the reshuffled series shows that GEV and POT estimated quantiles still are characterized by large oscillations, whereas MEV estimated values exhibit a remarkably smaller variability. This finding shows that the oscillations in the GEV-estimated quantiles are not generated by actual non stationarities of the original record, but by the high variance inborn in GEV estimations. On the contrary, MEV variability disappears when the original temporal sequence is destroyed, indicating that the oscillations observed in the MEV-estimated values reflects actual non stationarities (correspondent to low frequency oscillations) in both the yearly number of wet days and in the parameters of the daily dainfall distribution. The same can be observed for the Bologna dataset in Fig. 7.2: GEV and POT estimated quantiles are subject to random oscillations, whereas the extremes evalutated with the MEV approach are fairly stable in the randomly reshuffled series (right figure) and show a clear upward trend in the original series (left figure) due to trends in either the cardinality N or the Weibull parameters across the last two centuries.

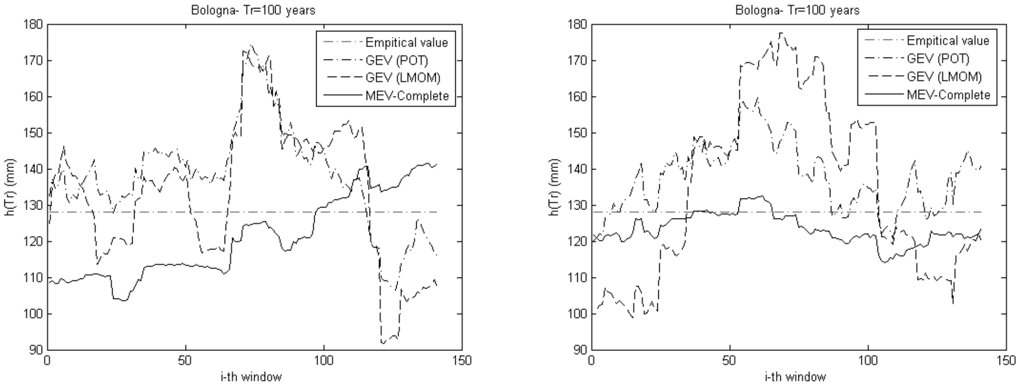


Figure 7.2: 50-years sliding windows analysis for the original (right) and reshuffled (left) Bologna dataset, for a given return time $T_r = 100$ years.

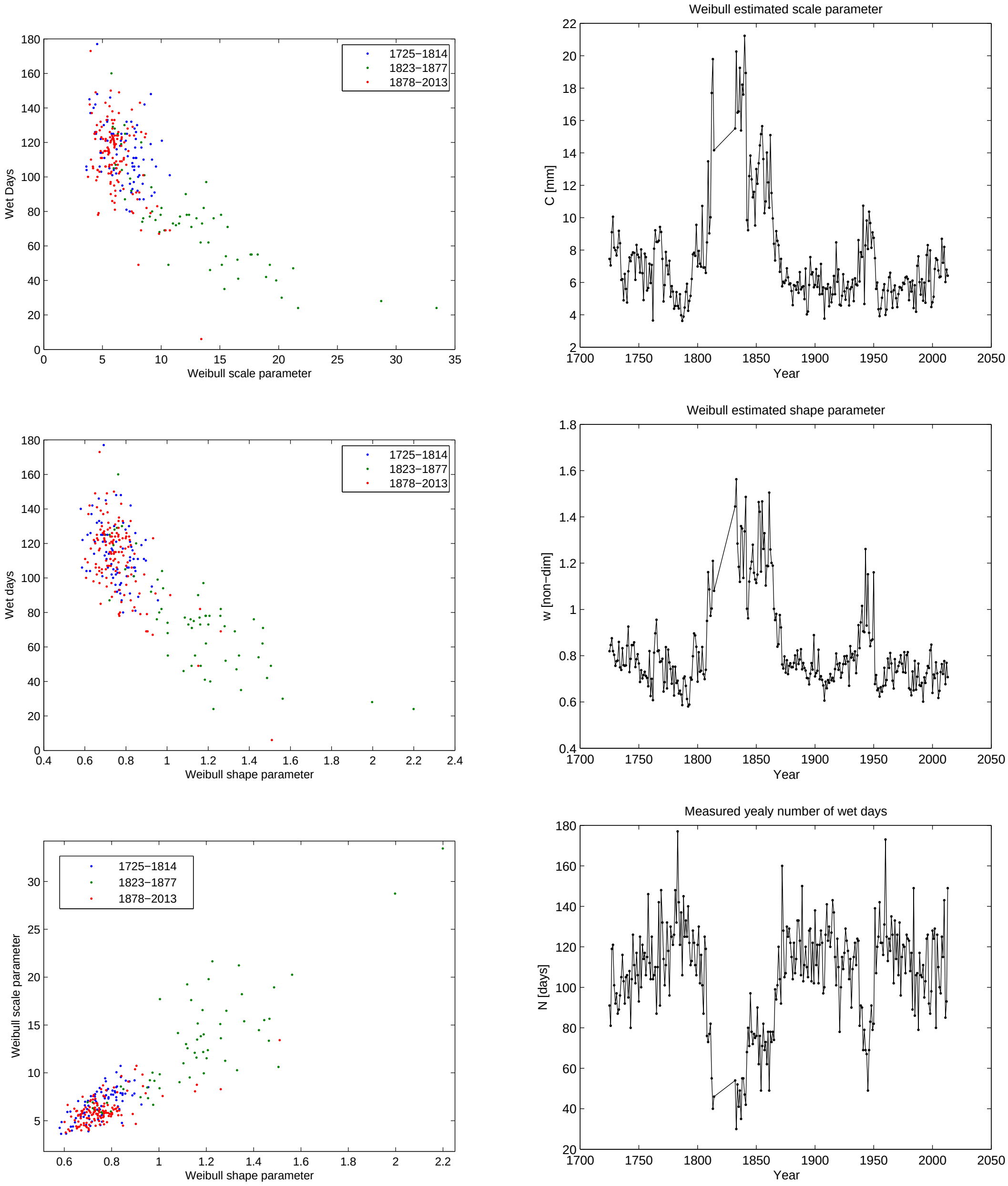


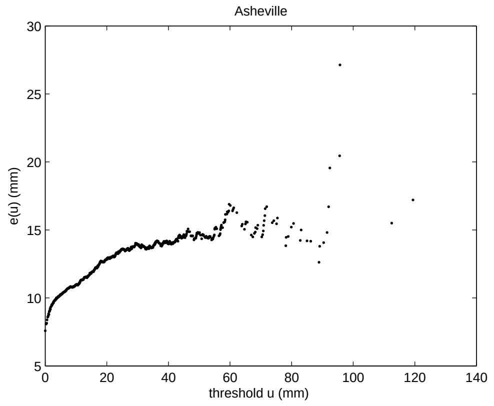
Figure 7.3: Scatter plots for shape w and scale C Weibull parameters (computed fitting Weibull to the single years of all the Padova record) and yearly number of wet days N (right figures). Distribution of w, C and N over time (on the left).

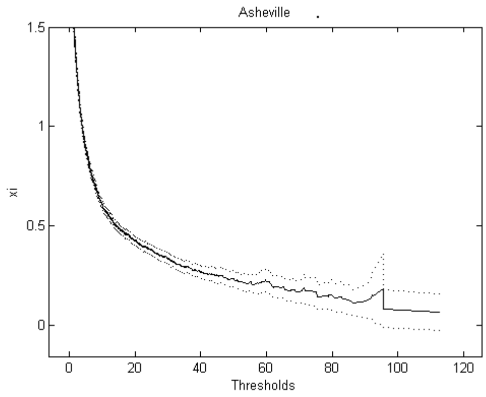
Conclusions

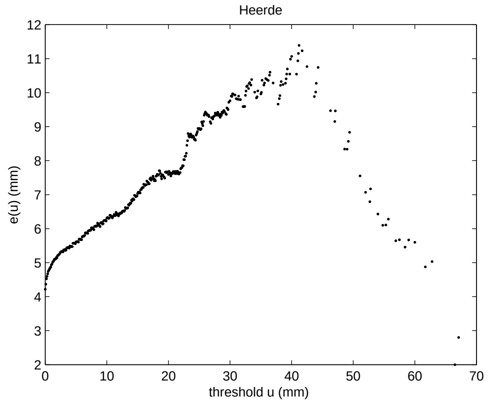
The present dissertation derives theoretically and applies the MEV compound distribution by comparing its performances with the traditional EV approaches, based on the fit of an asymptotic distribution (GEV) using Annual Maxima (AM) or Peak Over Threshold (POT) methods. A wide set of long observational series from stations situated in different locations was selected, spanning disparate climatic conditions in order to bestow general validity to the results of the analyses. The MEV distribution is derived removing the asymptotic hypothesis, on which classical EV theory is based, and accounting for stochastic variability in the yearly number of events. Hence, its practical application does not require a sufficiently large number of events per year to take place, as the classical GEV approach does. This is a conceptual advantage, since it has been shown *(Kudsoyiannis, 2004)* that GEV is a good approximation of the actual annual maximum cdf only for extremely large values of N. The use of MEV as a distribution for the annual maxima requires the adoption of a parent distribution for the daily rainfall values, whose parameters are considered random variables as well; in this study the two-parameters Weibull distribution was adopted and the corresponding parameters were estimated with Maximum Likelihood (ML) techniques. The stationary extreme values analysis was carried out by means of artificial datasets obtained from observed records by reshuffling the daily rainfall values, so as to eliminate serial correlation and to preserve their actual (unknown) distribution. In the synthetic data the distribution of the cardinality N(yearly number of rainy days) was preserved. For every rainfall record, the study was repeated for different return times and sample sizes, in order to evalutate the behaviour of the different EV distributions under a broad range of conditions. The analysis shows that the adoption of the MEV distribution allows a more accurate quantile estimation when extrapolation is required outside the range of data avilability. In fact, often the records available for practical applications have a limited length (30-50 years) and traditional methods used to esimate high quantiles typically provide extremely uncertain results. By repeating the analyses for 1000 successive reshuffling of all the datasets, the distribution of the error has been evalutated for each method. For the higher return times, MEV-estimated quantiles are characterized by a small RMSE due to smaller standard deviations of the error. On the contrary, GEV-AM and POT methods produce estimated quantiles whose standard deviation increase cosistenly with T_r. The advantage inborn in the MEV approach is that it considers not only the tail, but also the whole bulk of the distribution of daily rainfall for the definition of the cdf of the annual maximum. The GEV-AM is a good model for the sample of annual maxima on which it has been calibrated, but it may fail in describing properly the underlying process from which the annual maxima have been sampled, thereby leading to uncertain and misleading estimates when extraplations are required. In fact, GEV ensures a proper modeling of the "local"' statistical properties of a given samples, but fails when required to "'generalize"' and to capture the underlying statistical properties of the population from which the sample is drawn. The MEV approach overcomes this limitation of the traditional models: by using information from the bulk of the daily rainfall distribution, MEV is able to infer the general characters of the underlying stochastic process. As a consequence, MEV-estimated quantiles are not overly sensible neither to the specific sample nor to the sample size. The MEV approach takes into account the inhomogeneous behaviour of the daily rainfall, incorporating the inter-annual variability of the parameters of the Weibull distribution. The performances of the model showed a variability for different datasets, suggesting that the Weibull distribution may not be a model of general validity for daily rainfall dephts. In the stations where the Weibull goodness of fit is poor, a bias may enter in the MEV estimated quantiles. Nevertheless, for all the considered stations the bias remains a small fraction of the estimated quantile. The formulation of MEV explored in this dissertation is speculated to be the proper way to model extremes under a changing climate, as it properly accounts for the interannual variablity of the parameters of the daily rainfall, from which maxima are generated. In a more general framework, the annual set of parameter could be considered dependent on a set of climatic covariates with potential implications for the estimate of extremes and their temporal trajectories. The aforementioned results reveal that the MEV distribution surpasses the traditional GEV for distinct reasons. Firstly, it allows more accurate estimations of high quantiles, granting a reduction in the variance of the estimated values and thus diminishing the uncertainty in risk appraisals. Secondly, MEV is a more general tool than GEV and POT models, since it does not require the hypotheses of asymptoticity and Poissonian distribution of excesses on which GEV and POT lean on. Thirdly, MEV allows a superior description of the physics of the rainfall process and, as a consequence, is a more natural way to model nonstationary extremes. In conclusion, MEV approach constitutes a simple and realiable EV distribution which can be an alternative to the traditional GEV methods.

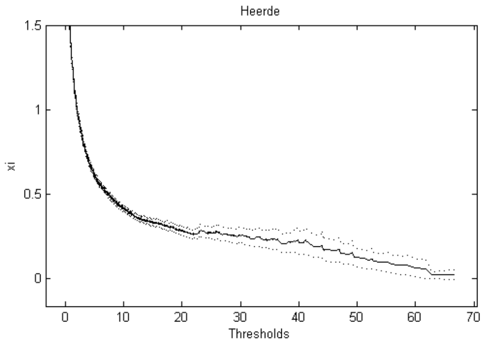
Appendix A

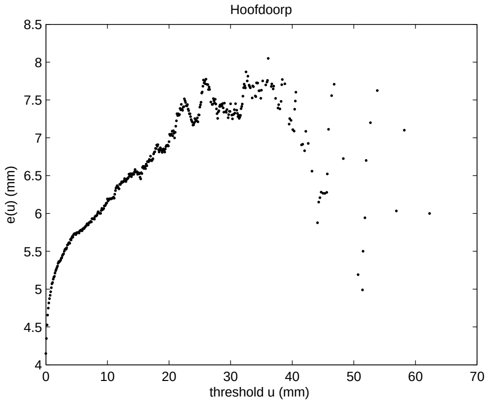
In this appendix Mean Excess Function (MEF) plots and Hill plots for all the stations are reported. The reader is referred to Chapter 4 for theoretical derivation and applications of the graphs for POT threshold selection and GPD shape parameter estimation. Hill plots shows Hill-estimated tail index, equivalent to the GEV-GPD shape parameter (solid line) and 95% confidence intervals (dashed lines).

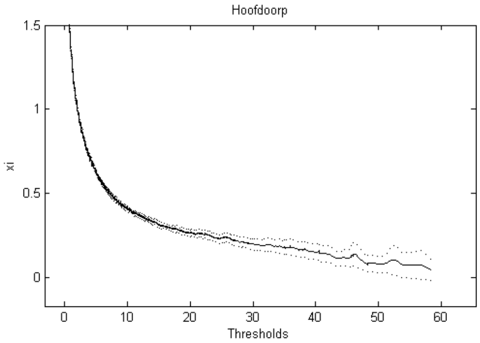


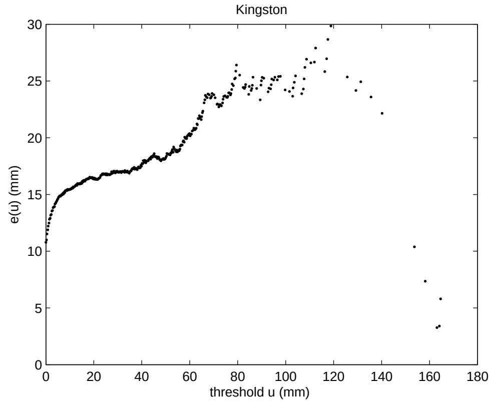


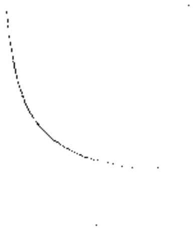


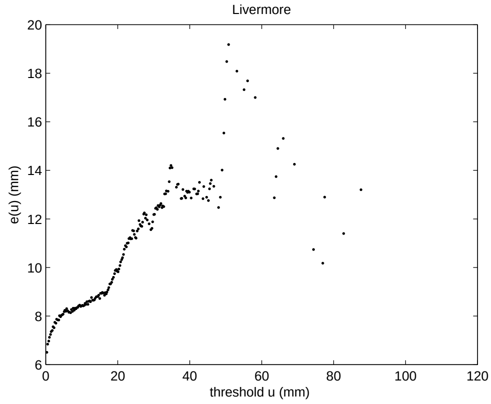


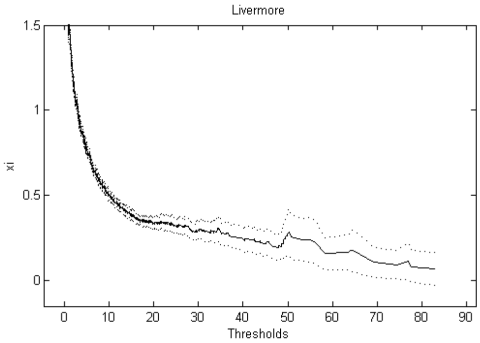

Hoofdoorp
e(u) (mm)
threshold u (mm)
4
4.5
5
5.5
6
6.5
7
7.5
8
8.5
0
10
20
30
40
50
60
70

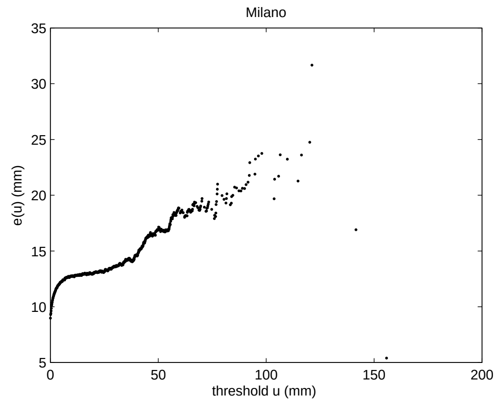

Hoofdoorp
xi
Thresholds
0
0.5
1
1.5
0
10
20
30
40
50
60

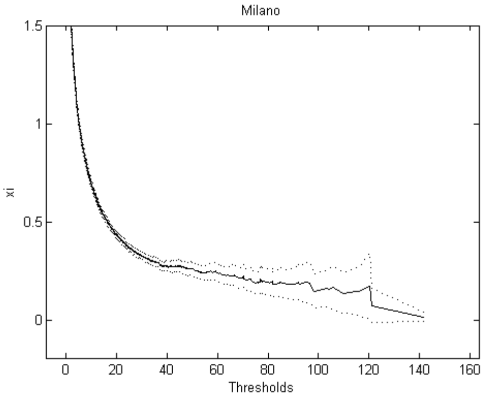

Kingston
e(u) (mm)
threshold u (mm)
0
5
10
15
20
25
30
0
20
40
60
80
100
120
140
160
180

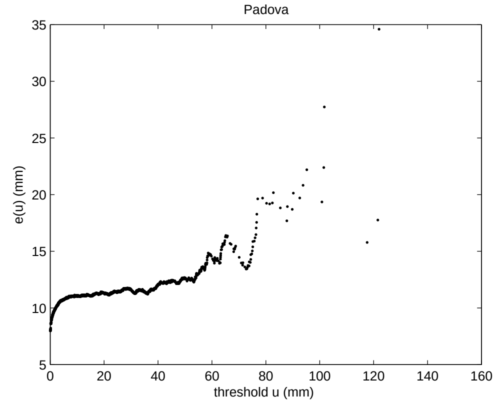


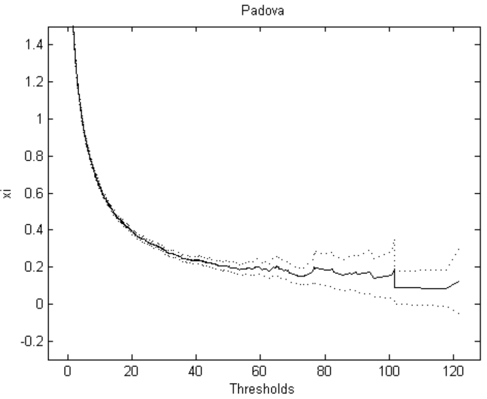

Livermore
e(u) (mm)
threshold u (mm)
6
8
10
12
14
16
18
20
0
20
40
60
80
100
120

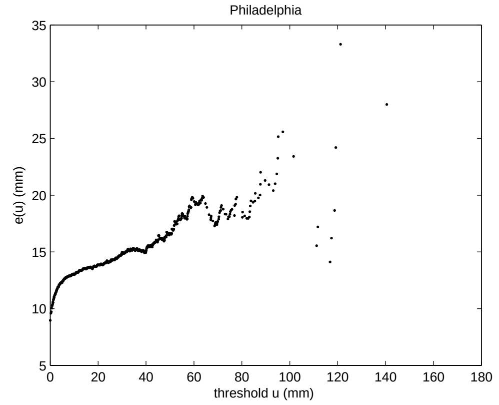

Livermore
xi
Thresholds
0
0.5
1
1.5
0
10
20
30
40
50
60
70
80
90

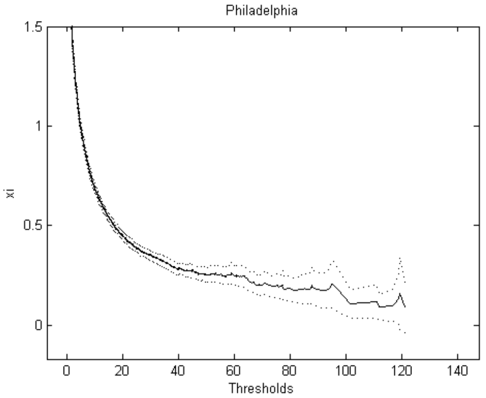

Milano
e(u) (mm)
threshold u (mm)
35
30
25
20
15
10
5
0
50
100
150
200

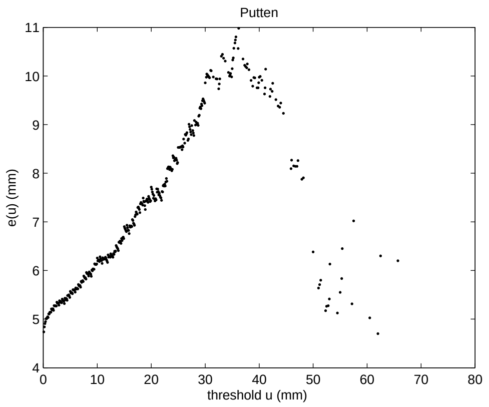

Milano
xi
Thresholds
1.5
1
0.5
0
0
20
40
60
80
100
120
140
160

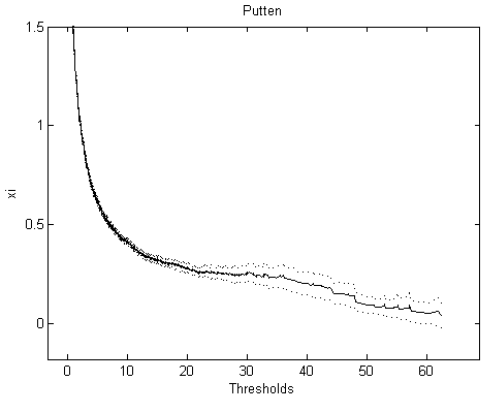

Padova
e(u) (mm)
threshold u (mm)
35
30
25
20
15
10
5
0
20
40
60
80
100
120
140
160

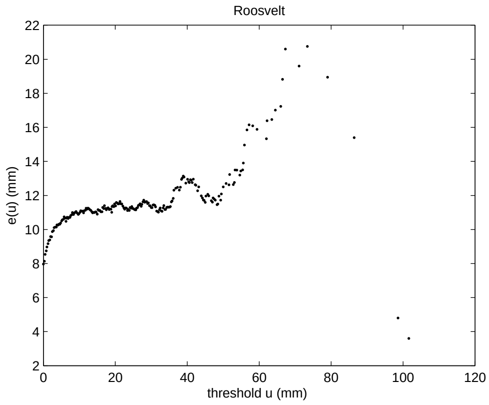

Padova
xi
Thresholds
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120

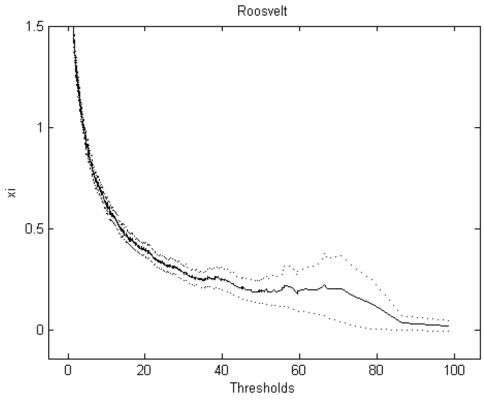

Philadelphia
e(u) (mm)
threshold u (mm)
35
30
25
20
15
10
5
0
20
40
60
80
100
120
140
160
180

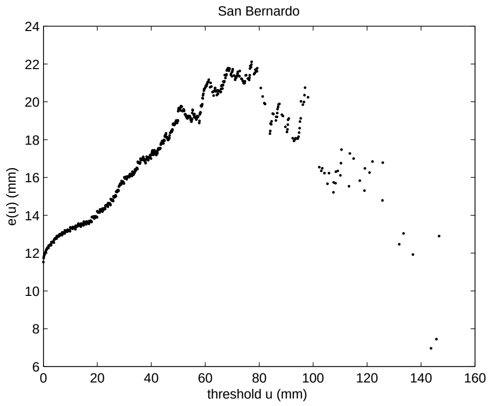

Philadelphia
xi
Thresholds
1.5
1
0.5
0
0
20
40
60
80
100
120
140

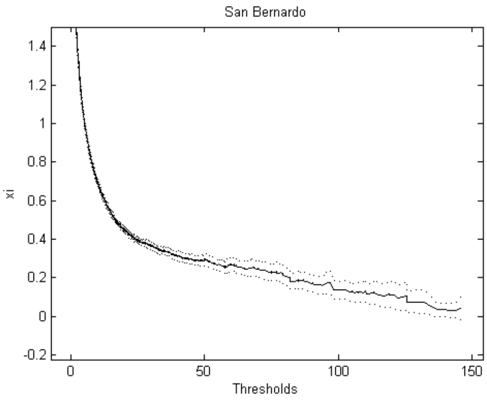

Putten
e(u) (mm)
4
5
6
7
8
9
10
11
0
10
20
30
40
50
60
70
80
threshold u (mm)

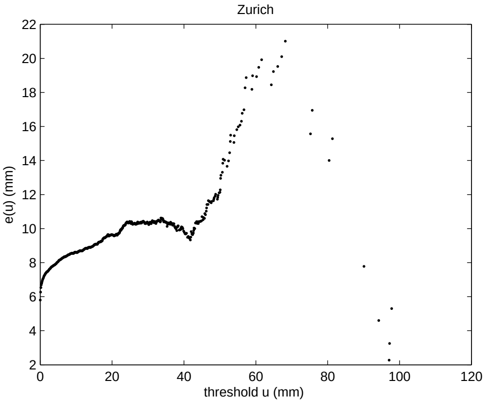

Putten
xi
0
0.5
1
1.5
0
10
20
30
40
50
60
Thresholds

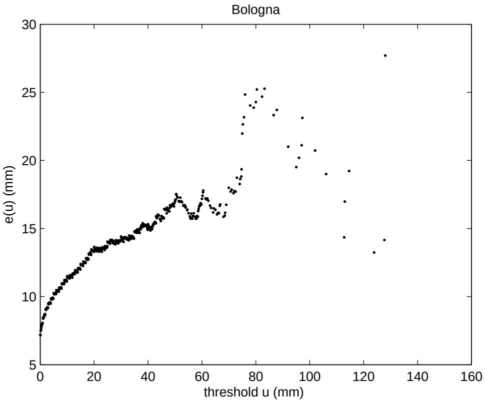

Roosvelt
e(u) (mm)
2
4
6
8
10
12
14
16
18
20
22
0
20
40
60
80
100
120
threshold u (mm)

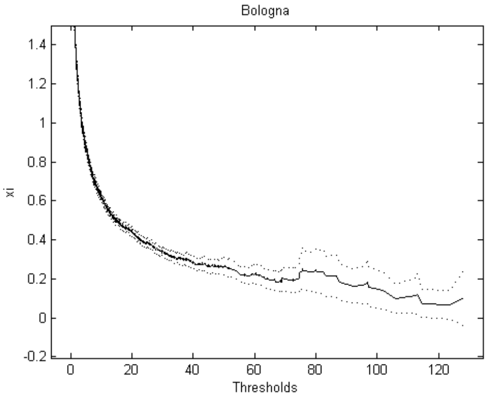

Roosvelt
xi
0
0.5
1
1.5
0
20
40
60
80
100
Thresholds

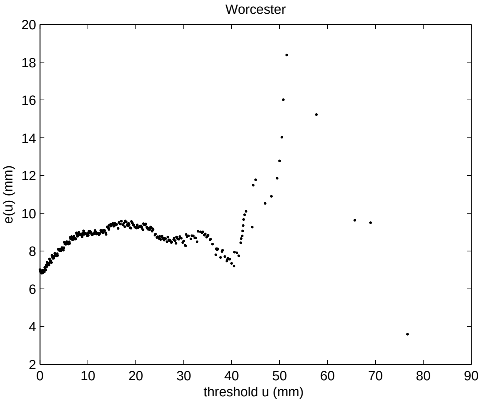

San Bernardo
e(u) (mm)
6
8
10
12
14
16
18
20
22
24
0
20
40
60
80
100
120
140
160
threshold u (mm)

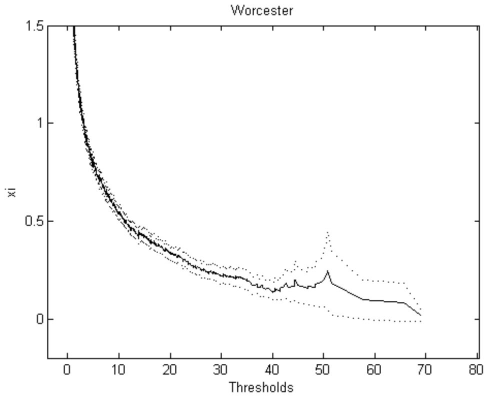

San Bernardo
xi
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0
50
100
150
Thresholds


Zurich
e(u) (mm)
threshold u (mm)
2
4
6
8
10
12
14
16
18
20
22
0
20
40
60
80
100
120


Bologna
e(u) (mm)
threshold u (mm)
5
10
15
20
25
30
0
20
40
60
80
100
120
140
160

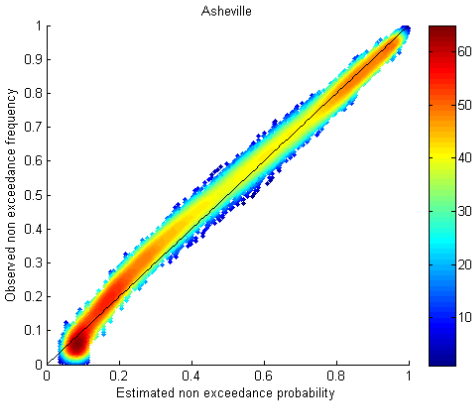

Bologna
xi
Thresholds
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0
20
40
60
80
100
120

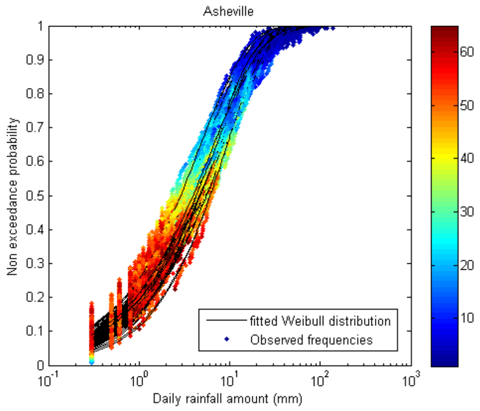

Worcester
e(u) (mm)
threshold u (mm)
2
4
6
8
10
12
14
16
18
20
0
10
20
30
40
50
60
70
80
90

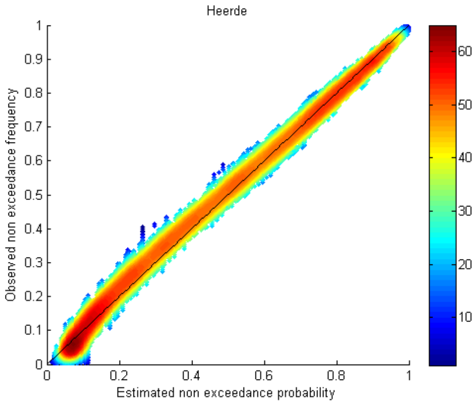

Worcester
xi
Thresholds
0
0.5
1
1.5
0
10
20
30
40
50
60
70
80

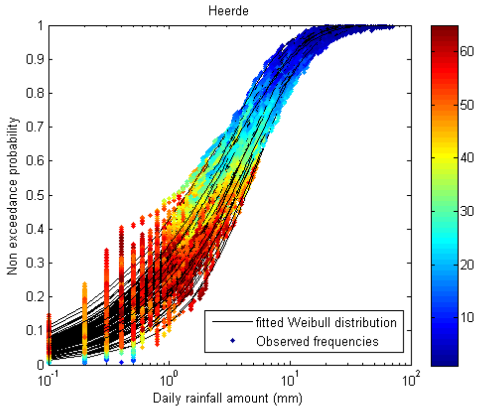
Appendix B

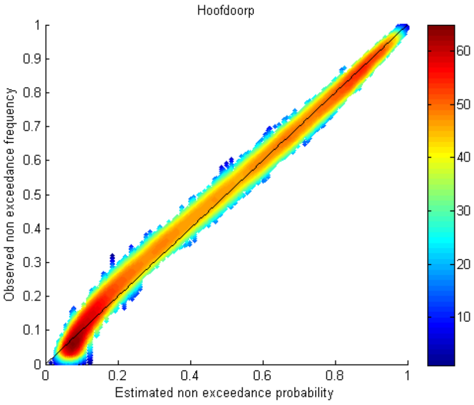
In this appendix are reported QQ plot obtained fitting Weibull distribution to the single years for all the datasets. Plots representing the observed empirical cdfs and the Weibull theoretical ones are also reported.

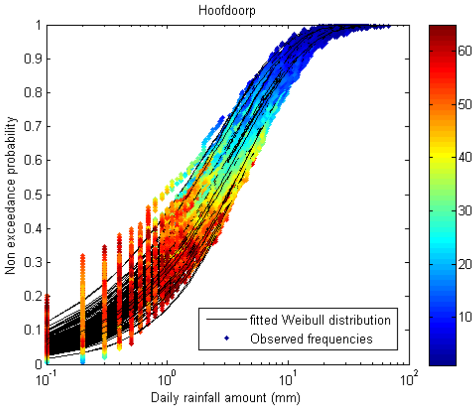


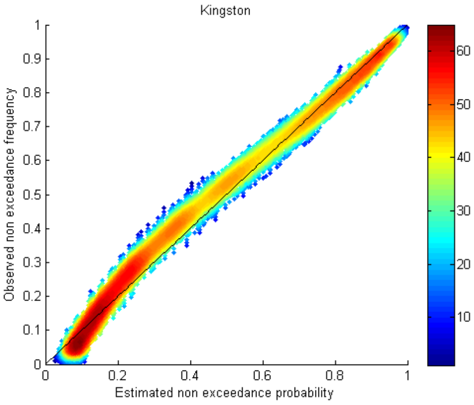


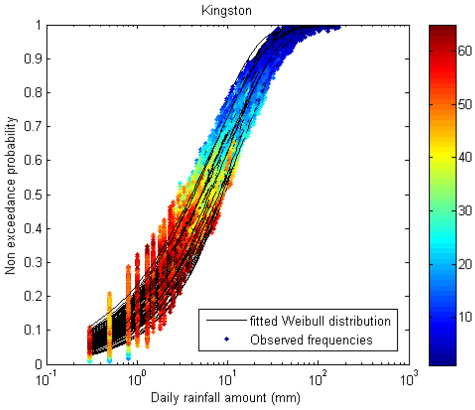


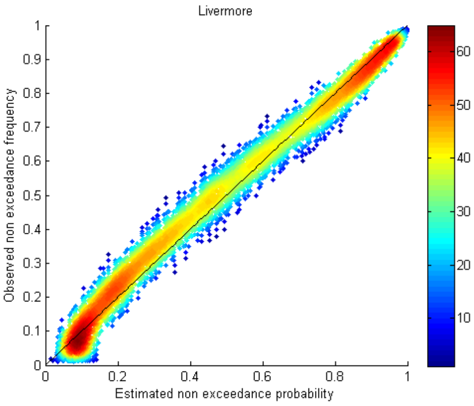


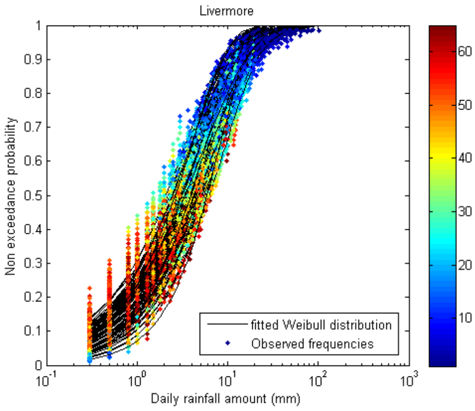

Hoofdoorp
Observed non exceedance frequency
Estimated non exceedance probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
10
20
30
40
50
60

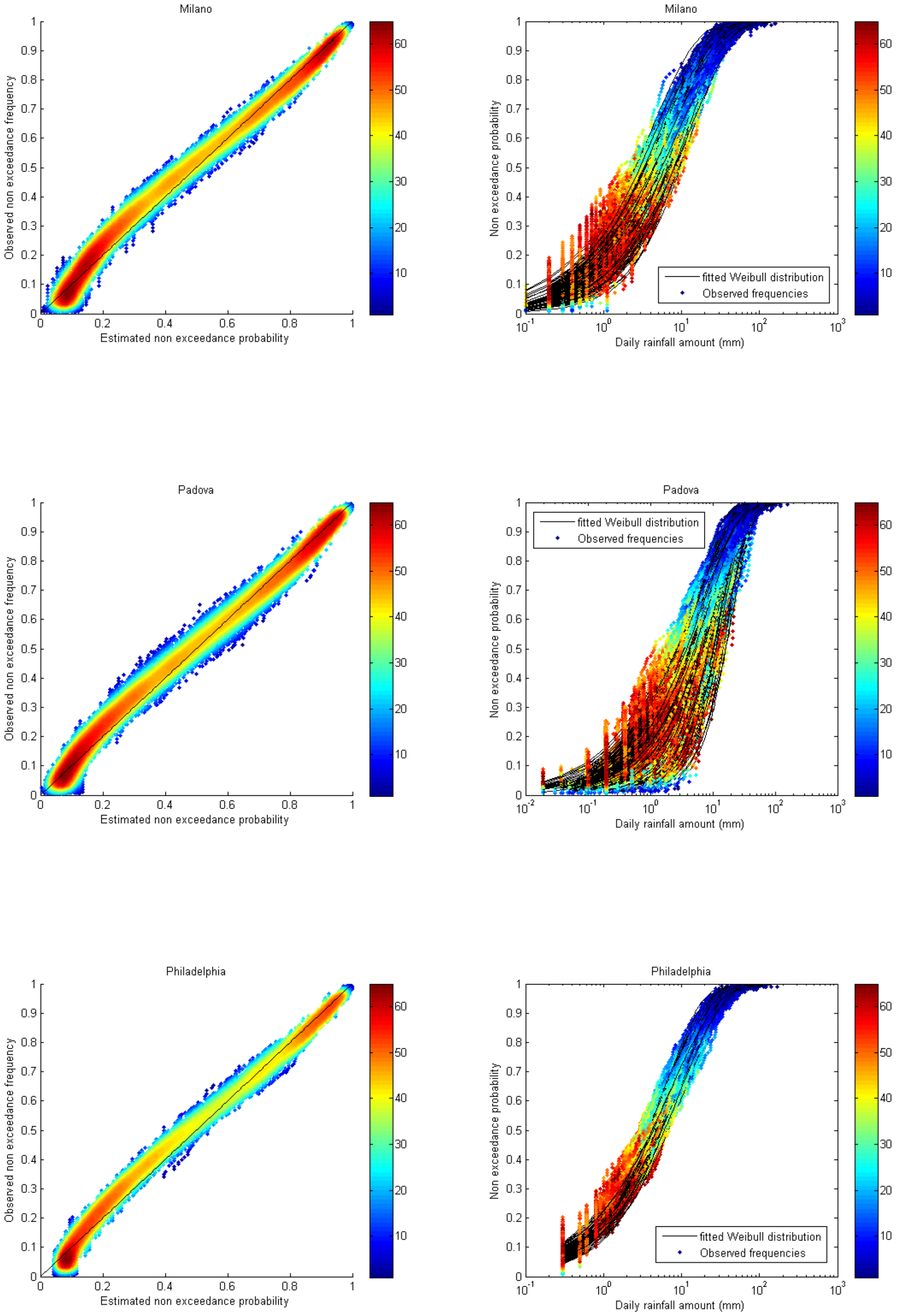

Hoofdoorp
Non exceedance probability
Daily rainfall amount (mm)
10^{-1}
10^{0}
10^{1}
10^{2}
fitted Weibull distribution
Observed frequencies
10
20
30
40
50
60

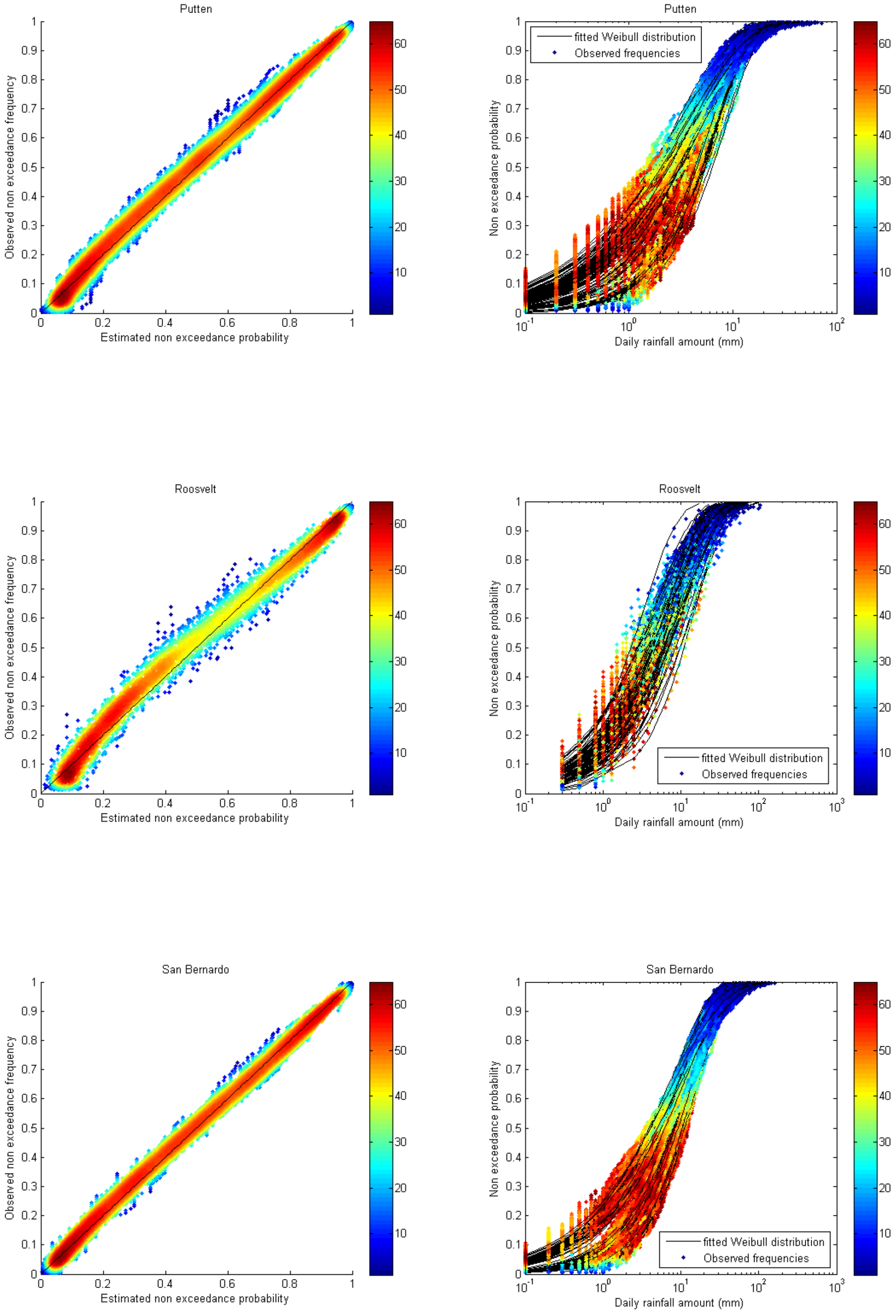

Kingston
Observed non exceedance frequency
Estimated non exceedance probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
10
20
30
40
50
60

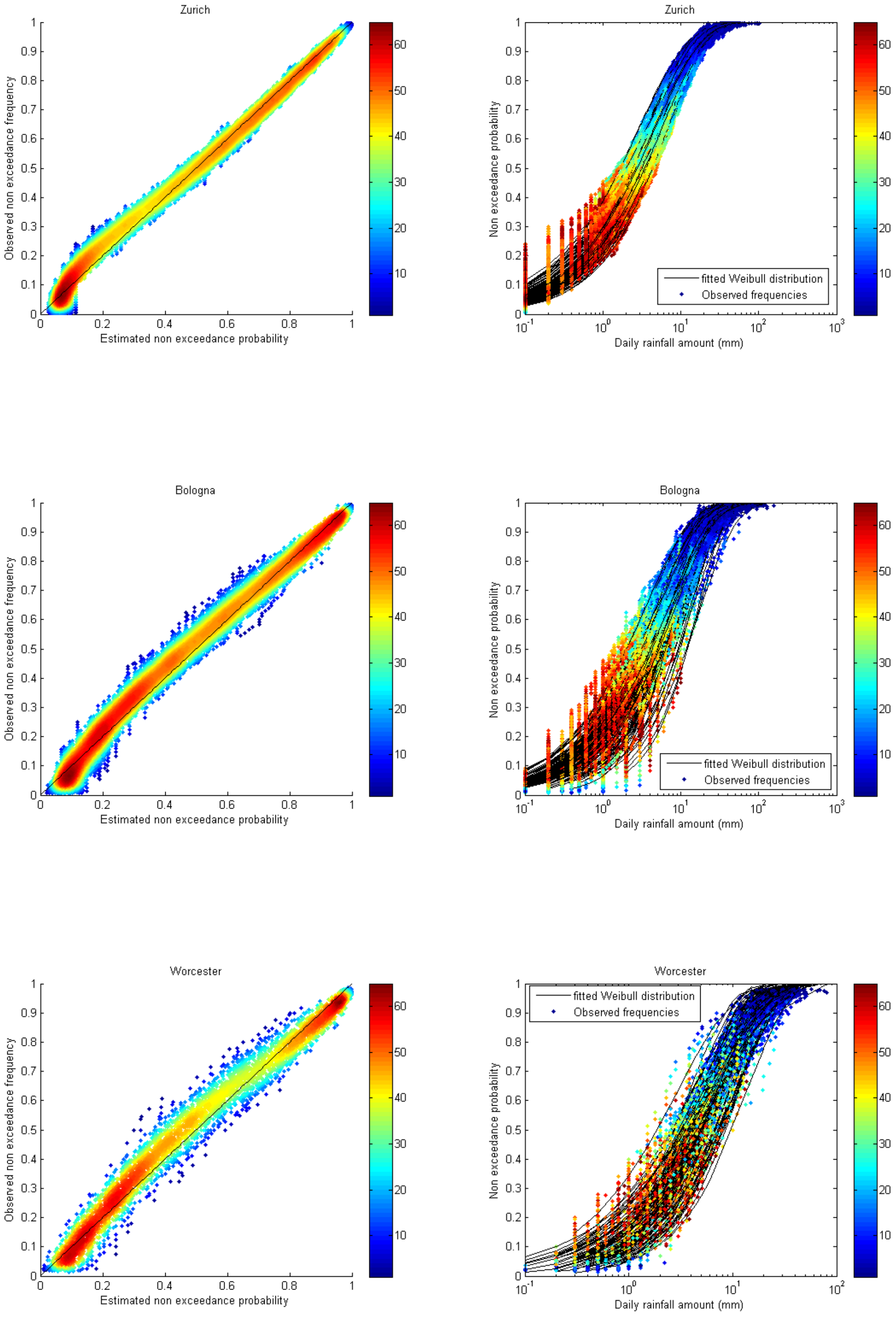

Kingston
Non exceedance probability
Daily rainfall amount (mm)
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
fitted Weibull distribution
Observed frequencies
10
20
30
40
50
60


Livermore
Observed non exceedance frequency
Estimated non exceedance probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
10
20
30
40
50
60


Livermore
Non exceedance probability
Daily rainfall amount (mm)
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
fitted Weibull distribution
Observed frequencies
10
20
30
40
50
60

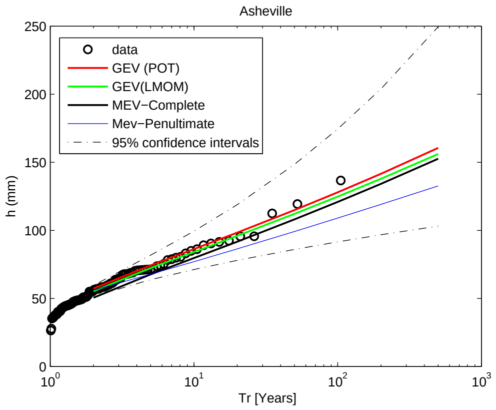

Milano
Observed non exceedance frequency
Estimated non exceedance probability
Milano
Non exceedance probability
Daily rainfall amount (mm)
fitted Weibull distribution
Observed frequencies
Padova
Observed non exceedance frequency
Estimated non exceedance probability
Padova
Non exceedance probability
Daily rainfall amount (mm)
fitted Weibull distribution
Observed frequencies
Philadelphia
Observed non exceedance frequency
Estimated non exceedance probability
Philadelphia
Non exceedance probability
Daily rainfall amount (mm)
fitted Weibull distribution
Observed frequencies
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
10
20
30
40
50
60
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}

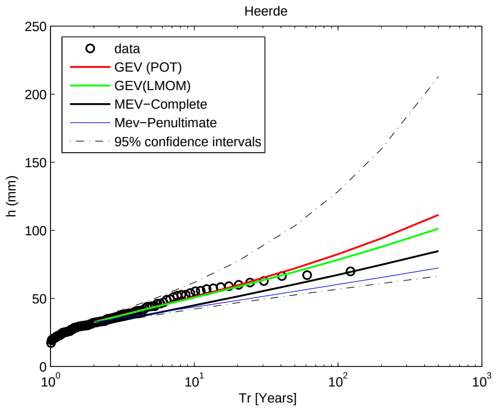

Putten
Observed non exceedance frequency
Estimated non exceedance probability
Putten
fitted Weibull distribution
Observed frequencies
Non exceedance probability
Daily rainfall amount (mm)
Roosvelt
Observed non exceedance frequency
Estimated non exceedance probability
Roosvelt
fitted Weibull distribution
Observed frequencies
Non exceedance probability
Daily rainfall amount (mm)
San Bernardo
Observed non exceedance frequency
Estimated non exceedance probability
San Bernardo
fitted Weibull distribution
Observed frequencies
Non exceedance probability
Daily rainfall amount (mm)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}

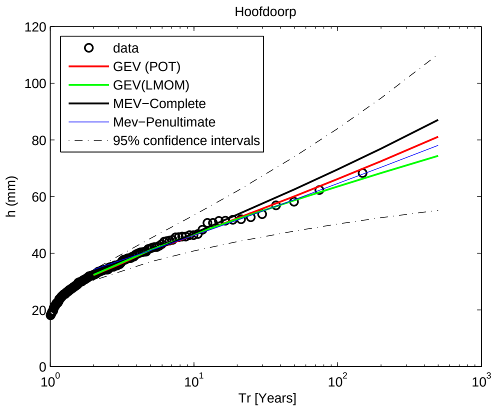

Zurich
Observed non exceedance frequency
Estimated non exceedance probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
Zurich
Non exceedance probability
Daily rainfall amount (mm)
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
fitted Weibull distribution
Observed frequencies
Bologna
Observed non exceedance frequency
Estimated non exceedance probability
Bologna
Non exceedance probability
Daily rainfall amount (mm)
fitted Weibull distribution
Observed frequencies
Worcester
Observed non exceedance frequency
Estimated non exceedance probability
Worcester
Non exceedance probability
Daily rainfall amount (mm)
10^{-1}
10^{0}
10^{1}
10^{2}
fitted Weibull distribution
Observed frequencies

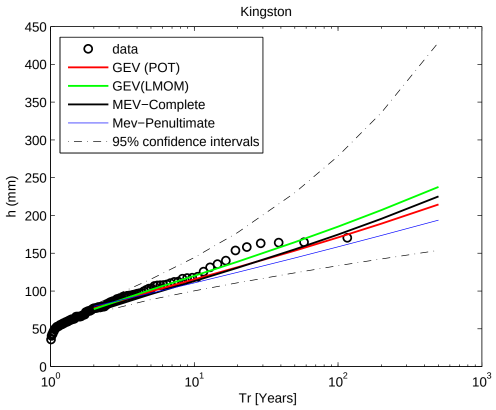
Appendix C

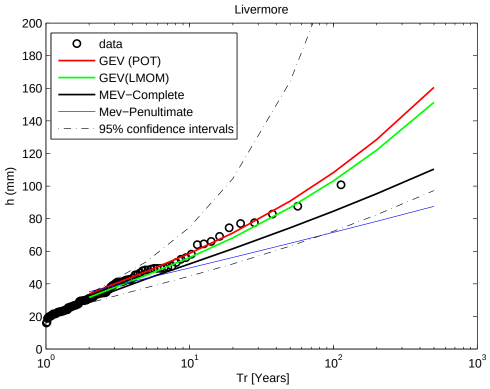
In this appendix we report the for all the selected datasets the results obtained by implementing GEV, POT and MEV approaches; the three distributions have been fitted to the whole records of all the 15 stations considered. The threshols selected for POT vary from 10mm to 25mm depending on the station. MEV Complete and MEV penultimate quantile estimation were obtained fitting Weibull distribution using ML to, respectively, single years and whole periods of records.

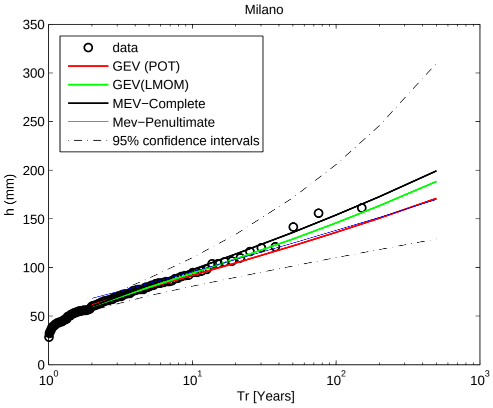


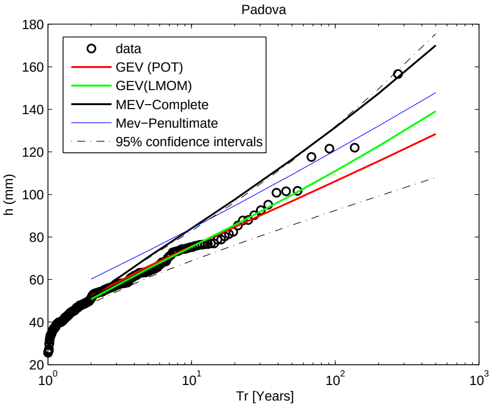


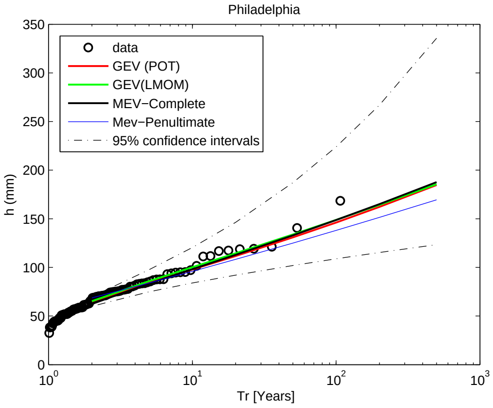


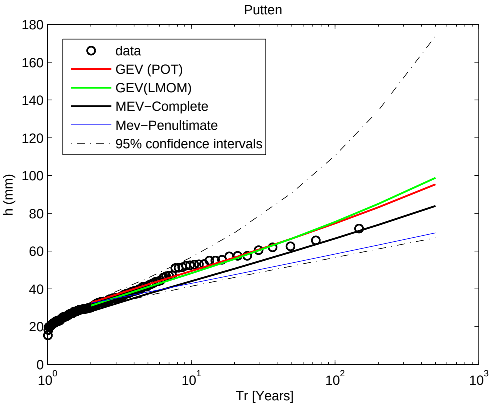


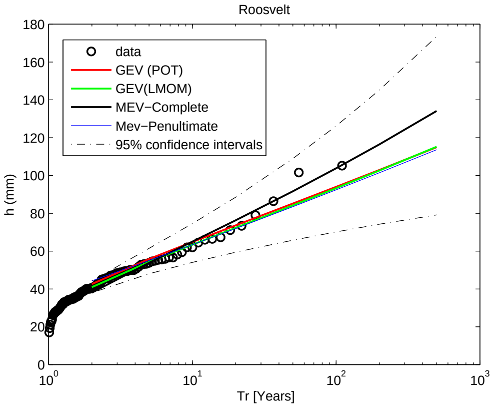

Livermore
h (mm)
Tr [Years]
0
20
40
60
80
100
120
140
160
180
200
10^0
10^1
10^2
10^3
data
GEV (POT)
GEV(LMOM)
MEV−Complete
Mev−Penultimate
95% confidence intervals

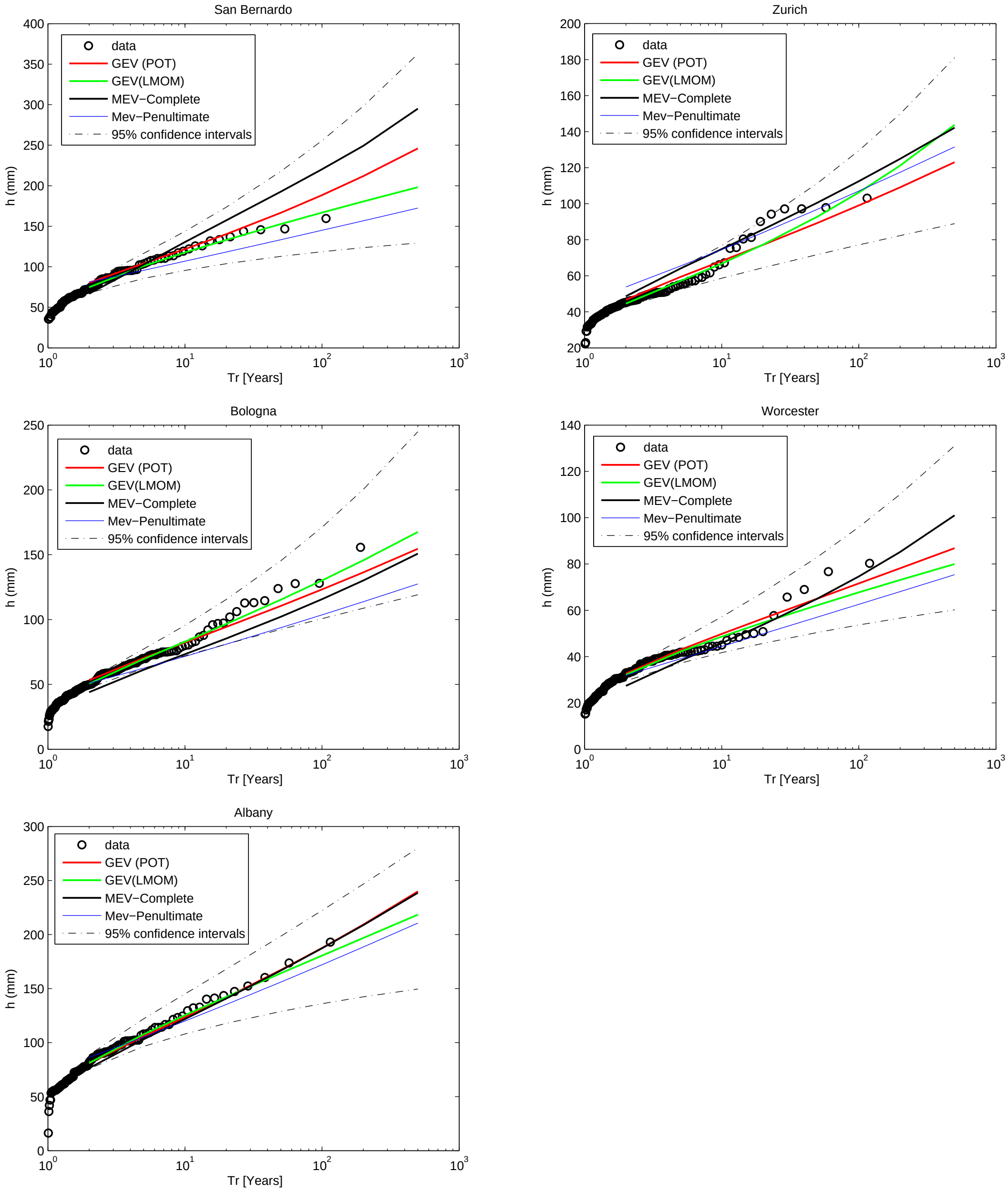

Milano
h (mm)
Tr [Years]
0
50
100
150
200
250
300
350
10^0
10^1
10^2
10^3
data
GEV (POT)
GEV(LMOM)
MEV−Complete
Mev−Penultimate
95% confidence intervals


Padova
h (mm)
Tr [Years]
20
40
60
80
100
120
140
160
180
10^0
10^1
10^2
10^3
data
GEV (POT)
GEV(LMOM)
MEV−Complete
Mev−Penultimate
95% confidence intervals


Philadelphia
h (mm)
Tr [Years]
0
50
100
150
200
250
300
350
10^0
10^1
10^2
10^3
data
GEV (POT)
GEV(LMOM)
MEV−Complete
Mev−Penultimate
95% confidence intervals

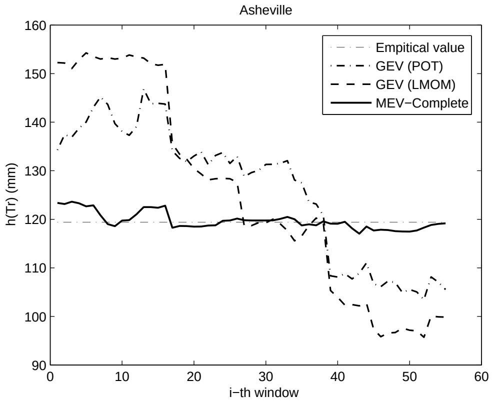

Putten
h (mm)
Tr [Years]
0
20
40
60
80
100
120
140
160
180
10^0
10^1
10^2
10^3
data
GEV (POT)
GEV(LMOM)
MEV−Complete
Mev−Penultimate
95% confidence intervals

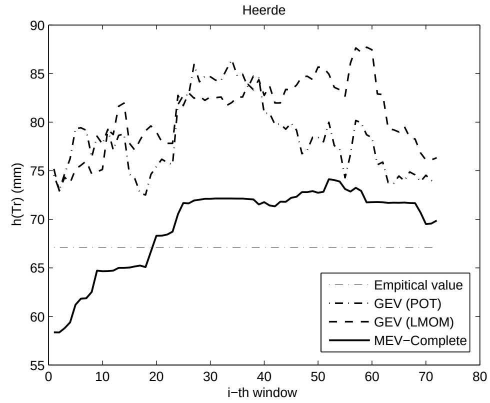

Roosvelt
h (mm)
Tr [Years]
0
20
40
60
80
100
120
140
160
180
10^0
10^1
10^2
10^3
data
GEV (POT)
GEV(LMOM)
MEV−Complete
Mev−Penultimate
95% confidence intervals

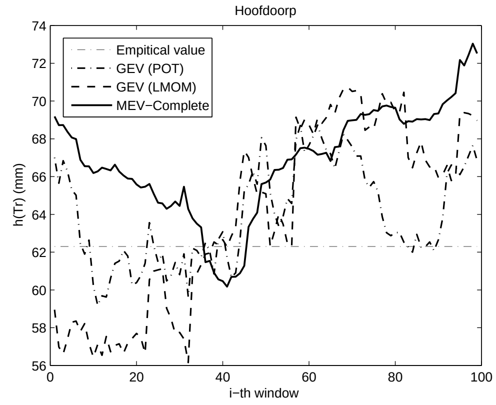

San Bernardo
Zurich
Bologna
Worcester
Albany
data
GEV (POT)
GEV(LMOM)
MEV−Complete
Mev−Penultimate
95% confidence intervals
h (mm)
Tr [Years]
10^0
10^1
10^2
10^3
0
50
100
150
200
250
300
350
400
20
40
60
80
120
140
160
180

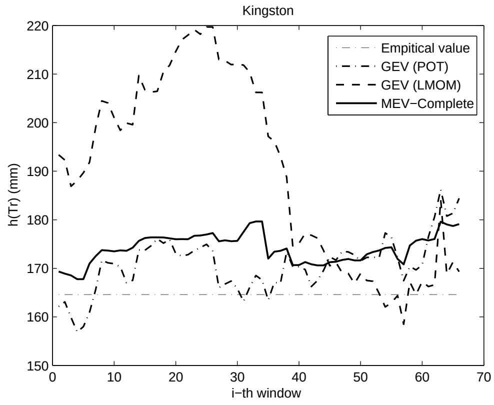
Appendix D

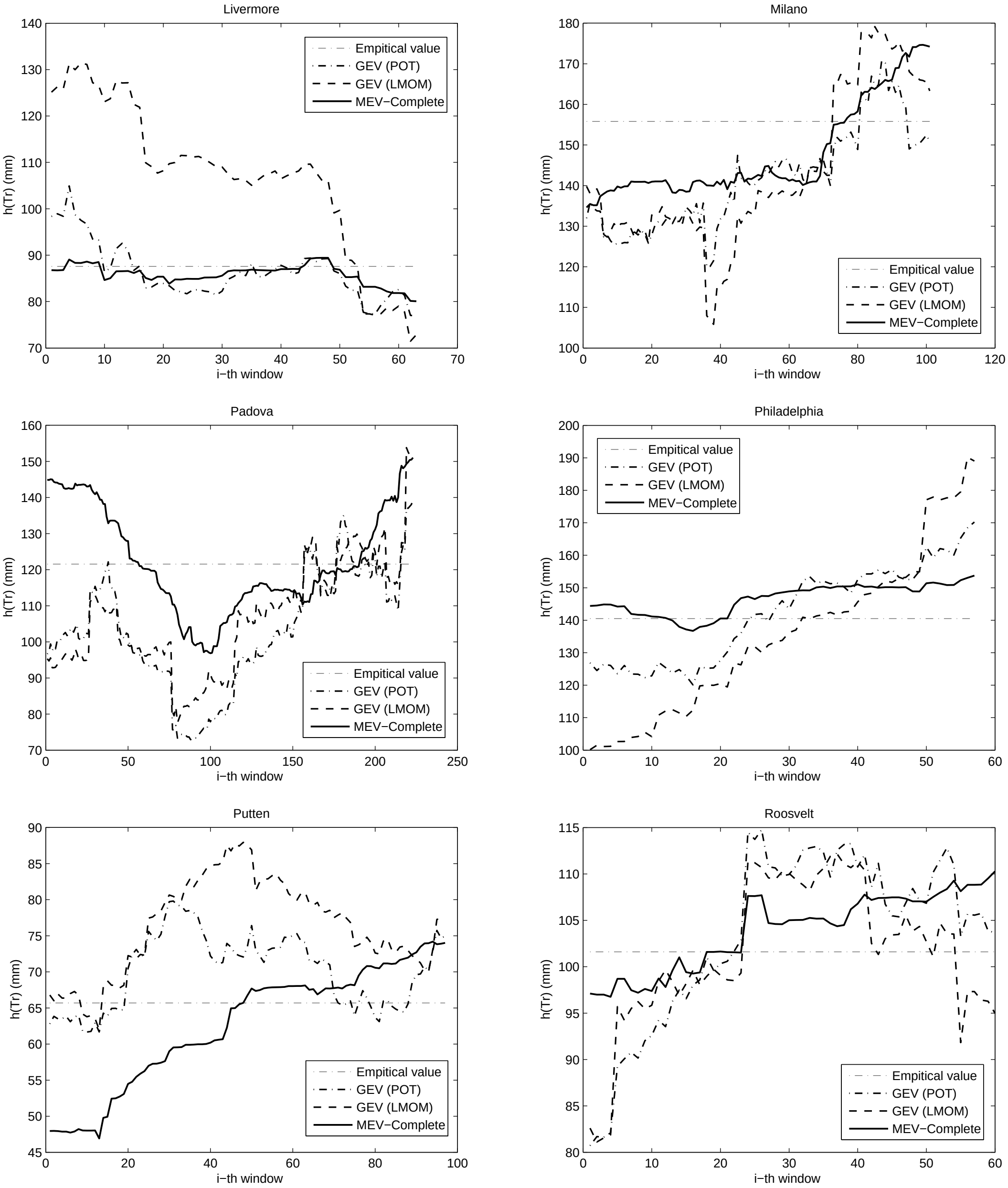
In this appendix the result obtained by implementing GEV, POT and MEV approaches to all the stations are reported. The quantile estimates were obtained using a sliding and overlapping windows approach. A window length of 50 years and a return time of 100 years have been fixed, in order to focus the analysis on the temporal variability of the estimated values.

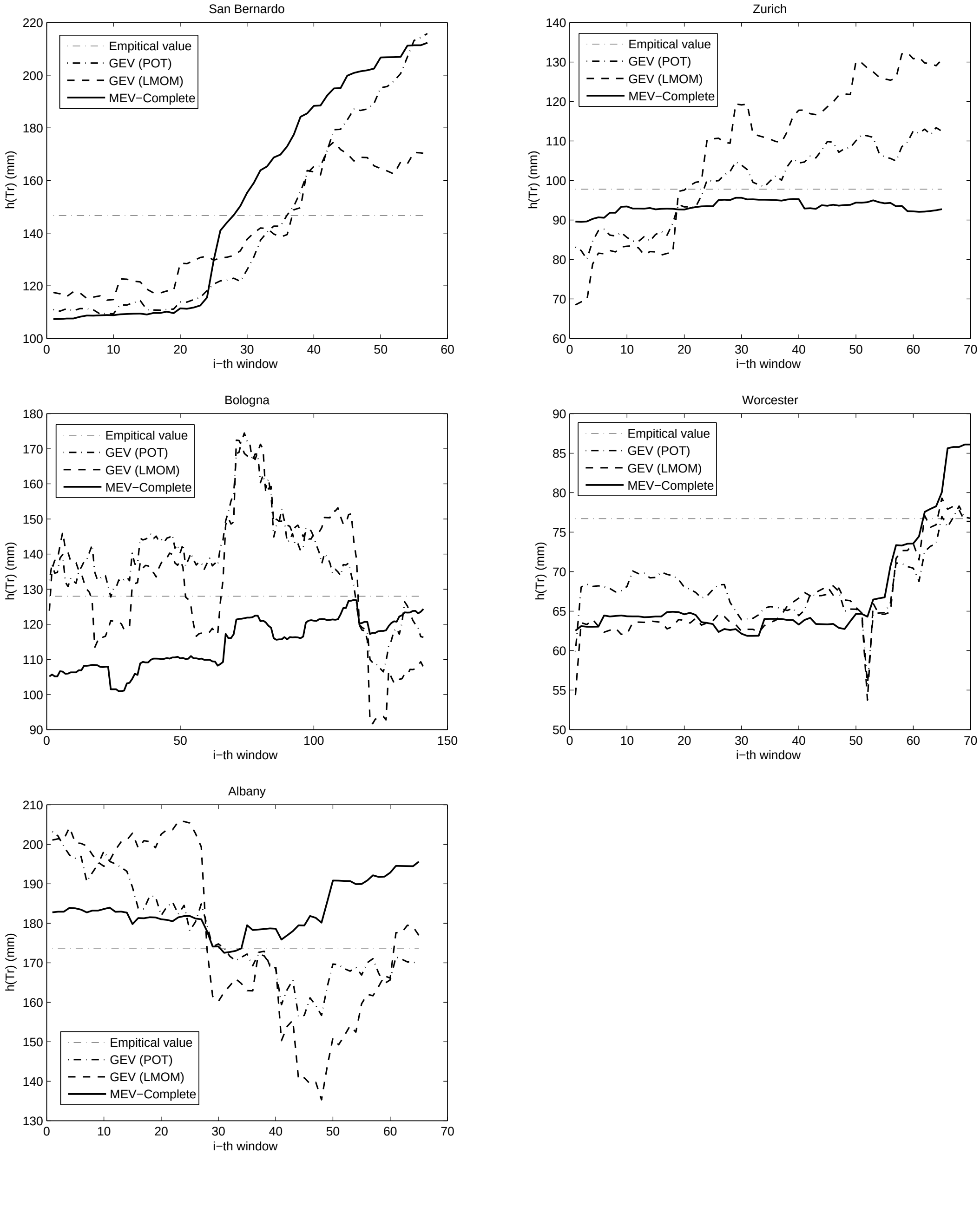









Livermore
h(Tr) (mm)
i−th window
140
130
120
110
100
90
80
70
0
10
20
30
40
50
60
70
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
Milano
h(Tr) (mm)
i−th window
180
170
160
150
140
130
120
110
100
0
20
40
60
80
100
120
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
Padova
h(Tr) (mm)
i−th window
160
150
140
130
120
110
100
90
80
70
0
50
100
150
200
250
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
Philadelphia
h(Tr) (mm)
i−th window
200
190
180
170
160
150
140
130
120
110
100
0
10
20
30
40
50
60
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
Putten
h(Tr) (mm)
i−th window
90
85
80
75
70
65
60
55
50
45
0
20
40
60
80
100
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
Roosvelt
h(Tr) (mm)
i−th window
115
110
105
100
95
90
85
80
0
10
20
30
40
50
60
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete


San Bernardo
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
h(Tr) (mm)
i−th window
100
120
140
160
180
200
220
0
10
20
30
40
50
60
Zurich
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
h(Tr) (mm)
i−th window
60
70
80
90
100
110
120
130
140
0
10
20
30
40
50
60
70
Bologna
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
h(Tr) (mm)
i−th window
90
100
110
120
130
140
150
160
170
180
0
50
100
150
Worcester
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
h(Tr) (mm)
i−th window
50
55
60
65
70
75
80
85
90
0
10
20
30
40
50
60
70
Albany
Empitical value
GEV (POT)
GEV (LMOM)
MEV−Complete
h(Tr) (mm)
i−th window
130
140
150
160
170
180
190
200
210
0
10
20
30
40
50
60
70

Bibliography

[1] Balkema, A. A., and L. D. Haan (1972), On R. Von Mises' Condition for the Domain of Attraction of exp(-e-x), *The Annals of Mathematical Statistics, 43(4), 1352-1354.*

[2] Balkema, A. A., and L. D. Haan (1974), Residual lifetime at great age, *Annals of Probability 2, 792-804.*

[3] Camuffo D., (1984), Analysis of the series of precipitation at Padova, Italy, *Climatic Change, 6, 57-77.*

[4] Camuffo D., Jones P., (2002), Improved Understanding of Past Climatic Variability from Early Daily European Instrumental Sources, *Kluwer Academic Publisher, Dordrecht.*

[5] Clauset A et al. (2007), Power law distributions in empirical data, SIAM, Vol. 51, No. 4.

[6] Coles S, (2001), An Introduction to Statistical Modeling of Extreme Values, *Springer Ser. in Stat., Springer, London.*

[7] Coles S, Pericchi R.L., Sisson S., (2002), A fully probabilistic approach to extreme rainfall modeling, *Journal of Hydrology 273 (2003): 35-50.*

[8] Cook, N.J., Harris, R. (2004), Exact and general ft1 penultimate distributions of extreme wind speeds drawn from tail-equivalent weibull parents, *Structural safety 26, 391-420.*

[9] Cramer H., (1946), Mathematical methods of statistics, *Princeton University Press, Princeton (NJ).*

[10] Cramer H., and Leadbetter M. F., (1967), Stationary and related stochastic processes, *John Wiley, New York.*

[11] De Haan L., (1971), A form of regular variation and its application to the domain of attraction of the double exponential, *Z. Wahrsch. Geb., (17), 241-258.*

[12] Embrechts P., Kluppelberg C., Mikosch T., (1996), Modelling extremal values for insurance and finance, *Springer Verlag,Berlin.*

[13] Fisher R.A., (1927), Sur la loi de probabilité de l'ecart maximum, *ann de la Soc Pol de Math 6:39-117.*

[14] Fisher R.A., Tippett L.H.C., (1928), Limiting forms of the frequency distribution of the largest or smaller member of a sample, *Math. Proc. Cambridge Philos. Soc. 24(02), 180-190.*

[15] Frisch U., and D. Sornette (1997), Extreme deviations and applications, *J. Phys. I, 7(9), 1155-1171.*

[16] Galambolos, J., (1972), On the Distribution of the Maximum of Random Variables, *The Annals of Mathematical Statistics, 43(2), 516-521.*

[17] Gencay R., Selcuk F., Ulugulyagci A., (2001), EVIM: A Software Package for Extreme Value Analysis in MATLAB, *Studies in Nonlinear Dynamics and Econometrics, 5(3) 213âĂŞ239.*

[18] Ghosh S, Resnick S., (2010), A discussion on mean excess plots, *Stochastic processes and their applications, 120, 1492-1517.*

[19] Gnedenko B., (1943), Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics 44(3) 423-453.*

[20] Gumbel E., (1954), Statistical theory of extreme values and some practical applications: a series of lectures. *Applied mathematics series; U. S. Govt. Print. Office.*

[21] Gumbel E., (1958), Statistics of extremes, *Columbia University Press, New York.*

[22] Harris, R. (2004), Extreme value analysis of epoch maxiconvergence, and choice of the asymptote *Journal of Wind Engineering and Industrial Aerodynamics 92(11): 897-918.*

[23] Hosking J.R.M. (1985), Algorithm as 215: Maximum likelihood estimation of the parameter of the generalized extreme-value distribution, *Appl. Stat., 34, 301-310.*

[24] Hosking J.R.M. and J. R. Wallis (1987), Parameter and quantile estimation for the Generalized Pareto Distribution, *Technometrics, 29, 339-349.*

[25] Hosking J.R.M. (1990), L-Moments: Analysis and estimation of distributions using linear combinations of order statistics, *J.R. Stat. Soc. Ser. B, 52(1), 105-124.*

[26] Katz R.W., Parlange M.B., Naveau P., (2002), Statistics of extremes in hydrology, *Advances in WaterResources 25, 1287-1304.*

[27] Koutsoyiannis D., (2004a), Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation *Hydrological Sciences Journal, 49(4), 575-590.*

[28] Koutsoyiannis D., (2004b), Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records *Hydrological Sciences Journal, 49(4), 591-610.*

[29] Ignaccolo M., Marani M. (2014), The Metastatistics of Daily Rainfall extremes: Non-asymptotic and Non stationary Extreme Value Analysis, *Submitted to Advances in Water Resources.*

[30] Leadbetter M. R., (1974), On extreme values in stationary sequences, *Probability theory and related fields, 38(4), 289-303.*

[31] Leadbetter M. R., Lindgren G., Rootzen H., (1983), Extremes and related properties of random sequences and processes, *Springer, New York.*

[32] Mandelbrot B., Nassim Nicholas Taleb (2005), How the finance gurus get the risk all wrong, *Fortune.*

[33] Marani M. (2003), Processi e modelli dell'idrometeorologia: Un' introduzione, *DICEA, Universita' degli studi di Padova.*

[34] Marani M., Zanetti S., (2014), Extreme events and NAO signatures in 268 years of daily rainfall observations in Padova (Italy), *Submitted to Advances in Water Resources.*

[35] Milly P., Betancourt J., Falkenmark M., Hirsch R., Kundzewicz Z. et al., (2008), Stationarity is dead: Whither water management?, *Science, 319.*

[36] Morrison J.E., Smith J.A., (2002), Stochastic modeling of flood peaks using the generalized extreme value distribution, *Water Resour. Res., Vol. 38, No. 12, 1305.*

[37] Papalexiou S.M., Koutsoyiannis D., (2013), Battle of extreme value distributions: A global survey on extreme daily rainfall, *Water. Resour. Res., 49, 187-201.*

[38] Papalexiou S.M., Koutsoyiannis D., Makropoulos C., (2013), How extreme is extreme? An assessment of daily rainfall distribution tails, *Hydrol. Earth. Syst. Sci., 17(2), 851-862.*

[39] Resnick S., (2006), Heavy-Tailed Phenomena: Probabilistic and Statistical Modeling, *Springer-Verlag, New York.*

[40] Pickands III, J., (1975), Statistical Inference Using Extreme Order Statistics, *The Annals of Statistics, 3(1), 119-131.*

[41] Salas J. D., Obeysekera J., (2014), Revisiting the Concepts of Return Period and Risk for Nonstationary Hydrologic Extreme Events, *ASCE, Journal of Hydrologic Engineering Vol. 19 No. 3, 554-568.*

[42] Scarrott, C., Mac Donald, A., (2012), A review of extreme value thresholds estimation and uncertainty quantification, *Statistical Journal Vol. 10, 33-60.*

[43] Serinaldi F., Kilsby C.G., (2014), Rainfall extremes: Toward reconciliation after the battle of distributions, *Water Resour. Res., 50, 336-352.*

[44] Serinaldi F., Kilsby C.G., (2014), Stationarity is undead: Uncertainty dominates the distribution of extremes, *Advances in Water Resources, 77, 17-36.*

[45] Smith, R.L. (2001), Evironmental statistics, *Dept. of Stat., University of North Carolina, USA.*

[46] Von Mises R., (1936), La distribution de la plus grande de n valeurs, *Rev. math. Union interbalcanique, 1(1).*

[47] Wilson P.S., Tuomi R. (2005), A fundamental probability distribution for heavy rainfall, *Geophys Res Lett 32(14): 1-4.*

[48] Yilmaz A.G., Hossain I. and Perera B.J.C., (2014), Effect of climate change and variability on extreme rainfall intensity-frequency-duration relationships: a case study in Melbourne, *Hydrol. Earth Syst. Sci., 18, 4065-4076.*