Modeling pairwise dependencies in precipitation intensities

M. Vrac, P. Naveau, P. Drobinski

Nonlin. Processes Geophys., 14, 789–797, 2007
www.nonlin-processes-geophys.net/14/789/2007/

Nonlinear Processes in Geophysics

Modeling pairwise dependencies in precipitation intensities

M. Vrac[1], **P. Naveau**[1], **and P. Drobinski**[2]

[1]Laboratoire des Sciences du Climat et de l'Environnement, IPSL-CNRS, Gif-sur-Yvette, France
[2]Service d'Aéronomie, IPSL-CNRS, Université Pierre et Marie Curie, Paris, France

Received: 29 August 2007 – Revised: 19 November 2007 – Accepted: 19 November 2007 – Published: 5 December 2007

Abstract. In statistics, extreme events are classically defined as maxima over a block length (e.g. annual maxima of daily precipitation) or as exceedances above a given large threshold. These definitions allow the hydrologist and the flood planner to apply the univariate Extreme Value Theory (EVT) to their time series of interest. But these strategies have two main drawbacks. Firstly, working with maxima or exceedances implies that a lot of observations (those below the chosen threshold or the maximum) are completely disregarded. Secondly, this univariate modeling does not take into account the spatial dependence. Nearby weather stations are considered independent, although their recordings can show otherwise.

To start addressing these two issues, we propose a new statistical bivariate model that takes advantages of the recent advances in multivariate EVT. Our model can be viewed as an extension of the non-homogeneous univariate mixture. The two strong points of this latter model are its capacity at modeling the entire range of precipitation (and not only the largest values) and the absence of an arbitrarily fixed large threshold to define exceedances. Here, we adapt this mixture and broaden it to the joint modeling of bivariate precipitation recordings. The performance and flexibility of this new model are illustrated on simulated and real precipitation data.

1 Introduction

There exists a wide range of distribution families to statistically model rainfall intensities. For example, Katz (1977), Vrac et al. (2007), and Wilks (2006) argued that most of the precipitation variability can be approximated by Gamma distributions. However, it is also well known (e.g. Katz et al., 2002) that the tail of the Gamma distribution can be too light to capture heavy rainfall intensities. This leads to the underestimation of return levels and others quantities linked to high percentiles of precipitation amounts. Consequently, the societal and economical impacts associated with heavy rains (e.g., floods) can be miscalculated. To solve this issue, an increasingly popular approach in hydrology (e.g. Katz et al., 2002) is to disregard small precipitation values and to focus only on the largest rainfall amounts. The advantage of this strategy is that an elegant mathematical framework called *Extreme Value theory* (EVT) developed in 1928 by Fisher and Tippett (1928) and regularly updated during the last decades (e.g. Resnick, 2007; De Haan and Ferreira, 2006) dictates the distribution of heavy precipitation. More specifically, EVT states that rainfall exceedances, i.e. amounts of rain greater than a given threshold u, can be approximated by a Generalized Pareto Distribution (GPD) if the threshold and the number of observations are large enough.

Past studies (e.g. Katz et al., 2002; Naveau et al., 2005; Cooley et al., 2007) have illustrated how univariate EVT can be applied to climate and hydrology sciences. Recently, there have been a few attempts at not only modeling extremes but the full range of the observations. For example, Frigessi et al. (2002) proposed a univariate mixture model with three components. The first one represents the bulk of the distribution and the second one focuses on the upper tail (i.e. extremes). The third one corresponds to a weight function that makes the connection between the first two parts of the Frigessi model. Vrac and Naveau (2007) applied this univariate model to downscale precipitation over the region of Illinois. Other univariate mixture models that take into account the EVT exist. Carreau and Bengio (2006) investigated a model that combines a non-parametric approach (neural networks) with EVT densities. The research developed therein can be viewed as an extension of these past approaches. We keep the idea of working with a mixture model that can characterize the full range of rainfall observations but we move from a univariate framework to a bivariate space. Such an

Correspondence to: M. Vrac
(mathieu.vrac@lsce.ipsl.fr)

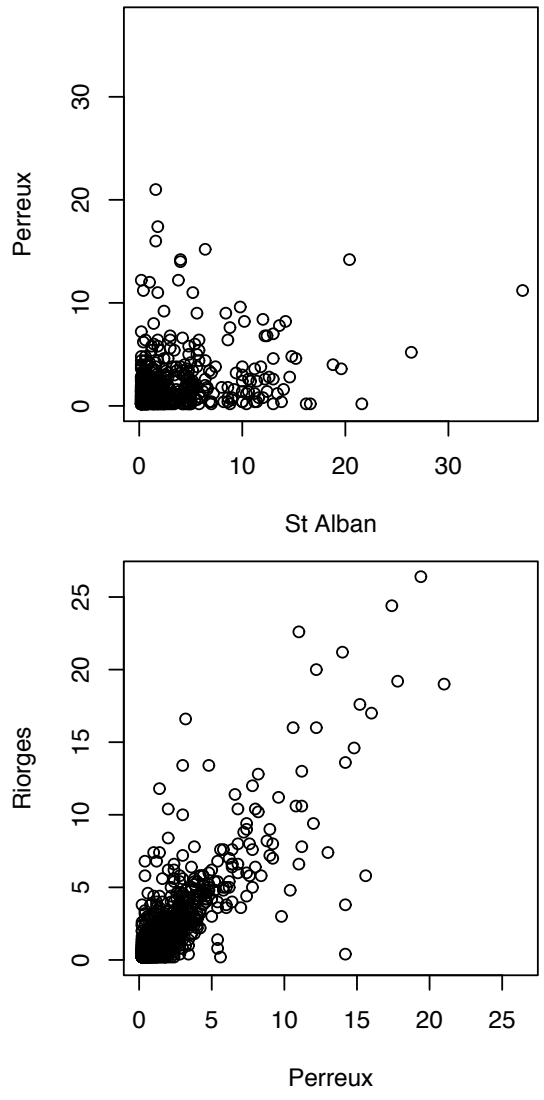

Fig. 1. The top panel shows the scatterplot between daily precipitation data in mm/day recorded at two French stations named St Alban and Perreux which are fairly far away from each other (about 300 km). In contrast, the bottom panel displays the same type of scatterplot but between two nearby stations, Perreux and Riorges (about 10 km).

extension can be trivial for some distributions. This is not the case here because the EVT is different in the 2-D case. While univariate EVT imposes a parametric form for the margins, bivariate EVT forces the dependence structure among extremes to be non-parametric and choices have to be made to deal with this problem (e.g. see Chapter 8 of Coles, 2001). In addition, modeling the transition from the bulk of the distribution to the extreme values represents an additional challenging task.

The paper is organized as follows. In Sect. 2 we give a brief overview of the precipitation measurements that will be used to illustrate and validate our approach. We also recall a few basic concepts used in bivariate EVT. Section 3 is divided into two parts. Firstly, we treat the univariate case by recalling the basic principles of the Frigessi mixture model Frigessi et al. (2002) and by applying it to univariate precipitation recordings. Secondly, we propose a bivariate model that combines the advantages of the Frigessi univariate mixture model and the principles of bivariate EVT. Section 4 focuses on applications. Our approach is tested on simulated data and applied to precipitation measurements. Finally, we summarize our results and discuss some future research directions in Sect. 5.

2 Precipitation data

To exemplify the methodologies proposed in this paper, we will analyze rainfall measurements coming from three weather stations located near the cities of St Alban, Perreux and Riorges that belong to the French Mediterranean region. In this section we present the basics statistical properties of these observations. The daily time series cover the time period from 1 January 1994 to 31 December 2004.

The top panel of Fig. 1 shows the scatterplot between daily precipitation data in mm/day recorded at two stations named St Alban and Perreux which are fairly far away from each other (about 300 km). In contrast, the bottom panel of Fig. 1 displays the same type of scatterplot but between two nearby stations, Perreux and Riorges (about 10 km). As expected, this figure indicates that nearby stations can provide strongly dependent recordings. In this example, this dependence still exists for large rainfall amounts. Consequently, the analysis of extremes should be improved if this dependence is taken into account. We have not yet tried to define the term "bivariate extreme event". To clarify this expression, we need to introduce a few notations. Let R_1 and R_2 be two positive, continuous and heavy tailed random variables that represent the rainfall recordings at two stations, say station 1 and 2. Here "heavy tailed" means that the upper tail distribution of R_1 and R_2 can be considered of the form of a power law, i.e. proportional to $x^{-1/\xi}$ for some $\xi>0$ as x gets large. This assumption is reasonable for our precipitation measurements because very strong rainfall occur frequently in the Mediterranean region due to local thunderstorms. Concerning the definition of a bivariate extreme event, does it mean that both R_1 and R_2 should be large at the same time or is it enough that only one of the two variables is very large? We opt for the latter case and we introduce the radius $|\mathbf{R}|=R_1+R_2$ to say that a bivariate extreme event occurs whenever the radius $|\mathbf{R}|$ is large. At this stage, we also need to better quantify the type of dependence among extremes that we have observed in the bottom panel of Fig. 1. This is a difficult task and a lot of schemes have been proposed to capture the relationship among extremes (e.g. Resnick, 2007; De Haan and Ferreira, 2006; Beirlant et al., 2004). Having already defined the radius $|\mathbf{R}|$, we recall the well-known bivariate EVT Pickands'

coordinates, a radius and a pseudo "angle" ω (e.g. see Chapter 8 of Coles, 2001) defined by

$$\omega = \frac{R_2}{|\mathbf{R}|}, \text{ with } |\mathbf{R}| = R_1 + R_2. \quad (1)$$

The effect of transforming the vector $(\boldsymbol{R}_1, \boldsymbol{R}_2)$ into $(|\mathbf{R}|, \omega)$ defined by Eq. (1) will be illustrated in Sect. 4 on simulated and real precipitation data. Because the angle takes its values between zero and one, one classical model is the Beta probability density function

$$b_{\boldsymbol{\beta}}(\omega) = \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)}\, \omega^{\beta_1 - 1}\, (1-\omega)^{\beta_2 - 1}, \quad (2)$$

with $\omega \in [0, 1]$ and where $\Gamma(.)$ is the Gamma function and β_i are positive reals. The Beta density offers a wide range of density shape while keeping the number of parameters under control. It also has the advantage to have well-known properties.

From a theoretical point of view, it is interesting to see that the joint probability of the angle and the radius $(\omega, |\mathbf{R}|)$ given that $|\mathbf{R}| > x$ for some large x can be written as

$$\frac{\mathbb{P}(|\mathbf{R}| > rx \text{ and } \omega \in [a, b])}{\mathbb{P}(|\mathbf{R}| > x)} = \frac{\mathbb{P}(|\mathbf{R}| > rx)}{\mathbb{P}(|\mathbf{R}| > x)} \mathbb{P}(\omega \in [a, b])$$

if $0 < a < b < 1$ and $|\mathbf{R}|$ is independent of ω. In addition, if the radius is assumed to be regularly varying with index $-1/\xi$, i.e.,

$$\lim_{x \to \infty} \frac{\mathbb{P}(|\mathbf{R}| > rx)}{\mathbb{P}(|\mathbf{R}| > x)} = c\, r^{-1/\xi}, \text{ for some constant } c > 0,$$

then it follows

$$\lim_{x \to \infty} \frac{\mathbb{P}(|\mathbf{R}| > rx \text{ and } \omega \in [a, b])}{\mathbb{P}(|\mathbf{R}| > x)} = \mathrm{c}\, r^{-1/\xi}\, \mathbb{P}(\omega \in [a, b]).$$

Hence, defining pseudo-coordinates is closely linked to the concept of *regular variation* and the latter has been increasingly popular in multivariate EVT during the last decades, specially in time series analyses for heavy-tailed models (e.g. Resnick, 2007; De Haan and Ferreira, 2006; Beirlant et al., 2004).

3 Our statistical models

3.1 The univariate case

According to basic univariate EVT (e.g. Coles, 2001; Embrechts et al., 1997), the probability that large rainfall amount, say the random variable R, is larger than the real r given that R is already larger than a fixed high threshold u can be approximated by a Generalized Pareto Distribution (GPD) tail defined as

$$\mathbb{P}(R > r | R > u) = \left(1 + \xi \frac{r-u}{\sigma}\right)_+^{-1/\xi}, \quad (3)$$

where $a_+ = \max(a, 0)$ and $\sigma > 0$ represents the scale parameter. The shape parameter ξ describes the GPD tail behavior. If ξ is negative, the upper tail is bounded. If ξ is zero, this corresponds to the case of an exponential distribution (all moments are finite). If ξ is positive, the upper tail is still unbounded but higher moments eventually become infinite. These three cases are termed "bounded", "light-tailed", and "heavy-tailed", respectively. The flexibility of the GPD to describe three different types of tail behavior makes it a universal tool for modeling exceedances. In our case, we assume that our rainfall data are heavy-tailed, i.e. ξ is assumed to be positive. We also note that the GPD belongs to the family of regularly varying function introduced at the end of Sect. 2

A possible drawback of EVT is that the GPD only models data exceeding a given high threshold, and one can wonder how to model the remaining data (i.e. lower than the threshold) or equivalently how to deal with the entire range of data. To answer these questions, Frigessi et al. (2002) proposed the following mixture model

$$f_{\boldsymbol{\theta}}(r) = c_{\boldsymbol{\theta}} \left[(1 - p_{\mu,\tau}(r))\, g_{\boldsymbol{\gamma}}(r) + p_{\mu,\tau}(r)\, h_{\sigma,\xi}(r) \right], \quad (4)$$

where $c_{\boldsymbol{\theta}}$ is a normalizing constant, $\boldsymbol{\theta} = (\mu, \tau, \boldsymbol{\gamma}, \sigma, \xi)$ encapsulates the vector of unknown parameters, $g_{\boldsymbol{\gamma}}$ corresponds to a light-tailed density with parameters $\boldsymbol{\gamma}$, the function $h_{\sigma,\xi}$ represents a heavy tailed Generalized Pareto (GP) density with threshold $u = 0$. One of the most interesting aspect of Eq. (4) is the weight function $p_{\mu,\tau}(.)$ defined by

$$p_{\mu,\tau}(r) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{r-\mu}{\tau}\right). \quad (5)$$

Because this weight function is non-decreasing, takes values in $[0, 1]$ and tends to 1 as r goes to ∞, it can play the role of an unsupervised threshold selection algorithm. This transition can be interpreted as the contribution of the GPD to the overall fit. For our case study, heavy rains are represented by the heavy tailed GPD density $h_{\sigma,\xi}(r)$, while low and medium precipitation are modeled by the light distribution $g_{\boldsymbol{\gamma}}$. Concerning the weight function $p_{\mu,\tau}(r)$, our past work (Vrac and Naveau, 2007) suggested that τ is almost equal to zero for rainfall data. This is also the case for our Mediterranean precipitation. Hence, τ is fixed to zero in the rest of this article. In other words, the weight function is set to the limit of Eq. (5) as τ tends to 0, i.e.

$$p_{\mu,0}(r) = \begin{cases} 1 & \text{, if } \mu \leq r, \\ 0 & \text{, otherwise.} \end{cases}$$

This special case has the advantage of reducing the number of parameters, although, for other applications, the more general form of $p_{\mu,\tau}(r)$ defined by Eq. (5) may be more appropriate. Compared to classical EVT where the threshold u is considered fixed, the parameter μ is not predetermined and has to be estimated.

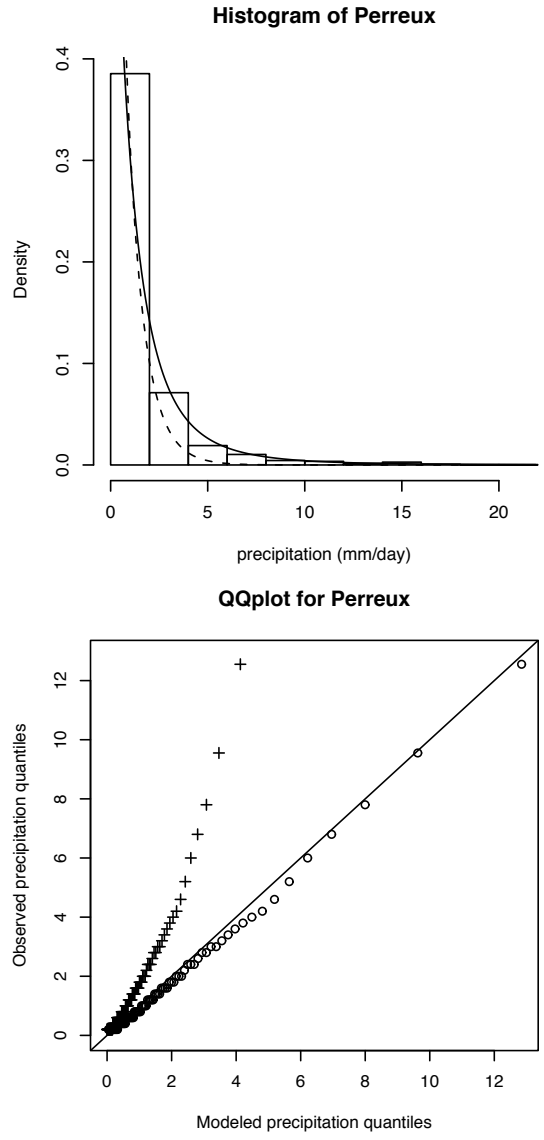

Fig. 2. Top panel: Histogram of the positive precipitation measurements from Perreux. Dashed and solid lines correspond to a Gamma fit and a mixture (Eq. 4) fit, respectively. Bottom panel: QQplots of the positive Perreux precipitation data (in mm/day) with the two fitted distributions. The x-axis corresponds to the expected quantiles and the y-axis represents the observed quantiles. The crosses and the circles correspond to the estimated Gamma and mixture quantiles, respectively.

While Frigessi et al. (2002) chose to parametrize the light density $g_{\boldsymbol{\gamma}}$ as a Weibull density in their fire loss application, we opt to represent $g_{\boldsymbol{\gamma}}$ by a Gamma density for our precipitation data. This choice appears to be in compliance with the past hydrological literature on precipitation modeling (e.g. Katz, 1977; Vrac et al., 2007; Wilks, 2006). The Gamma density is defined as

$$g_{\boldsymbol{\gamma}}(r) = \frac{\gamma_2^{\gamma_1}}{\Gamma(\gamma_1)} r^{\gamma_1 - 1} \exp(-r\gamma_2), \text{ for } r > 0. \qquad (6)$$

The mixture model between a light Gamma density and a heavy-tailed GPD has already been applied to downscale rainfall data over the state of Illinois (USA) (Vrac and Naveau, 2007). In this past study, weather stations were considered independent in space while the parameters of the mixture model were conditioned to large scale climatic information. In this respect, the present work represents a different direction because the pairwise spatial dependence will be directly addressed in the coming sections.

To establish the superiority of the mixture model (Eq. 4) for our data over a simple Gamma density, the histogram (top panel) and the quantile-quantile plot (QQplot) (low panel) of the positive rainfall amounts recorded at the Perreux weather station are shown in Fig. 2. In the top panel of Fig. 2, the dashed and solid lines correspond to a Gamma fit and a mixture fit, respectively. This indicates that the mixture model defined by Eq. (4) provides a reasonably good fit of the core of the rainfall distribution. Concerning the extremes, the bottom panel of Fig. 2 displays a QQ plot whose x-axis corresponds to the expected quantiles and the y-axis represents the observed quantiles. The crosses and the circles correspond to the estimated Gamma and mixture quantile fits, respectively. From these QQplots, it is clear that the Gamma density can not reproduce adequately the behavior of extreme precipitation for the station of Perreux. Similar figures were obtained for our two other stations. A possible danger of the mixture model is the risk of over-parameterization because six parameters have to be estimated to fit Eq. (4). To check this point the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978) have been calculated. These two classical statistical criteria for model selection are defined as $AIC = -2L(\theta) + 2p$ and $BIC = -2L(\theta) + p\log(n)$, where $L(\theta)$ is the log-likelihood of the model to be tested, p is the number of parameters, and n is the sample size. Based on the precipitation data from the Perreux station, the AIC values for the Gamma model (Eq. 6) alone and the mixture model (Eq. 4) are 4415 and 3254 respectively, and the BIC values are 4426 and 3285 in the same order. Hence, we can conclude that the mixture model (Eq. 6) improves sufficiently the likelihood with respect to the Gamma distribution alone to be selected as the "best" model among the two models, despite its larger number of parameters.

3.2 Our bivariate extension

Following the tactics developed in the previous section, it seems reasonable to assume that the core of the bivariate precipitation random vector $(\boldsymbol{R}_1, \boldsymbol{R}_2)$ can be modeled by a bivariate Gamma random vectors. As suggested by bivariate EVT, the extreme bivariate tails behavior could be

modeled in the Pickand's coordinate system $(|\mathbf{R}|, \omega)$ defined by Eq. (1). The difficulty in modeling the entire bivariate precipitation range comes from this difference between the cartesian coordinates $(\boldsymbol{R}_1, \boldsymbol{R}_2)$ necessary to model the distribution core by a bivariate Gamma density and the Pickand's coordinate system $(|\mathbf{R}|, \omega)$ needed to take advantage of bivariate EVT. Keeping Frigessi's approach in mind, we assume that a weight function can provide an elegant transition from the core of the distribution to the upper tails. As in the univariate case, this allows us to bypass the threshold selection problem which is even more difficult to apprehend in the bivariate case. To keep the number of parameters under control, we force the weight function to be univariate and to only vary in function of the radius $|\mathbf{R}|$. This condition allows us to define the probability density distribution of the vector $(\boldsymbol{R}_1, \boldsymbol{R}_2)$ as the following mixture

$$f_{\boldsymbol{\theta}}(r_1, r_2) = c_{\boldsymbol{\theta}}\Big[\left(1 - p_{\mu,0}(|\mathbf{r}|)\right) g_{\boldsymbol{\gamma}}(r_1, r_2) + p_{\mu,0}(|\mathbf{r}|)\, h_{\sigma,\xi}(|\mathbf{r}|)\, b_{\boldsymbol{\beta}}(\omega)\Big] \quad (7)$$

where $|\mathbf{r}|=r_1+r_2$, $\omega=r_2/|\mathbf{r}|$, c_θ is a normalizing constant, $h_{\sigma,\xi}(.)$ corresponds to the univariate GP density with parameters (σ, ξ) and threshold $u=0$, the univariate function $b_{\boldsymbol{\beta}}(.)$ represents a Beta probability density function with parameters $\boldsymbol{\beta}$ and $g_{\boldsymbol{\gamma}}(.,.)$ is a bivariate Gamma probability density function. There exists a wide variety of bivariate Gamma distribution, see the book by Kotz et al. (2000) for a review. In this study, we opt for the Cheriyan and Ramabhadran family (see Kotz et al., 2000) because of its large correlation range and its simplicity in terms of simulation and estimation. Each component of a bivariate Cheriyan and Ramabhadran vector is distributed following a Gamma distribution, and the components depend on each other by means of an auxiliary Gamma distributed variable. The joint distribution $g_{\boldsymbol{\gamma}}(r_1, r_2)$ is defined as

$$\int_0^{\min(r_1,r_2)} \frac{e^{-z} z^{\gamma_0-1}}{\Gamma(\gamma_0)} \prod_{i=1}^{2} \left[\frac{e^{-(r_i-z)}(r_i-z)^{\gamma_i-1}}{\Gamma(\gamma_i)} \right] dz, \quad (8)$$

where $\boldsymbol{\gamma}=(\gamma_0, \gamma_1, \gamma_2)$.

With our general mixture described by Eq. (7) whose elements are defined by the Eqs. (2), (3), (5) and (8), we can now investigate the practicability of such a model on simulated and real data.

4 Simulations, estimation and applications

4.1 Simulating bivariate samples from density (Eq. 7)

Our simulation algorithm can be viewed as an extension of the 1-D scheme suggested by Frigessi et al. (2002). It can be summarized by the following steps.

1. Draw U uniformely on $[0, 1]$.
2. If $U<1/2$, then sample $\mathbf{r}=(r_1, r_2)$ from g_γ defined by Eq. (8); return $\mathbf{r}$ with probability $1-p_{\mu,0}(|\mathbf{r}|)$ and stop; or, with probability $p_{\mu,0}(|\mathbf{r}|)$, return to 1.
3. If $U\geq 1/2$, then sample $|\mathbf{r}|$ from a GP density $h_{\sigma,\xi}$ and ω from $b_{\boldsymbol{\beta}}$ defined by Eq. (2); return $r_1=|\mathbf{r}|\times\omega$ and $r_2=|\mathbf{r}|-|\mathbf{r}|\times\omega$ with probability $p_{\mu,0}(|\mathbf{r}|)$ and stop; or, with probability $1-p_{\mu,0}(|\mathbf{r}|)$, return to 1.

In step 2 of this simulation scheme, a couple (r_1, r_2) has to be sampled from Eq. (8). By definition of the bivariate Cheriyan and Ramabhadran Gamma distribution, one can simulate (r_1, r_2) by first generating three independent univariate standard Gamma random variables (Y_0, Y_1, Y_2) with parameters γ_0, γ_1, and γ_2, respectively. Then, the sums $r_i=y_0+y_i$ $(i=1, 2)$ give the appropriate dependence between r_1 and r_2 (see Kotz et al., 2000).

We would like to explore two types of dependence (weak and strong) for two parts of the distribution (its core and its extremes). This provides four possible combinations. Hence, four samples of 1000 realizations are generated according to density (Eq. 7). These simulations have five common parameters ($\gamma_1=\gamma_2=0.3$, $\mu=2$, $\xi=0.8$ and $\sigma=0.9$) and different γ_0, β_1 and β_2 parameters. To inject a weak dependence (correlation<0.1) in the bivariate Gamma part of Eq. (7), we set $\gamma_0=10^{-3}$. In contrast, a strong dependence (correlation>0.9) in the bivariate Gamma is obtained by fixing $\gamma_0=3$. Concerning the extremes and the GPD, we also have two cases: $\beta_1=\beta_2=0.05$ and $\beta_1=\beta_2=5$. The latter provide a strong dependence in the upper tail, while the former produces a weak one. The next step is to determine if we can adequately estimate the parameters of these four combinations that represent a wide variety of dependencies.

4.2 The estimation procedure

Our bivariate model defined by Eq. (7) contains eight parameters $\boldsymbol{\theta}=(\gamma_0, \gamma_1, \gamma_2, \mu, \beta_1, \beta_2, \xi, \delta)$. A direct estimation of these parameters by a maximum likelihood approach can be tricky and computationally expensive. One of the main hurdles is the estimation of the parameter μ in the weight function $p_{\mu,0}$. To circumvent this difficulty, we develop an iterative estimation algorithm in which μ is updated at the end of each estimation cycle. To initialize our procedure, a first guest for μ is needed and it is set to a rather low value. This first estimate of μ is called $\hat{\mu}_{\text{first}}$ and is set to, say, the 75th percentile of the radius $|\mathbf{r}|=r_1+r_2$ from the sample under study. Then, we implement the following procedure.

(a) For all pairs (r_1, r_2) such that the radius $|\mathbf{r}|$ is smaller than $\hat{\mu}_{\text{first}}$, we estimate the parameters $\boldsymbol{\gamma}$ of the Cheriyan and Ramabhadran's bivariate Gamma distribution $g_{\boldsymbol{\gamma}}$, by maximizing a bivariate Gamma likelihood.

(b) For all $|\mathbf{r}|$ larger than $\hat{\mu}_{\text{first}}$, we estimate the GP density $h_{\sigma,\xi}$ parameters and the Beta density $b_{\boldsymbol{\beta}}$ to the $\omega=r_2/|\mathbf{r}|$ values by maximizing their respective likelihoods.

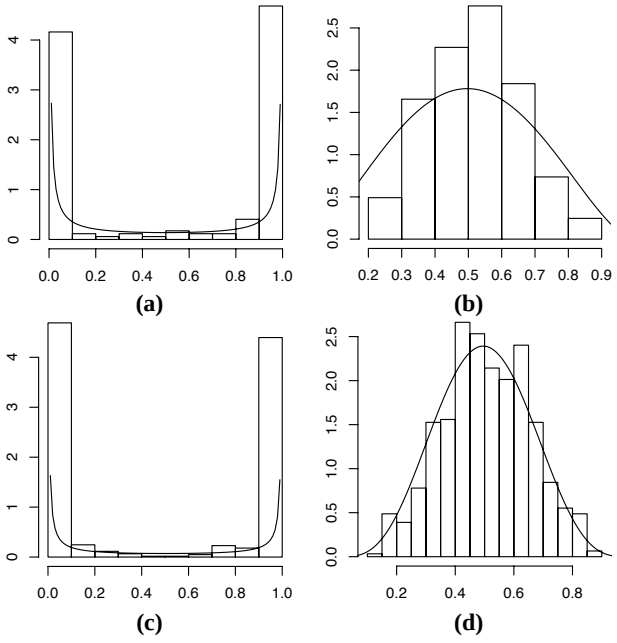

Fig. 3. Simulated data: histograms and its fitted Beta densities (solid lines) of the $\omega = r_2/(r_1+r_2)$ values conditionally on $|\boldsymbol{r}|>\hat{\mu}$. A U-shape histogram of this random variable indicates a strong independence between extreme rainfalls. On the opposite, a histogram centered around 0.5 shows a strong dependence. The left panels correspond to the two simulations with a weak dependence in the upper tail of the bivariate random variable defined by Eq. (7), i.e. β_1 and β_2 were set to 0.05 in Eq. (2). The right panels represent the two simulations with strong dependence, i.e. β_1 and β_2 were set to 5 in Eq. (2). The difference between the upper and lower panels resides in the pairwise dependence within the Gamma part of Eq. (7), weak for panels (a) and (b) and strong for panels (c) and (d), i.e. γ_0 was either set to 10^{-3} or to 3 in Eq. (8), respectively.

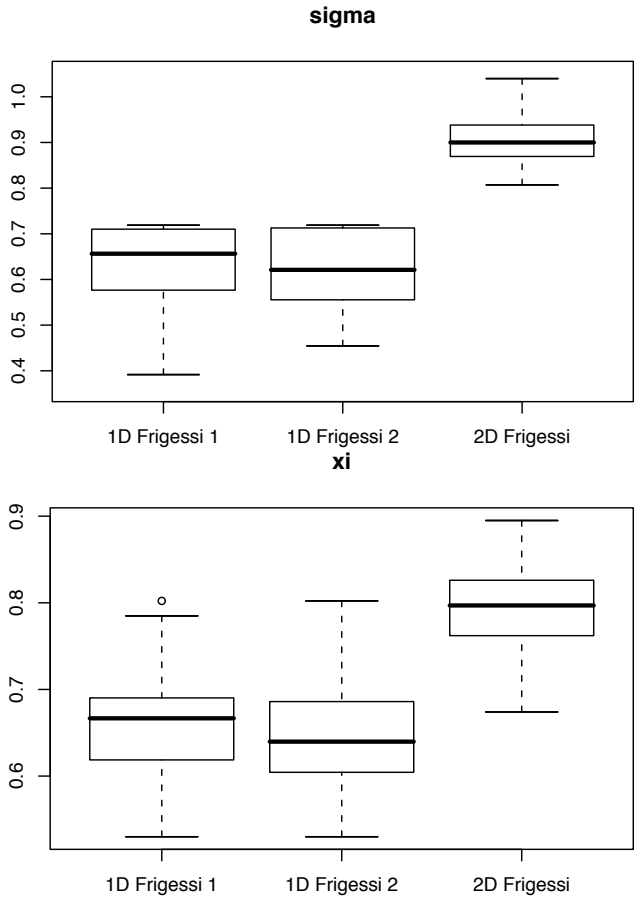

Fig. 4. Simulated data: The scale and shape parameter σ and ξ (set to the values 0.9 and 0.8, respectively) have been estimated from 50 realizations of length 1000 from our mixture model with a strong dependence in its core and tail. The so-called "1D Frigessi 1" and "1D Frigessi 2" case corresponds to the boxplots obtained when R_1 and R_2 are wrongly assumed to be independent and a classical univariate approach is applied on each rainfall compoment. The boxplot "2D Frigessi" displays the estimation result when the correct model is assumed.

(c) Based on the parameter estimates from steps (a) and (b), we estimate a new value for μ, say $\hat{\mu}_{\text{updated}}$, by fitting the full density (Eq. 7) to the whole sample through maximum likelihood estimation. All parameters but μ are fixed.

(d) Go back to step (a) as $\hat{\mu}_{\text{updated}}$ becomes $\hat{\mu}_{\text{first}}$ until $\hat{\mu}_{\text{updated}}$ and $\hat{\mu}_{\text{first}}$ are close enough.

The stopping criterion that defines the term "close enough" in step (d) translates into the condition $(\hat{\mu}_{\text{updated}}-\hat{\mu}_{\text{first}})/\hat{\mu}_{\text{first}}<0.02$. In this procedure, the final results can depend on the initial μ_{first} value. To overcome this potential weakness, several initial μ_{first} values are tested (the 70th, 75th, 80th, 85th, 90th, and 95th percentiles of the observed radius values), providing several results (usually equivalent) and the parameters associated with the highest log-likelihood are retained.

To assess the quality of our estimation procedure, we apply it to the four samples introduced at the end of Sect. 4.1. In Fig. 3, we look at the histograms of the ω values conditionally on $|\boldsymbol{r}|>\hat{\mu}$ and the solid lines correspond to the fitted Beta distributions. According to bivariate EVT (e.g. see Chapter 8 of Coles, 2001), a U-shape histogram of this random variable indicates a strong independence between extreme rainfalls. Conversely, a histogram centered around 0.5 shows a strong dependence. In Fig. 3, the left panels correspond to the two simulations with a weak dependence in the upper tail of the bivariate random variable defined by Eq. (7), i.e. β_1 and β_2 were set to 0.05 in Eq. (2). The right panels represent the two simulations with strong dependence, i.e. β_1 and β_2 were set to 5 in Eq. (2). The difference between the upper and lower panels resides in the pairwise dependence within the Gamma part of Eq. (7), weak for panels (a) and (b) and strong for panels (c) and (d), i.e. γ_0 was either set to 10^{-3} or to 3 in

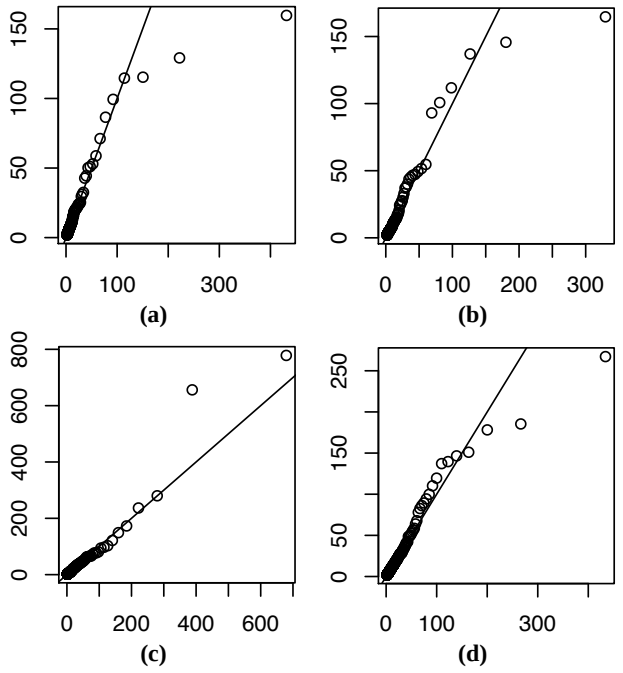

Fig. 5. Simulated data: QQplots of the radius variable $|\boldsymbol{r}|$ conditionally on $|\boldsymbol{r}|>\mu$. The x-axis corresponds to the true quantiles while the y-axis represents the estimated quantiles. See the caption of Fig. 3 to understand the meaning of the four different panels.

Eq. (8), respectively. As expected, the dependence among low and medium values generated by the Gamma density does not play a strong role in the upper tail. This explains the small difference in the histogram shapes between the upper and lower panels. The dissimilarity between the left and right panels is due to the strong disparity in the extreme behavior dependencies captured by the coefficients β_1 and β_2. Figure 4 compares the estimation result for $\sigma=0.9$ and $\xi=0.8$ from 50 realizations of length 1000 from our mixture model with a strong dependence in its core and tail. The so-called "1D Frigessi 1" and "1D Frigessi 2" case corresponds to the boxplots obtained when R_1 and R_2 are wrongly assumed to be independent. The boxplot "2D Frigessi" displays the estimation result when the correct model is assumed. In this latter case, the true values of σ and ξ are correctly estimated and the uncertainty spreads are reasonable. In contrast, wrongly assuming independence of R_1 and R_2 clearly underestimates the true value of σ and ξ and increases the boxplots width. This shows that applying a classical univariate approach and ignoring the dependence can mislead the practitioner.

Overall, Figs. 3 and 4 indicate three things: (1) our model is able to generate different types of dependencies in the upper tail, (2) the low and medium values do not influence the overall shape dependence in the extremes, and (3) our estimation procedure seems to work adequately.

Concerning the intensity of the extremes produced by our model and obtained by our estimation algorithm, we can observe Fig. 5 that displays four QQplots from our four samples. The x-axis corresponds to the true quantiles while the y-axis represents the estimated quantiles. These graphs indicate that very large extreme values can be adequately reproduced, e.g. see Panel (c). Still, the performance varies from samples to samples. For example, the largest value in Panel (a) is underestimated. Overall, the four fitted QQplots seem to capture most of the extreme behaviors. In order to confirm this result and to provide GPD goodness-of-fit tests, Andersen-Darling A^2 and Cramér-von Mises W^2 statistics were computed (e.g. Choulakian and Stephens, 2001). Both statistics show that the GPD fits can be considered as acceptable for a confidence level of 99% (i.e. with p-values$<$0.01).

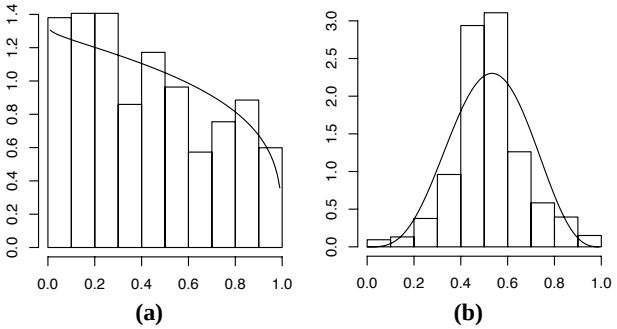

Fig. 6. Observed rainfalls: Histograms of the ω values conditionally on $|\boldsymbol{r}|>\hat{\mu}$ for the two weather stations pairs: Panel (a) the "St-Alban-Perreux" couple and Panel (b) the "Perreux-Riorges" pair. The solid lines correspond to the fitted Beta distributions.

Although this simulation study is limited to only four cases, the discrepancy among the four studied situations seems to indicate that our estimation procedure can cover a wide range of dependence cases, and therefore it can now be applied to real precipitation data.

4.3 Precipitation measurements

In Sect. 2, we described rainfall measurements from three weather stations located in the French Mediterranean region. Here, we apply our bivariate modeling strategy proposed in Sect. 3.2 to the two pairs of station presented on Sect. 2 and in Fig. 1. The estimation process detailed on Sect. 4.2 is implemented and a set of parameters for the model (7) is estimated for our two pairs of weather stations: $\theta_0=0.02$; $\theta_1=0.04$; $\theta_2=0.06$; $\mu=1.5$; $\tau=0$; $\beta_1=0.99$; $\beta_2=1.27$; $\xi=0.14$; $\sigma=6$ for the "St-Alban-Perreux" couple of stations, and $\theta_0=0.025$; $\theta_1=0.09$; $\theta_2=0.09$; $\mu=2$; $\tau=0$; $\beta_1=4.6$; $\beta_2=4.2$; $\xi=0.2$; $\sigma=4$ for the "Perreux-Riorges" couple. Based on these two sets of estimated parameters, the equivalent of Figs. 3 and 5 (histograms and Beta fit of the angle values, QQplots of the observed and fitted radius values, respectively) are shown in Figs. 6 and 7. For the former graph, the curve centered around 0.5 in panel (b) confirms the strong dependence among extremes recorded at the most nearby stations. This

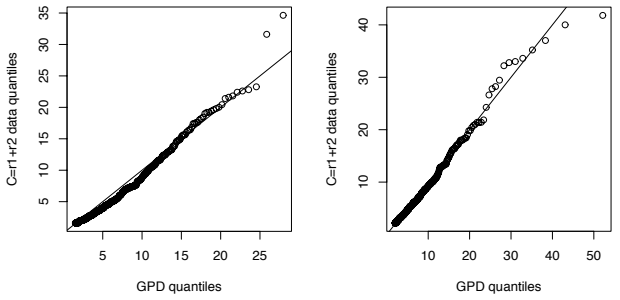

Fig. 7. Observed rainfalls: QQplots of the observed $|\boldsymbol{r}|$ quantiles conditionally on $|\boldsymbol{r}|>\hat{\mu}$ vs. estimated GPD quantiles for each pair: (a) corresponds to the couple "St-Alban-Perreux", (b) to "Perreux-Riorges".

dependence has already been observed in the bottom panel of Fig. 1. Panel (a) of Fig. 6 is more difficult to interpret. The histogram, as well as the Beta fit, seems to indicate a mild dependence, much weaker than in panel (b), but this is not the U-shape that characterizes the independence, e.g. the left panels of Fig. 3. Concerning large precipitation intensities, the QQplots in Fig. 7 indicate a good agreement between the estimated and observed quantiles for the two pairs of stations. The GPD goodness-of-fit tests performed through Andersen-Darling A^2 and Cramér-von Mises W^2 statistics (see Choulakian and Stephens, 2001), show that, as for simulated data, the GPD fits are considered as acceptable for a confidence level of 99% (i.e. with p-values$<$0.01), confirming the QQplots.

5 Conclusions and perspectives

We have presented a new statistical distribution that can model the entire range (i.e. low, medium and extreme values) of bivariate precipitation measurements. This model consists in a mixture between a bivariate Gamma distribution – representing the precipitation density core (i.e. the non extreme part) – and a product of GP and Beta densities in a Pickland's coordinates system – characterizing heavy rainfall density. The mixture is weighted through a function varying with the extremes strength within each pairwise rainfall data. A simulation scheme and an estimation procedure have been proposed and tested. Four simulated samples have been generated and studied. The dependence structure as well as the parameter values have been correctly retrieved for each simulated sample. Our estimation procedure has been applied to real precipitation measurements from three weather stations located in the South of France. Our statistical modeling confirms that nearby stations provide dependent recordings, not only for mean precipitation values but also among heavy rainfalls. This suggests that past studies that have completely ignored the spatial dependence between weather stations may have led to imprecise statistical outputs, specially in terms of extreme value analysis. More research is needed to extend our pairwise rainfall model into a fully multivariate framework. Besides the estimation problem beyond the 2-D case, the difficulty resides in proposing a parsimonious model that can be based on multivariate EVT and also offer enough flexibility to represent the dependencies within small precipitation, heavy rainfalls and between both.

Acknowledgements. This work was supported by the european E2-C2 grant, the National Science Foundation (grant: NSF-GMC (ATM-0327936)), by The Weather and Climate Impact Assessment Science Initiative at the National Center for Atmospheric Research (NCAR) and the ANR-AssimilEx project. The authors would also like to credit the contributors of the R project.

Edited by: B. D. Malamud
Reviewed by: two anonymous referees

References

Akaike, H.: A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19, 716–723, 1974.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J.: Statistics of Extremes: Theory and Applications, Wiley Series in Probability and Statistics, 2004.

Carreau, J. and Bengio, Y.: A hybrid Pareto model for asymmetric fat-tail data, Technical report 1283, Dept. IRO, Université de Montréal, 2006.

Choulakian, V. and Stephens, M. A.: Goodness-of-fit Tests for the Generalized Pareto Distribution, Technometrics, 43, 478–484, 2001.

Coles, S. G.: An introduction to statistical modeling of extreme values, Springer Series in Statistics, 2001.

Cooley D., Nychka, D., and Naveau, P.: Bayesian Spatial Modeling of Extreme Precipitation Return Levels, J. Am. Stat. Assoc., 102(479), 824–840, 2007.

De Haan, L. and Ferreira, A.: Extreme Value Theory: An Introduction, Springer Series in Operations Research and Financial Engineering, 2006.

Embrechts, P., Klüppelberg, C., and Mikosch, T.: Modelling Extremal Events for Insurance and Finance, Applications of Mathematics, vol. 33, Springer-Verlag, Berlin, 1997.

Fisher, R. A. and Tippett, L. H. C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proceedings of the Cambridge Philosophical Society, 24, 180–190, 1928.

Frigessi, A., Haug, O., and Rue, H.: A dynamic mixture model for unsupervised tail estimation without threshold selection, Extremes, 5, 219–235, 2002.

Katz, R., Parlange, M., and Naveau, P.: Extremes in hydrology, Adv. Water Resour., 25, 1287–1304, 2002.

Katz, R. W.: Precipitation as a chain-dependent process, J. Appl. Meteorol., 16, 671–676, 1977.

Kotz, S., Balakrishnan, N., and Johnson, N. L.: Continuous multivariate distributions, vol. 1, Models and applications, New York, Wiley, 2000.

Naveau, P., Nogaj, M., Ammann, C., Yiou, P., Cooley, D., and Jomelli, V.: Statistical methods for the analysis of climate extremes, C. R. Geoscience, 337, 1013–1022, 2005.

Resnick, S.: Heavy-Tail Phenomena: Probabilistic and Statistical Modeling, Springer Series in Operations Research and Financial Engineering, 2007.

Schwarz, G.: Estimating the dimension of a model, Ann. Statist., 6, 461–464, 1978.

Vrac, M. and Naveau, P.: Stochastic downscaling of precipitation: From dry events to heavy rainfalls. Water Resour. Res., 43, W07402, doi:10.1029/2006WR005308, 2007.

Vrac, M., Stein, M., and Hayhoe, K.: Statistical downscaling of precipitation through a nonhomogeneous stochastic weather typing approach. Climate Research, 34, 169–184, doi:10.3354/cr00696, 2007.

Wilks, D.: Statistical methods in the atmospheric sciences (second edition), Elsevier, Oxford, 2006.