Bayesian comparison of different rainfall depth-duration-frequency relationships

Aurélie Muller, Jean-Noël Bacro, Michel Lang

▶ To cite this version:

HAL Id: hal-00589787
https://hal.archives-ouvertes.fr/hal-00589787

Submitted on 3 May 2011

Bayesian comparison of different rainfall depth-duration-frequency relationships

Aurélie Muller [(1①)], Jean-Noël Bacro [(2)], Michel Lang [(1)]

(1) *Cemagref Centre de Lyon, U.R. Hydrologie-Hydraulique, 3 bis Quai Chauveau, CP 220, 69336 Lyon cedex 09, France*

(2) *Université Montpellier II, I3M, UMR CNRS 5149*

Tel. : 33 4 72 20 87 72

Fax : 33 4 78 47 78 75

e-mail : muller@lyon.cemagref.fr

Abstract

Depth-Duration-Frequency curves estimate the rainfall intensity patterns for various return periods and rainfall durations. An empirical model based on the Generalized Extreme Value Distribution is presented for hourly maximum rainfall, and improved by the inclusion of daily maximum rainfall, through the extremal indexes of 24 hourly and daily rainfall data. The model is then divided into two sub-models for the short and long rainfall durations. Three likelihood formulations are proposed to model and compare independence or dependence hypotheses between the different durations. Dependence is modelled using the bivariate extreme logistic distribution. The results are calculated in a Bayesian framework with a Markov Chain Monte Carlo algorithm. The application to a data series from Marseille shows an improvement of the hourly estimations thanks to the combination between hourly and daily data in the model. Moreover, results are significantly different with or without dependence hypotheses: the dependence between 24 hours and 72 hours durations is significant, and the quantile estimates are more severe in the dependence case.

Keywords

Depth-Duration-Frequency; Extreme value distributions; Bivariate extreme distributions; Extremal index; Bayesian framework

Abbreviations

DDF, Depth-Duration-Frequency; MCMC, Markov Chain Monte Carlo; GEV, Generalized Extreme Value Distribution; h, hour

1. Introduction

The rainfall intensity patterns for various return periods are required for designing hydraulic structures (dams, levees, drainage systems, bridges, etc.) or for flood mapping and zoning. The objective of the rainfall depth-duration-frequency (DDF) curves is to estimate the maximum amount of rainfall for any duration and return period. This frequency analysis uses annual or seasonal maximum series, or independent values above a high threshold selected for different durations. If each duration is treated separately, contradictions between rainfall estimates can occur. DDF analysis takes into account the different durations in a single study, and prevents curves from intersecting.

The first relationship goes back as early as 1932 (Bernard, 1932). The classical approach for building DDF curves has three steps (Chow et al., 1988). In the first step, a probability distribution function is fitted to each duration sample. In the second step, the quantiles of several return periods T are calculated using the estimated distribution function from step one. Lastly, the DDF curves are determined by fitting a parametric equation for each return period, using regression techniques between the quantile estimates and the duration. The disadvantages of this procedure are the need to have a large number of parameters, and the calculation of a regression based on dependent values (since the estimated quantiles come from the same observed series, but aggregated into different time scales). There are other more consistent approaches, using for example an extreme value distribution (e.g. Koutsoyiannis et al. (1998)).

Several empirical models have been proposed (see Garcia-Bartual and Schneider, 2001 for a review). More recently, some approaches have been derived from a multifractal process (Burlando and Rosso, 1996; de Lima and Grasman, 1999; Veneziano and Furcolo, 2002; Borga et al., 2005). All these approaches need fewer parameters than the classical one, but the dependence problem remains. In section 2, two models are presented: an empirical classical model and an improved empirical model including a relation between the daily and 24 hourly maximum rainfall distributions. Section 3 presents theoretical and practical methods for estimating model parameters, quantiles and confidence intervals in a

Bayesian framework, using a Markov Chain Monte Carlo (MCMC) algorithm. Section 4 gives an application to a rainfall series for Marseille, in southern of France. Section 5 gives the conclusions of this study.

2. Depth-Duration-Frequency (DDF) relationships

2.1. Distribution of annual maximum rainfall

If $X(t)$ is the rainfall intensity at time t, then $Y_i(\delta) = \int_i^{i+\delta} X(t)dt$ is the aggregated rainfall from time i over δ hours. Then the hourly and daily observations correspond to the time series $\{Y_i(1)\}$ and $\{Y_{24i}(24)\}$ respectively. The studied variables are $H_d = \max\{Y_i(d)\}$, the annual maximum rainfall depth measured in a moving window of d hours width, and $H_D = \max\{Y_{24i}(24)\}$ the daily annual maximum rainfall depth.

A traditional approach for estimating the annual maximum rainfall H in France is based on the Gumbel distribution (Gumbel 1958):

$$G(x) = P(H \leq x) = \exp\left(-\exp\left\{-(x-\beta)/\alpha\right\}\right), \qquad (1)$$

which is a particular form ($k = 0$) of the GEV distribution:

$$G(x) = P(H \leq x) = \exp\left(-\left\{1 - k(x-\beta)/\alpha\right\}^{1/k}\right), \quad \text{with } k(\beta - x) + \alpha > 0 \quad (2)$$

There are two conventions commonly used in the literature for the sign of the shape parameter k: we have chosen the same convention as Hosking et al. (1985) : $k<0$ is equivalent to a GEV unbounded from above, or equivalently, a GEV bounded from below.

Recently, Koutsoyiannis and Baloutsos (2000), Chaouche et al. (2002), Coles et al. (2003), Coles and Pericchi (2003), Sisson et al. (2006), Koutsoyiannis (2004a, b) and Bacro and Chaouche (2006) have shown that extreme rainfall quantiles can be seriously underestimated by the Gumbel distribution. This discussion has significant practical consequences, particularly for high return periods used for the design of major hydraulic constructions or the estimation of risk of extreme floods. This paper will show an example where a GEV distribution with a negative shape parameter k is more suitable than the Gumbel distribution.

2.1.1. The empirical DDF model

The following model attempts to estimate the behavior of the hourly variables H_d. Garcia-Bartual and Schneider (2001) give a review and a comparison of nine empirical models, with two or three parameters. Koutsoyiannis et al. (1998) give the general formula:

$$I_d(T)=a(T)/b(d) \qquad (3)$$

where $I_d(T)$ is the annual maximum rainfall intensity at the return period T for the duration d; $b(d)=(d+\theta)^\eta$, with $\theta>0$, $\eta\in(0,1)$, and $a(T)=F_Y^{-1}(1-1/T)$ where F_Y is a distribution function (for example GEV, lognormal, Gamma, log Pearson III, generalized Pareto distribution) of the normalized process of intensity $I_d(.)b(d)$.

In this study F_Y will be the GEV distribution of the annual or seasonal maximum rainfall. Then, H_d has a GEV distribution, with a quantile $H_d(T)$ given by:

$$H_d(T)=dI_d(T)=d\left(\beta+\alpha/k\left\{1-\left\{-\log(1-1/T)\right\}^k\right\}\right)/(d+\theta)^\eta . \qquad (4)$$

The parameters $\alpha_d,\ \beta_d,\ k_d$ of the distribution of H_d are simply expressed with $\alpha,\ \beta,\ k,\ \theta,\ \eta$:

$$\alpha_d = d\alpha /(d + \theta)^\eta, \quad \beta_d = d\beta /(d + \theta)^\eta, \quad k_d = k. \qquad (5)$$

Before using these relationships, it needs to be determined whether one DDF model can be applied to the whole range of durations, rather than several DDF sub-models on different sub-ranges of durations.

2.1.2. The extremal index DDF model

This second model improves the first one and attempts to estimate the behavior of the variables H_d and H_D. More particularly, H_{24} and H_D describe extremes of the same process, but with different sampling frequency: H_{24} is the annual maximum rainfall, aggregated over a 24 h period, starting from any calendar hour, whereas H_D is the annual daily maximum rainfall. When daily data are available, their series

are often longer, more reliable and the network of daily rain gauge is geographically denser. Therefore, they should be included in the model. An empirical relation can be used (Weiss 1964):

$$H_{24} = 1.14 H_D \qquad (6)$$

where 1.14 is an estimation of the Hershfield factor (Hershfield, 1961). Van Montfort (1997) proposed a method for estimating this factor. A theoretical relation between distributions of H_{24} and H_D, based on the extremal index, has been proposed by Robinson and Tawn (2000) to take account for the effect of sampling frequency on extreme values distributions. The extremal index EI is the primary measure of the degree of local dependence in the extremes of a stationary process. The extremal index is defined by the following result (Leadbetter, 1983): let $\{Z_i, i=1, 2, \ldots\}$ be a stationary sequence of random variables with marginal distribution function F, satisfying a strong mixing dependence condition. Stationarity is taken in the strict sense: a process $Z_1, Z_2,\ldots$ is said to be stationary if, for any subset of integers $\{i_1,\ldots,i_k\}$, and any integer m, the joint distributions of $(Z_{i_1},\ldots,Z_{i_k})$ and of $(Z_{i_1+m},\ldots,Z_{i_k+m})$ are identical. The strong mixing dependence condition limits the degree of long-term dependence at extreme levels and is defined by: for all $i_1 < \ldots < i_p < j_1 < \ldots < j_q$ with $j_1 - i_p > l_n$

$$\left| P(Z_{i_1} \leq u_n,\ldots, Z_{i_p} \leq u_n, Z_{j_1} \leq u_n,\ldots, Z_{j_p} \leq u_n) \right.$$
$$\left. - P\left(Z_{i_1} \leq u_n,\ldots, Z_{i_p} \leq u_n\right) P\left(Z_{j_1} \leq u_n,\ldots, Z_{j_q} \leq u_n\right) \right| \leq \alpha(n, l_n) \qquad (7)$$

where $\alpha(n,l_n) \to 0$ for a sequence l_n such that $l_n \to 0$ as $n \to \infty$, and a sequence of thresholds u_n that increase with n. Then, it can be shown (Leadbetter, 1983), that the distribution of maximum is approximated by:

$$P(\max\{Z_1,\ldots,Z_n\} \leq u_n) \approx F^{nEI}(u_n) \qquad (8)$$

for large n and u_n, where $0 \leq EI \leq 1$ is the extremal index of the process. EI plays an important role in extreme value analysis, with $EI=1$ indicating independence at

asymptotically high level. Robinson and Tawn (2000) have proposed the following relation, based on hypotheses of stationarity and strong-mixing dependence of the series:

$$P(H_{24} \leq x) = P(H_D \leq x)^{24EI_{24}/EI_D} \tag{9}$$

where $0 \leq EI_D$, $EI_{24} \leq 1$ are the extremal indexes of the daily and 24 hourly series. The extreme values can be measured through the size of clusters of extreme values. A cluster definition is the following: a cluster of extreme values begins with a value above a high threshold u, and finishes when r consecutive values are under the threshold u (Beirlant et al. 2004). Let n_u denote the number of times an upper threshold u is exceeded, and n_c the number of clusters above u; n_c depends on u and r. Careful choices of u and r are needed, as if r is too small, clusters can be dependent and if r is too large, n_c becomes too small. Several methods exist to estimate the extremal index of a stationary series (Ancona-Navarrete and Tawn, 2000; Coles 2001; Beirlant et al., 2004). According to Robinson and Tawn (2000), the following estimator generally produces good estimates:

$$\hat{EI}(u, r) = n_c / n_u . \tag{10}$$

The asymptotic value $EI = \lim_{u \to \infty} \hat{EI}(u, r_u)$ can be approached using a sequence of thresholds $(u_1, \ldots, u_n)$ that increase with n, and r_u such that $r_u/u \to 0$ as $u \to \infty$. The limit is considered to have been reached when estimations of $\hat{EI}(u_n, r_u)$ are stable for u_n above some threshold u.

Let $\Theta = 24EI_{24}/EI_D$, the equation (9) implies relations between GEV parameters of both distributions (Ancona-Navarrete and Tawn, 2000; Coles, 2001):

$$\text{if } k_D = 0: \quad \beta_{24} = \beta_D + \log(\Theta)\alpha_D, \quad \alpha_{24} = \alpha_D, \quad k_{24} = 0$$
$$\text{if } k_D \neq 0: \quad \beta_{24} = \beta_D + \alpha_D/k_D(1 - \Theta^{-k_d}), \quad \alpha_{24} = \alpha_D\Theta^{-k_d}, \quad k_{24} = k_D \tag{11}$$

The daily data are included in the model (4). A new model is then defined, whose parameters are α_D, β_D, k_D, Θ, θ and η. All the parameters α_d, β_d and k_d of the

GEV distribution of H_d are simple functions of the model parameters. For example, in the case $k_D \neq 0$, model (4) becomes:

$$H_d(T)=(d/24)\left[\beta_D+\alpha_D/k_D\left\{1-\Theta^{-k_D}\left(-\log(1-1/T)\right)^{k_D}\right\}\right](24+\theta)^\eta/(d+\theta)^\eta. \qquad (12)$$

In the model, the shape parameter k_d is constant for the different durations, and equal to the shape parameter k_D. Nadarajah et al. (1998) showed theoretically, with a study of ordered multivariate extremes, that the relationship $H_d \leq H_{d'} \leq (d'/d)\,H_d$ imposes restrictions on the marginal distributions. In particular,

$$k_d = k_d' \leq 0 \quad \text{or} \quad k_d > 0, k_d' > 0 \qquad (13)$$

In our case, the rainfall is assumed not to be upwardly bounded, thus $k_d \leq 0$, and all the shape parameters are equal. Moreover, the relationship (9) between daily and 24 hours maximum rainfall implies equality between k_{24} and k_D.

2.2. Selection of two duration ranges

The model (12) has been firstly applied to model the DDF for all durations between 1 hour and 72 hours, but the estimated shape and location parameters (α_d, β_d) were outside of their marginally estimated 95% confidence intervals, for the durations 3 hours to 12 hours.

Then, since extreme cumulative rainfalls on short and long durations are derived from different meteorological processes (convective rainfalls for short durations: Llasat, (2001); Garcia-Bartual and Schneider, (2001)), two duration ranges will be considered. The empirical model from eq. (4) is chosen for the short duration rainfalls. Since long duration rainfalls are assumed to contain daily rainfall, the extremal index model from eq. (12) is used for the long duration rainfalls. Let d_b be the boundary duration that separates the short and long durations. To ensure consistency between short and long durations, the estimated parameters of both ranges have to satisfy continuity in d_b. The shape parameter is constant in both ranges, according to the theoretical study of Nadarajah et al. (1998).

Let $f_d(x; \alpha_d, \beta_d, k_d)$ be the GEV density of the maximum annual or seasonal rainfall in d hours, where α_d, β_d and k_d are the scale, location and shape parameters. Therefore, the relationships between the parameters (α_d, β_d, k_d) and the duration d are as follows:

- for short durations, $d \leq d_b$, and $\alpha_s, \beta_s, \eta_s, \theta_s$ denote the parameters of eq. (4):

$$\alpha_d = d\alpha_s /(d + \theta_s)^{\eta_s} \quad ; \qquad \beta_d = d\beta_s /(d + \theta_s)^{\eta_s} ; \qquad k_d = k_D \quad (14)$$

- for long durations, $d \geq d_b$, and α_D, β_D, k_D, Θ, θ, η denote the parameters of eq. (12), for example if $k_d \neq 0$:

$$\alpha_d = (d/24)\alpha_D \Theta^{-k_D}(24+\theta)^{\eta}/(d+\theta)^{\eta}; \ \beta_d = (d/24)\left\{\beta_D + \alpha_D/k_d(1-\Theta^{-k_D})\right\}(24+\theta)^{\eta}/(d+\theta)^{\eta}; \ k_d = k_D \quad (15)$$

Continuity hypotheses on the boundary d_b imply that α_{d_b}, β_{d_b} have the same values in both equations (14) and (15). This implies:

$$\beta_s = \alpha_s \beta_{24}/\alpha_{24}$$
$$\eta_s = \left\{\log[24\alpha_s(d_b+\theta)^{\eta}] - \log[\alpha_{24}(24+\theta)^{\eta}]\right\}/\log(d_b + \theta_s) \qquad (16)$$

With two ranges of durations, eight parameters (α_D, β_D, k_D, Θ, θ, η, α_s and θ_s) are sufficient to calculate α_d, β_d, k_d, for all d in the ranges of durations. Then, the cost to paid for this improvement is only the add of two extra parameters (α_s and $\theta_{s)}$, and the choice of a boundary duration d_b.

3. Bayesian framework

3.1. Choice of the estimation method

Many techniques exist for parameter estimation in extreme value models. For the rather complex models presented here, likelihood based techniques are particularly attractive. Different methods of inference can be drawn from the likelihood function : the procedure of maximum likelihood, but also the Bayesian

techniques. The Bayesian analysis of extreme value data is desirable since, owing to scarcity of data, they allow to include other sources of information through a prior distribution of the parameters. Moreover, the output of the Bayesian analysis, the posterior distribution of the parameters, provides an useful inference framework for the extreme quantiles through the predictive distributions. In a practical point of view, the posterior distribution are estimated by simulation, using in our paper a Markov Chain Monte Carlo algorithm (MCMC). The different steps of the algorithm are clearly described in the paper of Renard et al. (2006). The predicted quantiles are then calculated for every simulated parameters set, giving finally an estimation of the predicted distribution of the quantiles.

3.2. Prior elicitation

Table 1 gives the different prior distributions, which are similar to the choice of Coles and Pericchi (2003) for the GEV parameters.

Table 1: A priori distribution of the parameters.

Parameter	Distribution
α_D	lognormal with mean 0 and variance 100
β_D	normal with mean 0 and variance 100
k_D	uniform on [-1,1]
Θ	uniform on [1,24]
θ	normal with mean 0 and variance 100, truncated at 0
η	lognormal with mean 0 and variance 100
Φ	uniform on [0,1]

Since the clusters are larger in the time series of 24 hourly rainfall data than in the time series of daily rainfall, $EI_{24} \leq EI_D$, therefore $\Theta = 24EI_{24}/EI_D \leq 24$. Moreover, $\Theta \geq 1$. Indeed, let $N_{r_n}(u_n)$ denote the number of excess of u_n in r_n consecutive measures of 24 hourly rainfall ($H_{24,1},...,H_{24,r_n}$), for r_n such that $\lim_{n\to\infty} r_n/n = 0$. A cluster of extremes is defined to occur when $N_{r_n}(u_n) > 0$, with the values in the cluster being the excedances of u_n. The cluster size distribution $\pi_{24,n}$ is defined by:

$$\pi_{24,n}(j) = P(N_{r_n}(u_n) = j \mid N_{r_n}(u_n) > 0), \; for \; j = 1,...,r_n \qquad (17)$$

The limiting cluster size distribution is:

$$\pi_{24}(j)=\lim_{n\to\infty}\pi_{24,n}(j), \; for \; j=1,...,\infty \tag{18}$$

Robinson and Tawn (2000) showed that:

$$EI_D \le 24EI_{24}\left[1-\sum_{i=1}^{23}(1-i/24)\pi_{24}(i)\right] \tag{19}$$

Therefore, as:

$$1-\sum_{i=1}^{23}(1-i/24)\pi_{24}(i) \ge 1-\sum_{i=1}^{\infty}\pi_{24}(i)+\sum_{i=1}^{23}i/24\pi_{24}(i)=\sum_{i=1}^{23}i/24\pi_{24}(i)>0 \tag{20}$$

this implies that:

$$\Theta = 24EI_{24}/EI_D \ge 1/\left(1-\sum_{i=1}^{23}(1-i/24)\pi_{24}(i)\right)\ge 1. \tag{21}$$

Therefore $\Theta \in [1,24]$.

As θ must be positive, a normal distribution truncated at 0 is chosen. Moreover, parameters of both models have to satisfy some physical constraints:

$$d<d' \; \Rightarrow \; H_d \le H_{d'} \le (d'/d)H_d$$
$$H_D \le H_{24} \le 2H_D \tag{22}$$

If the quantiles calculated for different return periods T=2, 5, 10, 100 and 1,000 years do not verify these two relations, the parameters are rejected in the MCMC algorithm.

3.3. Likelihood definition

Three different likelihood formulations L_1, L_2 and L_3 will be used.

- Likelihood L_1: independence between seven durations.

Firstly, independence is supposed between durations d = 1 h, 6 h, 12 h, 24 h, 48 h, 72 h and daily (D) observations. The likelihood is expressed as

$$L_1 = \prod_{d=1,6,12,24,48,72,D} \prod_{i=1}^{i_d} f_d(x_i^{(d)}; \alpha_d, \beta_d, k_d) \qquad (23)$$

where $x_i^{(d)}$, i=1,…,i_d are the annual or seasonal maximum rainfall measured in d hours and α_d, β_d, k_d are given by equations (14) to (16). The advantage of this likelihood is the use of a large set of available data.

- Likelihood L_2: independence between four durations

Secondly, since independence between all these durations is an unlikely hypothesis, only four durations will be considered. Since the one-hour maximum rainfall generally occurs during a thunderstorm, whereas the 72-hours maximum rainfall occurs generally during a frontal rainfall event, both maxima are assumed to originate from different processes. Both durations may be considered independent, as shown by Kieffer Weisse (1998). Moreover, 24 hourly data, and daily data of years without hourly measurements will be used. The likelihood formula is given by:

$$L_2 = \prod_{d=1,24,72,D} \prod_{i=1}^{i_d} f_d(x_i^{(d)}; \alpha_d, \beta_d, k_d) \qquad (24)$$

where α_d, β_d, k_d are given by equations (14) to (16), $i'_1 = i_1$, $i'_{24} = i_{24}$ $i'_{72} = i_{72}$ and i'_D is the number of years of daily measurements without hourly measurements.

- Likelihood L_3: four durations, with dependence between both of them.

Lastly, a generalization of the second likelihood is introduced: the dependence between 24 hours and 72 hours maximum rainfall is considered through a bivariate extreme distribution, from the logistic family (Coles, 2001), defined for marginal Fréchet distributions:

$$G(x,y)=P(u_{24}(H_{24})\leq x, u_{72}(H_{72})\leq y)=\exp\{-(x^{-1/\Phi}+y^{-1/\Phi})^{\Phi}\}, \; x>0, \; y>0, \quad (25)$$

for a dependence parameter $\Phi\in(0,1)$. The transformed variables $u_d(H_d)=-1/\log(G_d(H_d))$, where G_d is the GEV distribution of rainfall H_d, are standard Fréchet distributed, because $P(-1/\log(G_d(H_d)) \leq x) = P(G_d(H_d) \leq \exp(-1/x)) = \exp(-1/x)$ since $G_d(H_d)$ is uniformly distributed between 0 and 1. As $\Phi\rightarrow 1$, $G(x,y)\rightarrow\exp\{-(x^{-1}+y^{-1})\}$, corresponding to independent variables; as $\Phi\rightarrow 0$, $G(x,y)\rightarrow\exp\{-\max(x^{-1}, y^{-1})\}$, corresponding to perfectly dependent variables. This is the most widely used model in bivariate extreme value analysis (Coles, 2001). The dependence structure of any bivariate extreme value distribution function G can be described in several ways. A quite popular way is the Pickands dependence function A, satisfying some properties (Beirlant et al., 2004). The Pickands dependence function A(t), is defined for $t\in[0,1]$ by:

$$A(t)=-\log\left(G\left[G_1^{-1}\{\exp(-1+t)\}, G_2^{-1}\{\exp(-t)\}\right]\right) \quad (26)$$

where G_1, G_2 are the two marginal distributions. G is completely determined by its margins G_1, G_2 and its Pickands dependence function A(t) through equation (26). A can be estimated by non-parametric (Pickands 1981, 1989; Capéraà et al. 1997), or by parametric methods. The comparison between non-parametric and parametric estimators of A is a way to validate the parametric model.

The likelihood is given by:

$$L_3=\prod_{i=1}^{i24} g\left(u_{24}(x_i^{(24)}), u_{72}(x_i^{(72)}); \alpha_{24}, \beta_{24}, k_{24}, \alpha_{72}, \beta_{72}, k_{72}, \Phi\right) u'_{24}(x_i^{(24)}) u'_{72}(x_i^{(72)})$$
$$\prod_{i=1}^{i_D} f_D(x_i^{(D)}; \alpha_D, \beta_D, k_D) \prod_{i=1}^{i_1} f_1(x_i^{(1)}; \alpha_1, \beta_1, k_1) \quad (27)$$

where u'_d is the derivative function of u_d, and g is the density function of the bivariate logistic distribution. Note that if $k_d=0$, $u_d(x)=\exp((x-\beta)/\alpha)$, and if $k_d\neq 0$, $u_d(x)=(1-k_d(x-\beta)/\alpha)^{-1/k_d}$.

4. Applications of the DDF models to the Marseille rainfall data series

4.1. Presentation of the series

Two data series at Marseille are available, with 67 years of hourly data (1918-2002) and 122 years of daily data (1882-2003). The daily series have been reviewed by Météo-France through the European project IMFREX for the study of climate changes, using homogeneity criteria. In both hourly and daily series, each year has less than 10% of missing values. In order to exclude problems of non-homogeneity due to seasonality, two different seasons have been considered based on the mean of monthly maximum rainfall (Kieffer Weisse, 1998) (see Figure 1). As heavy hourly and daily rainfalls occur within the September-January period, this period will be chosen as representative of extremal events. All the results presented are calculated within this period.
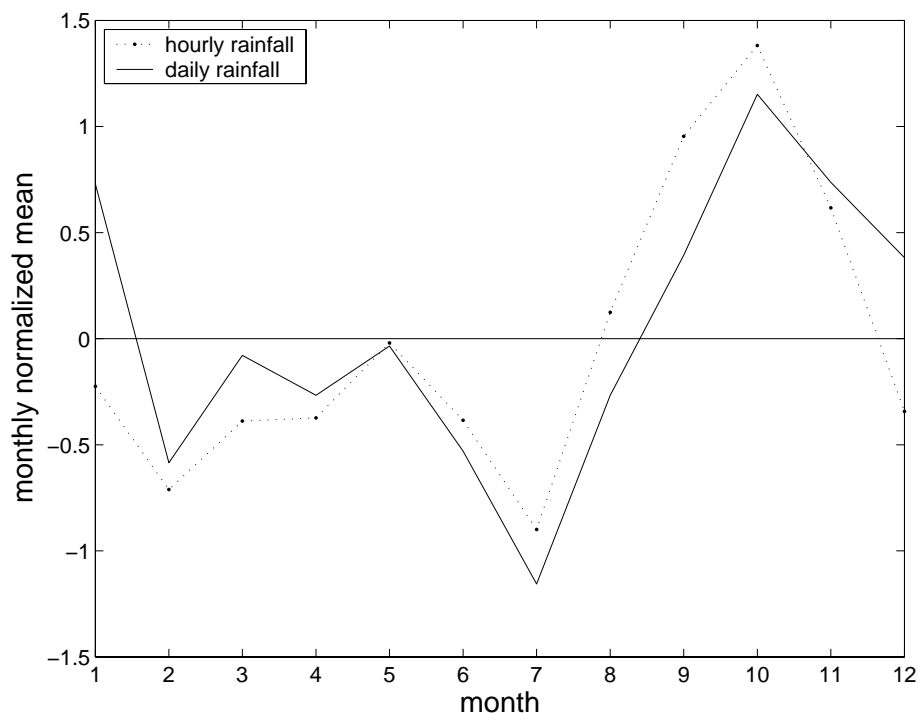

Figure 1: Seasonal fluctuation of monthly maximum rainfall: normalized and centered mean of monthly maximum rainfall for the daily and one hour rainfall data.
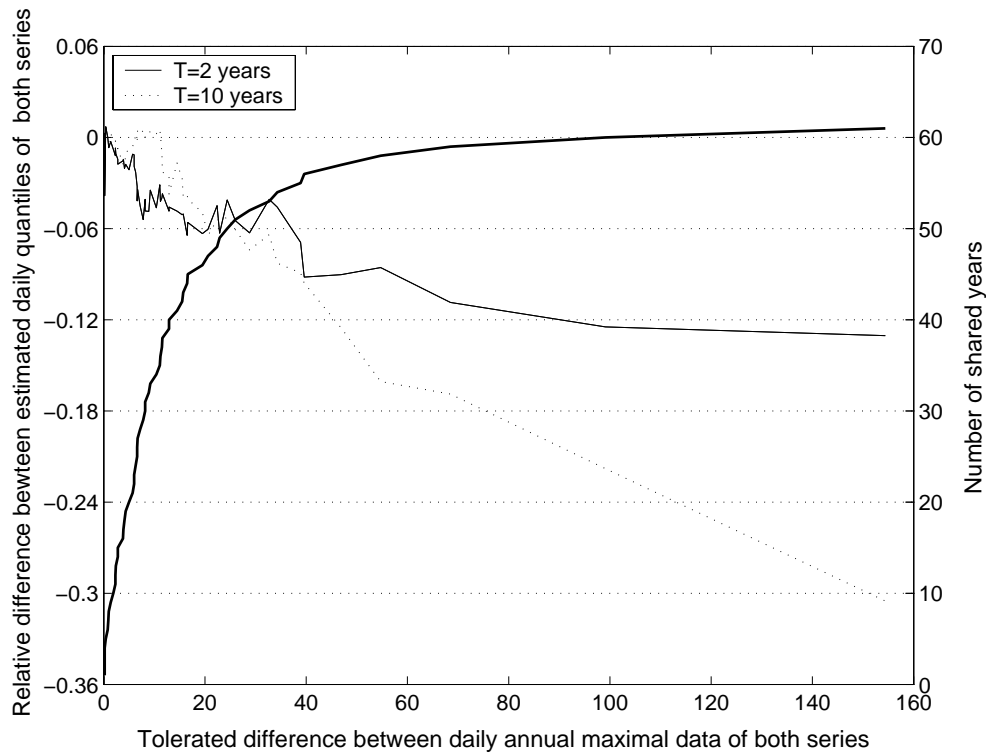
Figure 2: Selection of the tolerance threshold for the difference between daily annual maxima of both daily and hourly series. Left Y axis: relative difference between daily quantiles of both series (corresponding curves: solid and dotted lines for $T = 2$ and 10 years). Right Y axis: number of tolerated years in the hourly series, in common with the daily series (corresponding curve: bold line).

The consistency between the two data series has been tested, comparing their daily annual maximum values and their associated quantiles. The daily series is taken as the reference series, since it has been reviewed by Météo-France. With perfect measures in both series, the annual maxima would be equal. But because of some failures or missing values of the hourly measures, particularly in the extreme events, the annual maxima can be different. Here, a simple method is employed: retaining the years of the hourly series where the absolute value of the difference between daily annual maxima is less than some tolerance threshold. This tolerance threshold is calculated by the following method: for any tolerance threshold between 1 mm and around 160 mm, two hourly and daily sub-series are defined. Both sub-series contain only those years whose difference between rainfall maximum amounts is lower than the tolerance. The optimum tolerance threshold is a compromise between the number of selected years in the sub-series (bold curve in Figure 2) and the relative difference between daily quantiles estimated for both sub-series (solid and dotted lines in Figure 2, for return periods

14

2 and 10 years). The chosen threshold is 19 mm, which corresponds to a relative difference less than 6%, and 45 validated years in the hourly series.
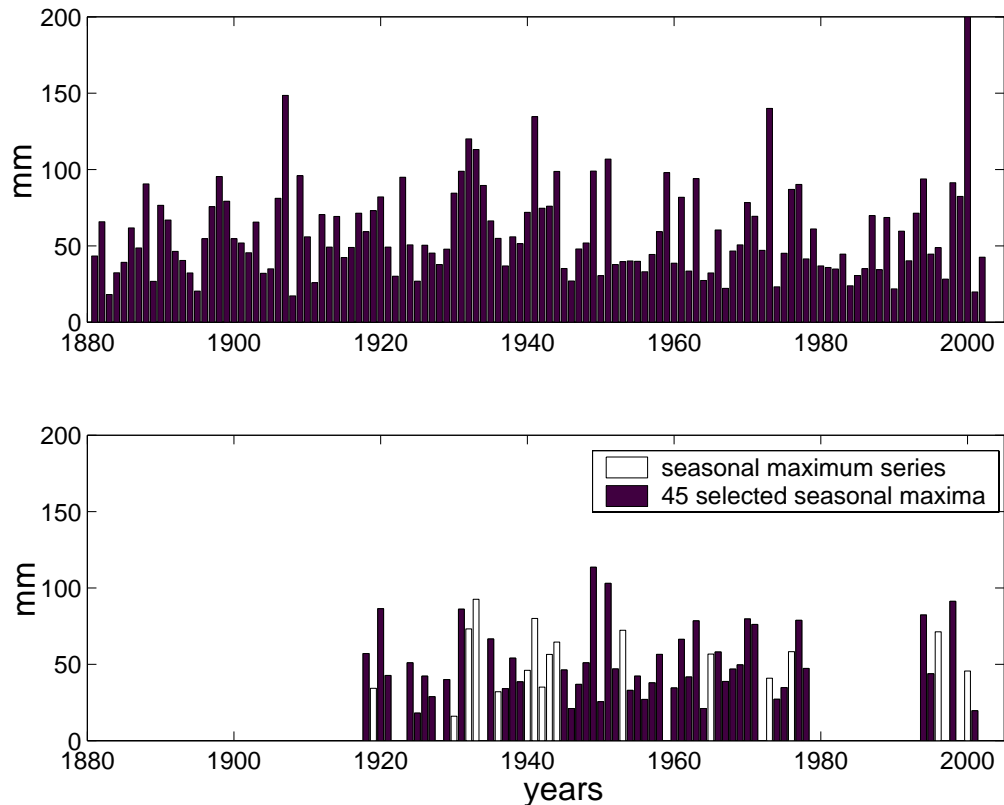

Figure 3: Top: daily seasonal maximum rainfall amount of the daily Marseille series (122 years). Bottom: daily seasonal maximum rainfall depth calculated on the hourly series (67 years). Black bars represent the 45 selected years of hourly series, after check of the consistency between series.

A comparison between the extreme values of the two data sets (daily and hourly values on Figure 3) shows that the hourly recording rain gauge has some difficulties in recording very extreme rainfall events. For example, the extreme events in 2000 (200 mm), 1973 (140 mm) and 1932 (120 mm) are missing values in the hourly rainfall series. The differences in reliability of daily rain gauge data and hourly recording data are very common, but not widely recognized. Similar observations and methodologies are discussed in Koutsoyiannis et al. (1998).

Some aspects of stationarity of the 122 years series have been checked. Firstly, a likelihood ratio test has been applied between a GEV distribution with a temporal trend in scale and position parameters, and a GEV distribution with fixed parameters. Both distributions are nested models $M_s \subset M_t$, where M_s, M_t are the stationary and trend models. Therefore the deviance statistic:

$$D = 2(l_t(M_t) - l_s(M_s)) \qquad\qquad (28)$$

is χ_2^2 distributed, where $l_s(M_s)$, and $l_t(M_t)$ are the maximized log-likelihoods for models M_s and M_t respectively. As the computed statistic $D = 0.186$, the stationary model is not rejected for every level lower than 25%. Secondly, stationarity has been tested on different annual variables: mean annual rainfall of wet days (with more than 1 mm precipitation), annual maximum, annual ratio of zero rainfall, annual ratio of values above upper thresholds. The non-parametric Mann-Kendall test (Mann, 1945; Kendall, 1975) is used to detect monotonic trends in series of independent data. With a tacit hypothesis of independence of the four annual series, no significant trend was detected for the four variables by the Mann-Kendall test, for a level $\alpha=10\%$. These stationarity results are in agreement with the European project IMFREX (Dubuisson and Moisselin, 2006).

4.2. Dependence between rainfall depths

The correlation coefficient between 1 hourly and 72 hourly maximum rainfall is equal to 0.41. This value is quite high, therefore the independence hypotheses is only justified by physical reasons, as the corresponding rainfall processes are considered to be different (Kieffer Weisse, 1998). The correlation coefficient is 0.56 between 1 hour and 24 hours, and 0.90 between 24 hours and 72 hours, justifying the bivariate distribution in likelihood L_3 between 24 hourly and 72 hourly rainfalls.
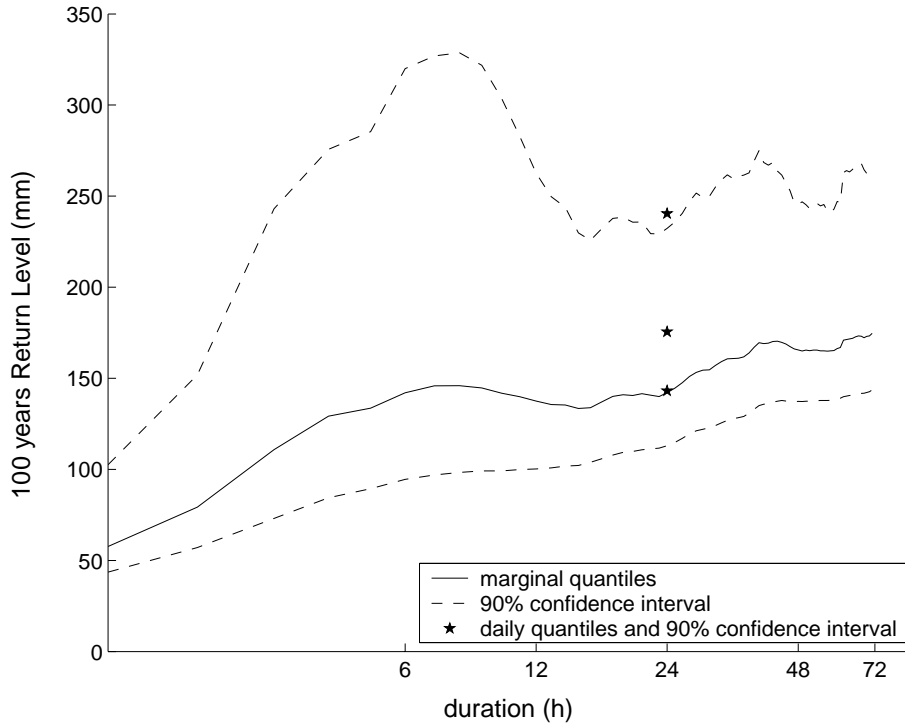
Figure 4: Marginal estimates and 90% confidence intervals (estimated by a Bayesian analysis) of $H_D(100)$ (stars) and $H_d(100)$ for d=1 h to 72 h.

From Figure 4, it is seen that the estimated value for $H_6(100)$ is too large to meet the required marginal ordering constraint with $H_{12}(100)$. This incoherence is due to sampling hazard: the maximal rainfall of the year 1951 was 103 mm in six hours, and 104 mm in 12 hours. Those values are also the maximal values of the series of annual maximal in six and 12 hours. Then estimates of six hours quantiles can be higher than estimates of 12 hours quantiles for high frequencies. Then the marginal estimate $\hat{k}_{12}$ is larger than $\hat{k}_6$. $H_{24}(100)$, $H_{48}(100)$, $H_{72}(100)$ are too small to meet the ordering constraints with H_D. This is explained by the fact that the hourly series contains only 45 years, whereas the daily series is 122 years long, and the largest rainfall values are not included within the hourly series.

The logistic bivariate extreme distribution is fitted to the bivariate rainfall data (24 h, 72 h). The Pickands dependence function of this particular distribution compares well with its non-parametric estimators by Pickands (1981, 1989) and Capéraà et al. (1997) (see Figure 5).
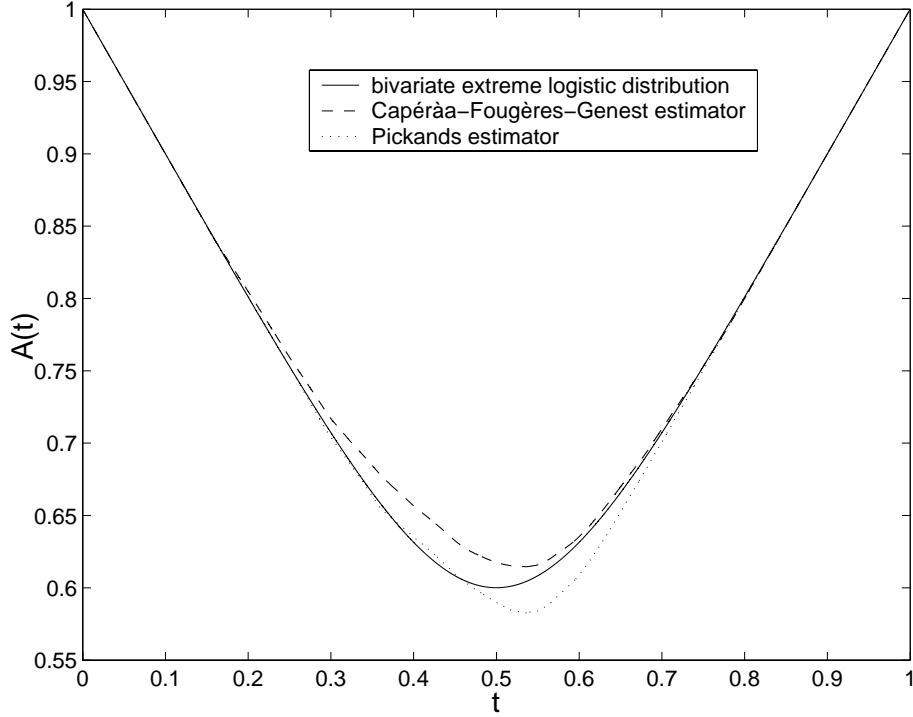
17

Figure 5: Estimation of the Pickands dependence function $A(t)$: comparison between non-parametric Pickands and Capéraà *et al.* estimations and logistic estimation.

Moreover, a likelihood ratio test has been applied between three nested models, fitted to the bivariate data (24 h, 72 h): independence case, logistic bivariate distribution and a logistic asymmetric bivariate distribution, whose bivariate distribution is:

$$G(x,y)=P(H_{24}\leq x, H_{72}\leq y)=\exp\left(-(1-\Psi_1)x^{-1}-(1-\Psi_2)y^{-1}-\left\{(\Psi_1 x^{-1})^{1/\Phi}+(\Psi_2 y^{-1})^{1/\Phi}\right\}^{\Phi}\right)$$

$$\text{with } \Psi_1,\Psi_2 \in [0,1]$$

(29)

The logistic asymmetric bivariate distribution is a more general model than the logistic symmetric one: the two variables are exchangeable in the symmetric case, but not in the asymmetric case. No significant difference was detected at the 10% level between the logistic and asymmetric logistic models ($\hat{\psi}_1=1, \hat{\psi}_2=1$), but the logistic model was better than the independence model, with a significant ratio test (the *p*-value is lower than 0.1%). The dependence parameter Φ of the logistic distribution was estimated to be 0.24 by likelihood maximization, implying a high level of dependence between 24 hourly and 72 hourly rainfall data. The effect of

the bivariate logistic distribution is to change the shape parameter estimation: marginally, $\hat{k}_{24}=0$, $\hat{k}_{72}=0.04$, but with the bivariate logistic distribution, applied on 24 and 72 hourly data and without the constraint $k_{24}=k_{72}$:

$\hat{k}_{24}=-0.14$, $\hat{k}_{72}=-0.12$, which is close to the daily marginal estimator: $\hat{k}_D=-0.13$.

4.3. Choice of duration ranges

The durations are separated into two ranges. The boundary duration d_b between short and long durations is added as an extra parameter, in likelihood definitions L_1, L_2, and L_3. Then the boundary duration is estimated by likelihood maximization, under the constraints (22). The maximum likelihood estimator of d_b is 5.6, with likelihood L_1, whereas likelihoods L_2, L_3 are not discriminant for d_b, and give equal maximum likelihoods with d_b=5, 6 or 7 hours, while other parameters change slightly.
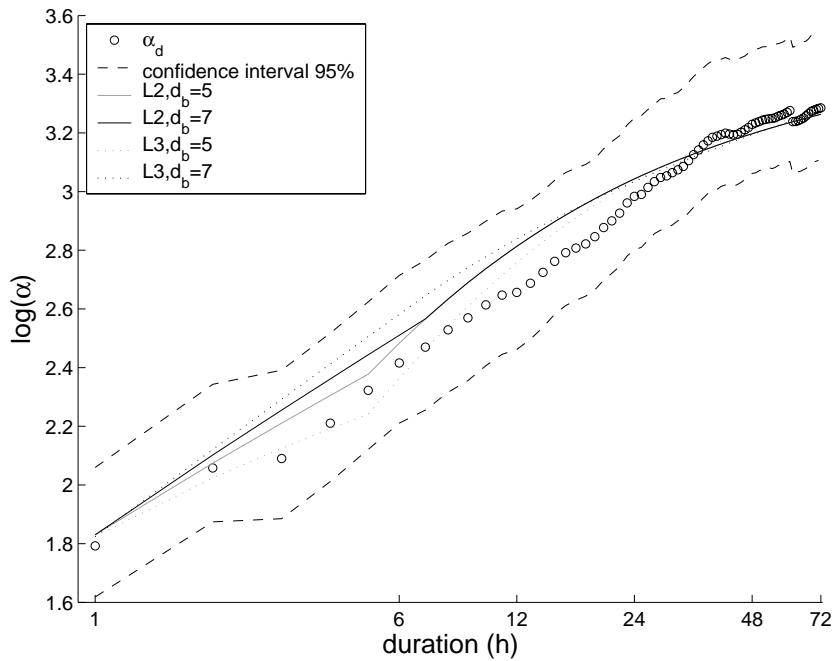

(a)
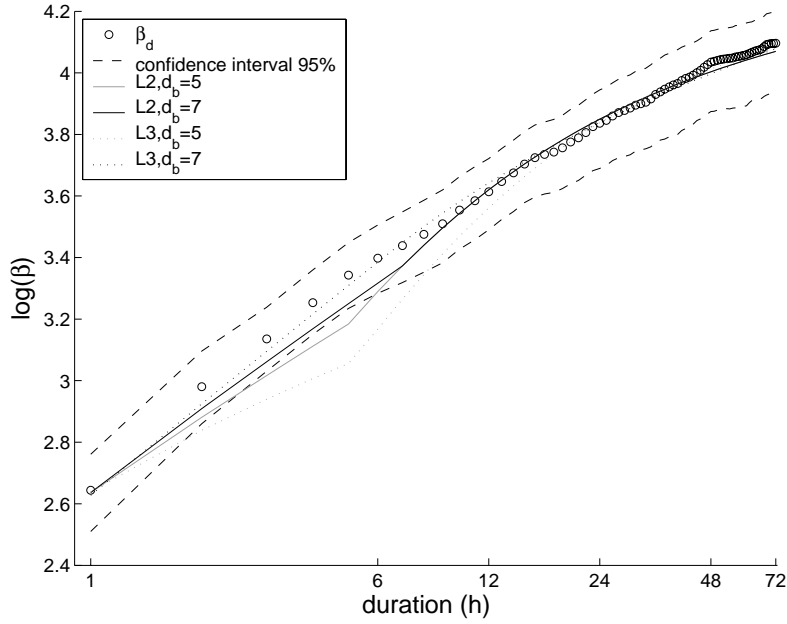
(b)

Figure 6: Choice of the duration d_b of the boundary between short and long durations. Comparison under L_2 and L_3 maximizations and constraints (22), with two splits (d_b = 5 or 7 hours), of: (a) α_d estimates; (b) β_d estimates.

Figure 6 shows that d_b=7 hours gives a better fit of the parameters, with estimations inside the 95% confidence for α_d and β_d. Estimations with d_b =5 hours are closed to the marginal estimate $\hat{\alpha}_d$, but outside the 95% confidence interval of $\hat{\beta}_d$. Results with d_b =6 hours are intermediate, but outside the 95% confidence interval of $\hat{\beta}_d$. The parameter k_d is not used for the choice of d_b, since k_d is constant in the model (k_d=k_D), and d_b does not affect the parameter k_D of the daily data. The chosen value for d_b is therefore 7 hours.
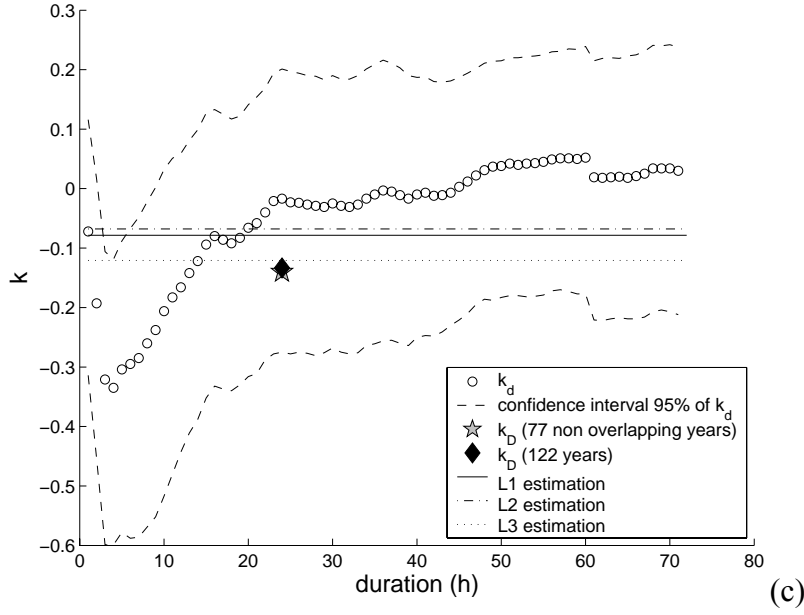
(a)

(b)
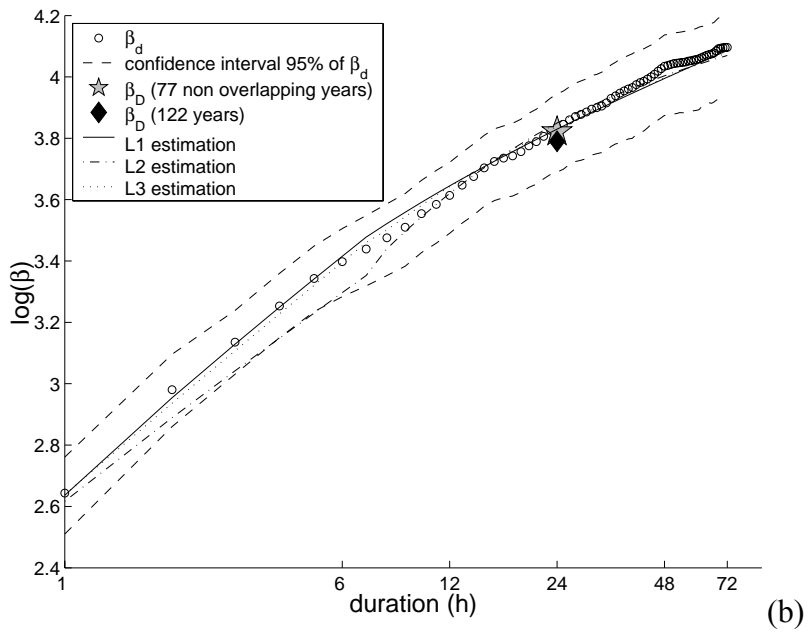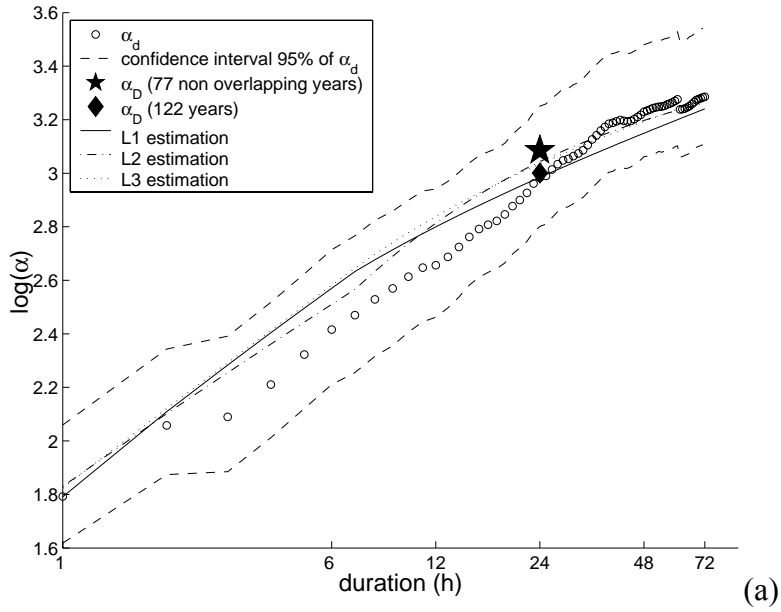
Figure 7: Comparison of α_d, β_d, k_d estimates under L_1, L_2 and L_3 maximizations and constraints (22): (a) α_d estimates; (b) β_d estimates; (c) k_d estimates.

Figure 7 shows the good fit of the maximum likelihood estimations under constraints (22), for L_1, L_2, L_3. The estimations of $\hat{\alpha}_d$, $\hat{\beta}_d$ with likelihoods L_1, L_2 and L_3 are close to the marginally estimated parameters. The marginal estimates of $\hat{k}_d$ are approximately constant after about 24 hours, and present a minimum for the 4 hours duration. The variability of $\hat{k}_d$ is due to the sampling sensitivity of the estimator, and to the fact that the maximum rainfall in six hours (103 mm) is very close to the maximum rainfall in 15 hours (104 mm).

4.4. Comparison between the three likelihood definitions

The results are presented after the run of 80 000 MCMC simulations. The parameters were computed on the last 40 000 iterations, thus allowing 40 000 burning iterations. The convergence of the MCMC algorithm is assessed by the R statistic (Gelman et al., 1997) calculated for each parameter in the second half of the burning iterations. Eight parallel sequences of Metropolis algorithm have been considered, with a random starting point, sampled in the prior distribution. As the computed ratio R is very close to one, the convergence of the MCMC simulations can be accepted.

22

The estimated parameters are presented in the Table 2, with the median of the 40 000 last simulated parameters. The 90% confidence intervals are calculated by sorting each marginal simulated parameter, and excluding the values lower and larger than the 5% and 95% empirical quantiles. The comparison between the median and the middle of the 90% confidence interval shows that the posterior distribution of (α_D, β_D, k_D) is symmetric, but this was not the case for the other parameters.

Table 2: Estimated parameters and 90% confidence intervals.

	L_1	L_2	L_3
α_D	44.06(41.07,46.99)	44.82(40.92,48.65)	44.47(40.44,48.39)
β_D	20.05(18.45,21.83)	21.38(19.16,23.87)	21.23(18.89,23.89)
k_D	-0.083(-0.158,-.014)	-0.076(-0.183,0.015)	-0.131(-0.236,-0.035)
Θ	1.17(1.02,1.42)	1.18(1.02,1.48)	1.19(1.02,1.54)
θ	4.09(-1.03,11.88)	5.61(-1.33,21.66)	9.22(-1.42,32.96)
η	0.89(0.69,1.12)	0.96(0.74,1.25)	0.96(0.75,1.41)
α_s	10.15(4.41,16.89)	10.23(4.38,21.24)	9.36(4.27,18.51)
θ_s	0.94(-0.50,2.08)	0.88(-0.55,2.95)	0.68(-0.55,2.45)
Φ			0.265(0.212,0.335)

The MCMC estimates of the parameters α_d, β_d, k_d, for d between 1 h and 72 h (not graphically shown) are similar to those obtained with the maximum likelihood (presented in Figure 7). In the three cases, the k_d estimations are negative, implying unbounded quantiles, when the return period becomes infinite.
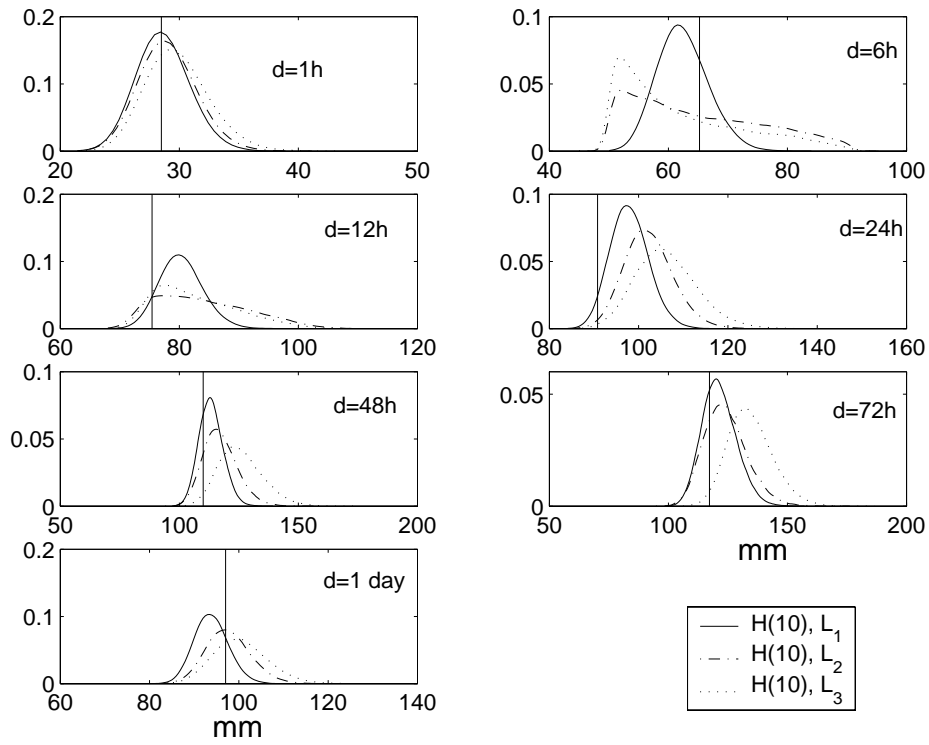
Figure 8: Posterior distributions of the quantiles $H_d(10)$ for d=1 h, 6 h, 12 h, 24 h, 48 h, 72 h and d=1 day (with the last 40 000 simulations of the MCMC algorithm). The vertical lines are the marginally estimated quantiles $H_d(10)$.
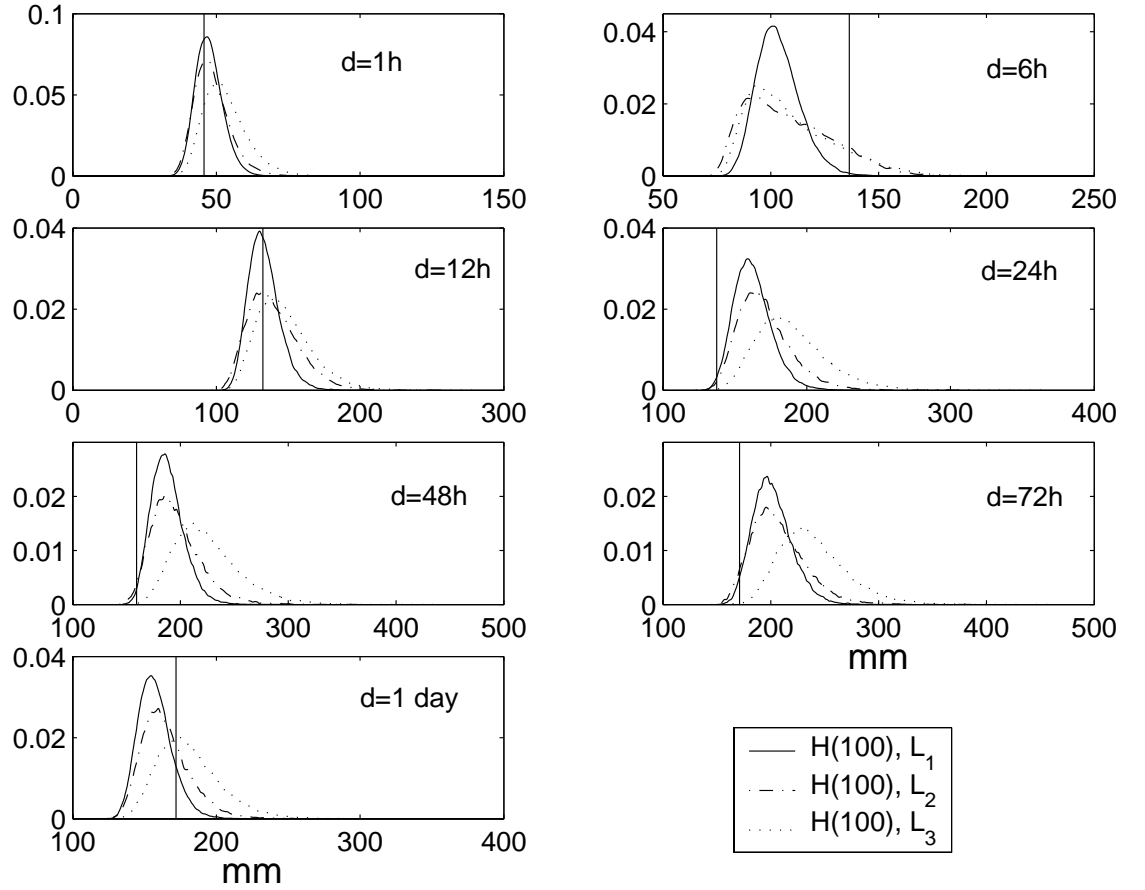
Figure 9: Posterior distributions of the quantiles H_d (100) for d=1 h, 6 h, 12 h, 24 h, 48 h, 72 h

and d=1 day (with the last 40 000 simulations of the MCMC algorithm). The vertical lines are

the marginally estimated quantiles H_d (100).

The posterior distributions of the quantiles $H_d(T)$ are presented in Figure 8 and Figure 9 for the return periods T = 10 and 100 years. The shape of the distributions is generally skewed, and the supports become larger with the successive likelihood definitions (L_1 to L_3). Namely, less data are used in the L_2 definition than in the L_1 one, and the L_3 definition includes an extra dependence parameter Φ.

As the daily series contains a large number of extreme values (200 mm, 148 mm, 140 mm, 138 mm, etc.) and as the hourly series does not contain the most extreme values, the estimated long duration quantiles are significantly larger than the marginal estimates (vertical lines in Figure 8 and Figure 9), especially when fewer hourly data are included in the estimation procedure (L_2, L_3). This is due to the link between long durations and daily rainfalls, by Θ in equation (12). For the

same reason, daily quantiles are slightly lower than their marginal estimates. Moreover, the short durations are linked to the long durations only by k_D, and by continuity hypotheses of the parameters α_d, β_d at the boundary between short and long durations (cf. equations (14) to (16)). Thus daily data produce less effect on the short duration estimations. The six hours quantiles estimations are lower than their marginal estimations, because of the linkage between 1 hourly and 6 hourly rainfalls. The six hourly marginal quantiles are high because of the presence in the six hours series of the maximum rainfall falling in 15 hours, not present in the one hourly series.
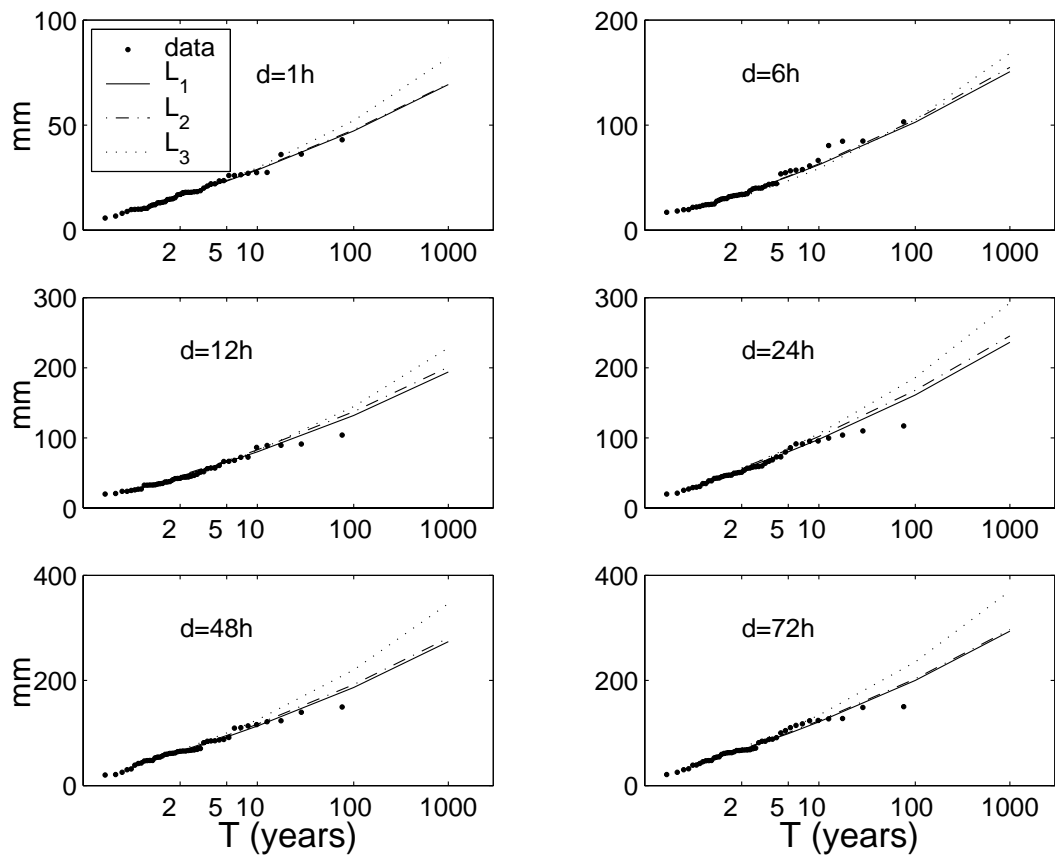

Figure 10: Median of the simulated distributions of maximum rainfall amount H_d, d=1 h to 72 h (with the last 40 000 MCMC simulations).
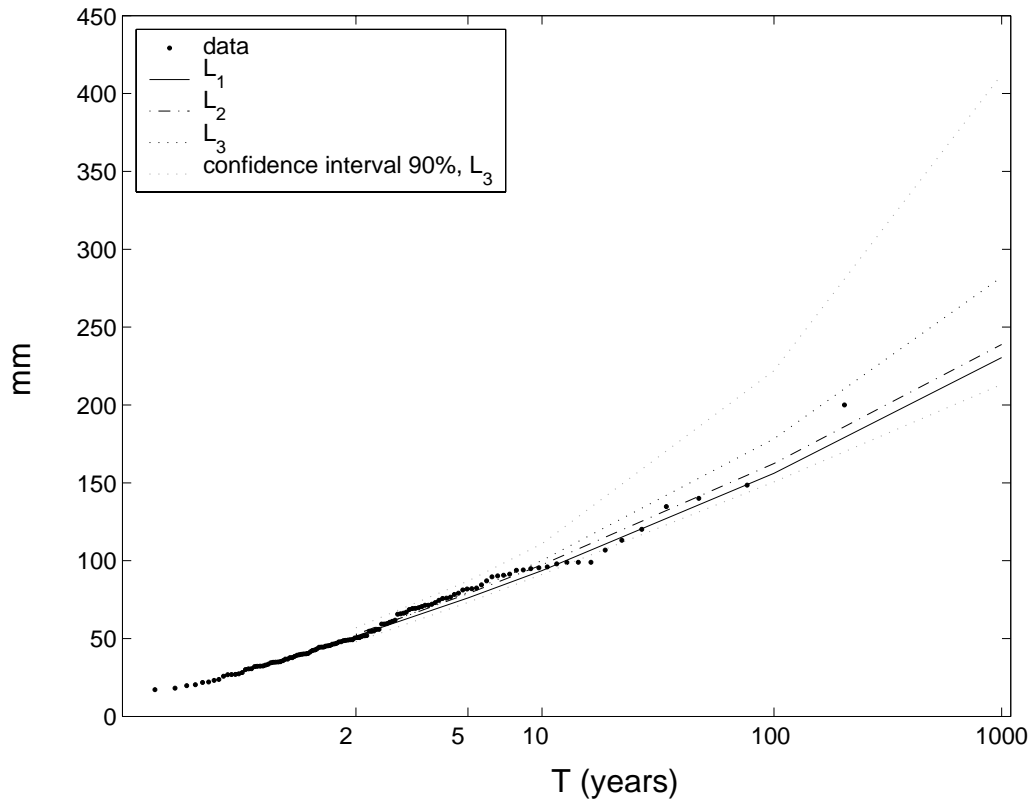
Figure 11: Median of the simulated distribution of maximum daily rainfall amount (with the last 40 000 MCMC simulations).

Figure 10 and Figure 11 show the DDF curves: the median quantiles are quite similar with the L_1 and L_2 likelihood definitions, and are larger with L_3, due to the k_D value in this last case (cf. Table 2). Thus the logistic dependence model has a significant influence on the estimations, with a dependence parameter Φ equal to 0.26: its 90% confidence interval does not contain 1, which corresponds to the independent case.

5. Conclusions

Based on hypotheses of GEV or Gumbel distribution for the seasonal maximum rainfall distribution, the DDF models agree with a GEV distribution, with negative shape parameter. The quantiles of any duration between 1 hour and 72 hours, and any return period between 2 years and 1000 years have been estimated

by the proposed empirical model, associated with three different likelihood definitions. The durations have been separated into short (less than 7 hours) and long (above 7 hours) durations. The proposed model has eight parameters, or nine if the 24 hours and 72 hours rainfalls are modeled by a bivariate logistic distribution. The likelihood choice formulates hypotheses of independence or dependence between data, and needs a choice of the most representative and non-redundant data. Both independence hypotheses and the bivariate distribution have been used to define likelihoods. In a future work, multivariate distributions with more than two dimensions would provide the model with more information.

The daily rainfall data have been included into the DDF study, through the extremal indexes. Then estimates have been improved since daily series are often longer than hourly series. Namely, the daily Marseille series contains measurements that have not been recorded by the hourly rain gauge, and particularly extreme values. The rainfall intensities on durations 12 hours to 72 hours are linked together and with daily rainfalls, by the extremal indexes of the daily and 24 hourly series, and by the same shape parameter. The quantiles estimated by the models are thus significantly larger than those marginally estimated, in the case of the long durations, proving the important effect of adding the daily data into the model.

A significant difference between independence hypotheses and bivariate logistic distribution has been shown in the case of the 24 hourly and 72 hourly rainfall data. The dependence case, treated with the bivariate logistic distribution, gives a stronger negative scale parameter, close to the parameter $\hat{k}_D$ estimated on the daily maximal data. The dependence parameter estimated in this case is about 0.26 with a 90% confidence interval equal to [0.21, 0.33], showing the strong dependence between these two durations. Moreover, the likelihood ratio test between models with or without dependence shows that the bivariate logistic distribution is significantly valid relative to the more general asymmetric bivariate logistic distribution, and significantly better than the model under independence hypotheses.

These results have been allowed by the Bayesian framework, which gives a method for defining the posterior distribution of parameters, and includes the prior

knowledge on the parameters and the physical behavior of rainfall. Estimations and confidence intervals of parameters have been calculated through a two-step MCMC algorithm. The posterior distributions are generally far from normal and reproduce the heavy tail of the quantiles, proving the usefulness of the Bayesian approach instead of a maximum likelihood estimation of confidence intervals, based on the asymptotic normality of the estimators. Bayesian and maximum likelihood estimations of the medians are quite similar.

Acknowledgement

References

Ancona-Navarrete, M. A. and J. A. Tawn (2000). A comparison of methods for estimating the extremal index. Extremes 3(1): 5-38.

Bacro, J.N. and A. Chaouche (2006). Incertitude d'estimation des pluies extrêmes du pourtour méditerranéen: illustration par les données de Marseille. Hydrological Sciences Journal 51(3): 389-405.

Beirlant, J., Y. Goegebeur, J. Segers and J. Teugels (2004). Statistics of Extremes, Theory and Applications.

Bernard, M. M. (1932). Formulas for rainfall intensities of long durations. Trans. ASCE 96: 592-624.

Borga, M., C. Vezzani and G. Dalla Fontana (2005). Regional rainfall depth-duration-frequency equations for an alpine region. Natural Hazards 36(1-2): 221-235.

Burlando, P. and R. Rosso (1996). Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. Journal of Hydrology 187(1-2): 45-64.

Capéraà, P., A. L. Fougères and C. Genest (1997). A non-parameteric estimation procedure for bivariate extreme value copulas. Biometrika 84: 567-577.

Chaouche, K., P. Hubert and G. Lang (2002). Graphical characterization of of probability distribution tails. Stochastic Environmental Research and Risk Assessment 16(5): 342-357.

Chow, V. T., D. R. Maidment and L. W. Mays (1988). Applied hydrology.

Coles, S. (2001). An introduction to statistical modeling of extreme values. London.

Coles, S. and L. Pericchi (2003). Anticipating catastrophes through extreme value modelling. Journal of the Royal Statistical Society. Series C: Applied Statistics 52(4): 405-416.

Coles, S., L. Pericchi and S. Sisson (2003). A fully probabilistic approach to extreme rainfall modelling. Journal of Hydrology 273(1-4): 35-50.

de Lima, M. I. P. and J. Grasman (1999). Multifractal analysis of 15-min and daily rainfall from a semi-arid region in Portugal. Journal of Hydrology 220(1-2): 1-11.

Dubuisson, B. and J.M. Moisselin (2006). Evolution des extrêmes climatiques en France à partir des séries observées, Colloque SHF, Lyon, France.

Garcia-Bartual, R. and M. Schneider (2001). Estimating maximum expected short-duration rainfall intensities from extreme convective storms. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 26(9): 675-681.

Gelman, A., J. B. Carlin, H. S. Stren and D. B. Rubin (1997). Bayesian Data Analysis. London.

Gumbel, E. J. (1958). Statistics of Extremes. New York, Columbia University Press.

Hershfield, D. M. (1961). Rainfall frequency atlas of the United States for durations from 30 minutes to 24 hours and return periods from 1 to 100 years. U. S. D. o. C. Weather Bureau Technical Paper 40. Washington D.C.

Hosking, J., J. Wallis, and E. Wood (1985). Estimation of the generalized extreme value distribution by th method of probability-weighted moments. Technometrics, 27(3):251-261.

Kendall MG. (1975). Rank correlation methods. London: Griffin.

Kieffer Weisse, A. (1998). Etude des précipitations exceptionnelles de pas de temps court en relief accidenté (Alpes françaises). Méthode de cartographie des précipitations extrêmes. Ph. D. Mécanique des milieux géophysiques et Environnement. Grenoble, Institut National Polytechnique de Grenoble.

Koutsoyiannis, D. (2004a). Statistics of extremes and estimation of extreme rainfall. I. Theoretical investigation. Hydrological Sciences Journal 49(4): 575-590.

Koutsoyiannis, D. (2004b). Statistics of extremes and estimation of extreme rainfall. II. Empirical investigation of long rainfall record. Hydrological Sciences Journal 49(4): 591-610.

Koutsoyiannis, D. and G. Baloutsos (2000). Analysis of a long record of annual maximum rainfall in Athens, Greece, and design rainfall inferences. Natural Hazards 22(1): 31-51.

Koutsoyiannis, D., D. Kozonis and A. Manetas (1998). A comprehensive study of rainfall intensity-duration-frequency relationships. Journal of Hydrology 206(1-2): 118-135.

Leadbetter, M. R. (1983). Extremes and local dependence in stationary sequences. Zeit. Wahrscheinl. -theorie 65: 291.

Llasat, M.C. (2001). An objective classification of rainfall events on the basis of their convective features. Application to rainfall intensity in the north-east of Spain. International journal of climatology, 21: 1385-1400.

Mann, HB. (1945). Nonparametric tests against trend. Econometrica. 13:245-259.

Nadarajah, S., C. W. Anderson and J. A. Tawn (1998). Ordered multivariate extremes. J.R. Statistic. Soc. B 60(2): 473-496.

Pickands, J. (1981). Multivariate extreme value distributions. Bulletin of the International Statistical Institute, Proceedings of the 43rd Session, Buenos Aires.

Pickands, J. (1989). Multivariate negative exponential and extreme value distributions. Extreme Value Theory: Proceedings, Oberwolfach.

Renard, B., V. Garreta and M. Lang (2006). An empirical comparison of MCMC methods used in bayesian inference. Application for regional trend detection. Water Resources Research, in press.

Robinson, M. E. and J. A. Tawn (2000). Extremal analysis of processes sampled at different frequencies. Journal of the Royal Statistical Society Series B-Statistical Methodology 62: 117-135.

Sisson, S.A., L.R. Pericchi, S.G. Coles (2006). A case for a reassessment of the risks of extreme hydrological hazards in the Caribbean. Stochastic Environmental Research and Risk Assessment 20(4): 296-306.

van Montfort, M. A. J. (1997). Concomitants of the Hershfield factor. Journal of Hydrology 194(1-4): 357-365.

Veneziano, D. and P. Furcolo (2002). Multifractality of rainfall and scaling of intensity-duration-frequency curves. Water Resources Research 38(12): 421-4212.

Weiss, L. L. (1964). Ratio of true to fixed-interval maximum rainfall. Journal of the Hydraulic Division of ASCE 90: 77-82.

32