

Bayesian comparison of different rainfall depth-duration-frequency relationships

Aurélie Muller ⁽¹⁾, 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):

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

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:

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

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 $I_d(\cdot)/b(d)$.

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:

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

The parameters α_d , β_d , k_d of the distribution of H_d are simply expressed with α , β , k , θ , η :

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

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):

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

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 $\{Z_i, i=1, 2, \dots\}$ 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, \dots is said to be stationary if, for any subset of integers $\{i_1, \dots, i_k\}$, and any integer m , the joint distributions of $(Z_{i_1}, \dots, Z_{i_k})$ and of $(Z_{i_1+m}, \dots, Z_{i_k+m})$ are identical. The strong mixing dependence condition limits the degree of long-term dependence at extreme levels and is defined by: for all $i_1 < \dots < i_p < j_1 < \dots < j_q$ with $j_1 - i_p > l_n$

$$\left| P(Z_{i_1} \leq u_n, \dots, Z_{i_p} \leq u_n, Z_{j_1} \leq u_n, \dots, Z_{j_q} \leq u_n) - P(Z_{i_1} \leq u_n, \dots, Z_{i_p} \leq u_n)P(Z_{j_1} \leq u_n, \dots, Z_{j_q} \leq u_n) \right| \leq \alpha(n, l_n) \quad (7)$$

where $\alpha(n, l_n) \rightarrow 0$ for a sequence l_n such that $l_n \rightarrow 0$ as $n \rightarrow \infty$, 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:

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

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:

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

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:

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

The asymptotic value $EI = \lim_{u \rightarrow \infty} \hat{EI}(u, r_u)$ can be approached using a sequence of thresholds (u_1, \dots, u_n) that increase with n , and r_u such that $r_u/u \rightarrow 0$ as $u \rightarrow \infty$.

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):

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

The daily data are included in the model (4). A new model is then defined, whose parameters are $\alpha_D, \beta_D, k_D, \Theta, \theta$ 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:

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

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,

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

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):

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

- for long durations, $d \geq d_b$, and $\alpha_D, \beta_D, k_D, \Theta, \theta, \eta$ denote the parameters of eq. (12), for example if $k_d \neq 0$:

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

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

$$\beta_s = \alpha_s \beta_{24} / \alpha_{24} \quad (16)$$

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

With two ranges of durations, eight parameters $(\alpha_D, \beta_D, k_D, \Theta, \theta, \eta, \alpha_s$ and $\theta_s)$ 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 $(\alpha_s$ 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

