Flexible semiparametric Generalized Pareto modeling of the entire range of rainfall amount

Patricia Tencaliec, Anne-Catherine Favre, Philippe Naveau, Clémentine Prieur, Gilles Nicolet

Flexible semiparametric Generalized Pareto modeling of the entire range of rainfall amount

Tencaliec, P.[1], Favre, A.-C.[2], Naveau, P.[3], Prieur, C.[1], and Nicolet, G.[2]

[1]Univ. Grenoble Alpes, CNRS, INRIA, LJK, F-38000 Grenoble, France
[2]Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, F-38000 Grenoble, France
[3]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CNRS-CEA-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France

Abstract

Precipitation amounts at daily or hourly scales are skewed to the right and heavy rainfall is poorly modeled by a simple gamma distribution. An important, yet challenging topic in hydrometeorology is to find a probability distribution that is able to model well low, moderate and heavy rainfall. To address this issue, we present a semiparametric distribution suitable for modeling the entire-range of rainfall amount. This model is based on a recent parametric statistical model called the class of Extended Generalized Pareto Distributions (EGPD). The EGPD family is in compliance with Extreme Value Theory for both small and large values, while it keeps a smooth transition between these tails and bypasses the hurdle of selecting thresholds to define extremes. In particular, return levels beyond the largest observation can be inferred. To add flexibility to this EGPD class, we propose to model the transition function in a nonparametric fashion. A fast and efficient nonparametric scheme based on Bernstein polynomial approximations is investigated. We perform simulation studies to assess the performance of our approach. It is compared to two parametric models: a parametric EGPD and the classical Generalized Pareto Distribution (GPD), the latter being only fitted to excesses above a high threshold. We also apply our semiparametric version of EGPD to a large network of 180 precipitation time series over France.

Keywords— precipitation, Extreme Value Theory, Extended Generalized Pareto Distribution, semiparametric, Bernstein polynomials, maximum likelihood estimator

1 Introduction

Modeling the distribution of precipitation data is needed in many applications regarding water resources management, design, or planning, such as urban water supplies, hydropower, forecast of flood or droughts events, irrigation systems. A first and essential step in the statistical modeling is to find probability distributions that can describe correctly the occurrences and the intensities of precipitation. As the process of rainfall occurrences is discrete, while its amount is a continuous one, the most common approach is to have a different model for these two features. In this work, we only focus on the second part, *i.e.,* the statistical modeling of strictly positive rainfall amounts, and we refer to Wilks (1999) and Apipattanavis *et al.* (2007) to model occurrence processes.

Fitting accurately the full spectrum of rainfall amounts has proven to be a challenging task, mainly due to the fact that they are heavily skewed to the right. Different distributions, such as gamma (see *e.g.*, Katz, 1977; Stern and Coe, 1984; Wilks, 1989), mixed exponential (see *e.g.*, Woolhiser and Pegram, 1979; Richardson, 1981; Wilks, 1999; Garavaglia *et al.*, 2010), Weibull (see *e.g.*, Zucchini and Adamson, 1984) or lognormal (see *e.g.*, Apipattanavis *et al.*, 2007) have been considered as possible candidates. As suggested by Vrac *et al.* (2007) and Wilks (2011), gamma and mixed exponential are typically the preferred choices, but, as pointed out by Katz *et al.* (2002), the tail of a gamma distribution can be too light to model heavy rainfall and underestimation of extreme values can occur, an undesirable feature in any hydrological risk analysis.

As mentioned by Evin *et al.* (2018), stochastic precipitation generators have become useful tools in risk assessment studies for two reasons. Realistic simulated precipitation draws are needed as inputs of conceptual hydrological models. In particular, the observed series of streamflows are too short to estimate the very high floods return levels. In this context, simple but rich probability density functions (pdf) to generate precipitation draws, extreme included, are needed. In this framework, the work in this paper can also be viewed as proposing a new and flexible tool (the precipitation building unit) to researchers interested by constructing such stochastic rainfall weather generators.

As the upper tail of the distribution holds crucial information, the Generalized Pareto Distribution (GPD) is nowadays the common choice for modeling heavy rainfall in the statistical climatological community (see, *e.g.*, Katz *et al.*, 2002; Nadarajah, 2005). GPD is defined by the cumulative distribution function (cdf) $H_\xi(x/\sigma)$ as

$$H_\xi(z) = \begin{cases} 1 - (1+\xi z)_+^{-1/\xi}, & \text{for } \xi \neq 0, \\ 1 - e^{-z}, & \text{for } \xi = 0, \end{cases}$$

where ξ is the shape parameter, $\sigma > 0$ is the scale parameter and $a_+ = \max(a, 0)$. It is mathematically justified by extreme value theory (EVT) (see, *e.g.*, Coles, 2001). In hydrology, ξ is often assumed to be non negative for daily rainfall (see, *e.g.*, Evin *et al.*, 2018). We keep this hypothesis of $\xi \geq 0$ in this work.

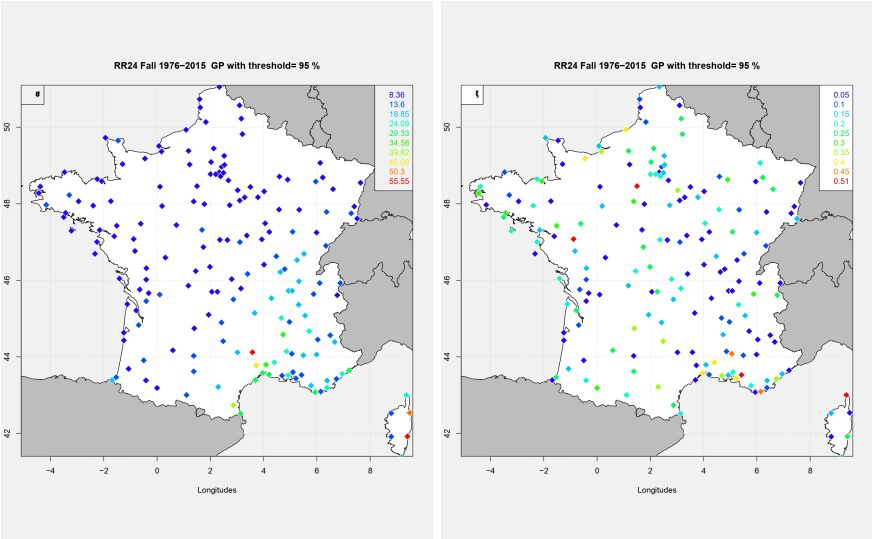
A practical limitation of the GPD is that it can be only applied to "extreme precipitation" and this leads to the question of how to set a threshold that differentiates heavy and moderate rainfall. Answering this question becomes delicate when the number of time series under study increases, say in a climate model output analysis with thousands of grid points. In such situations, graphical device tools like a Quantile-Quantile plot (QQ-plot) (see, *e.g.*, Coles, 2001; Katz *et al.*, 2002) cannot be visually checked anymore. Hence, the threshold in hydrological instances dealing with a large number of datasets is chosen arbitrarily, classically the 95% quantile of each time series.

To illustrate this issue, a classical GPD analysis was applied to daily Fall rainfall (1976-2015) at 180 French weather stations, see Figure 1. As often done for such large sets of weather stations, a site dependent threshold is set and, in this introductory and motivating example, it corresponds to the 95% quantile at each location. For such a threshold choice strategy, the left panel provides the expected scale parameter spatial structure for extreme French precipitation, with a large variability over the Mediterranean coast (this is mainly due to the orography and weather patterns produced by the Rhone valley and the influence of the Mediterranean Sea). The shape parameter displayed in the right panel also follows a typical behavior in the sense that the spatial variability is strong with some climatological incoherent features. For example, the weather station of Chartes (coordinates: 48.46 Lat, 1.5 Lon) located South-West of Paris has a shape parameter of 0.47. This is extremely high with respect to its neighbouring stations (which have an estimated ξ ranging from 0 to 0.2). This is also climatological inconsistent in this Parisian region where weather and climatological features for extreme rainfall should be spatially coherent due to the lack of orography and of specific local storms/wind patterns.

To explore this issue in more details, the QQ-plot in the upper-right panel of Figure 2 shows that, despite this high value of ξ around 0.47, the departure from the red diagonal is not pronounced, and this indicates a rather good fit for extreme rainfall at this station. The histogram in the lower-left panel also appears to be in compliance with the estimated GP density. The red flag pointing towards an overestimation of ξ could be the scatterplot between the estimates of σ and ξ obtained by bootstrapping the original rainfall data (before thresholding). This last plot indicates that the variability in the estimation of ξ can be large, ranging from zero to 0.8, and the clear negative correlation between the estimators of σ and ξ is a well known feature of the GP parameter inference, i.e. an overestimation of ξ is coupled by an underestimation of σ and *vice-versa* (see, e.g. Ribereau *et al.*, 2011). As it is operationally impossible to repeat this visual inspection for our 180 weather stations and for different threshold values, this leaves us to wonder if the large estimate of ξ for the Chartres station is just due to the inherent variability of GP fit for small samples (here 40 exceedances above the 95% threshold). Later on in Section 5, we will compare this GP analysis of this rainfall dataset of 180 weather stations with our proposed approach. In particular, the example of the Chartres station will be revisited in detail.

Besides extremal behaviors, practitioners can be interested in summarizing the full rainfall range, *e.g.*, to determine in a climate change Detection & Attri-

Figure 1: Scale (left panel) and shape (right panel) parameters from a classical GPD analysis applied to daily Fall French rainfall (1976-2015). The threshold is chosen to be equal to the 95% quantile at each of 180 weather stations.



bution study if rainfall (extremes included) have changed over time (see, *e.g.*, Hegerl and Zwiers, 2011). In recent years, a few attempts have been made to bypass the threshold choice and to characterize the full precipitation spectrum. Mixture and hybrid models have been proposed, such as the dynamic weighted model used by Vrac and Naveau (2007), or the hybrid model based on a gamma and GPD distributions of Furrer and Katz (2008). In practice, these models have a large number of parameters and the inference remains a challenge (see also Carreau and Bengio, 2008; MacDonald *et al.*, 2011, for mixture approaches).

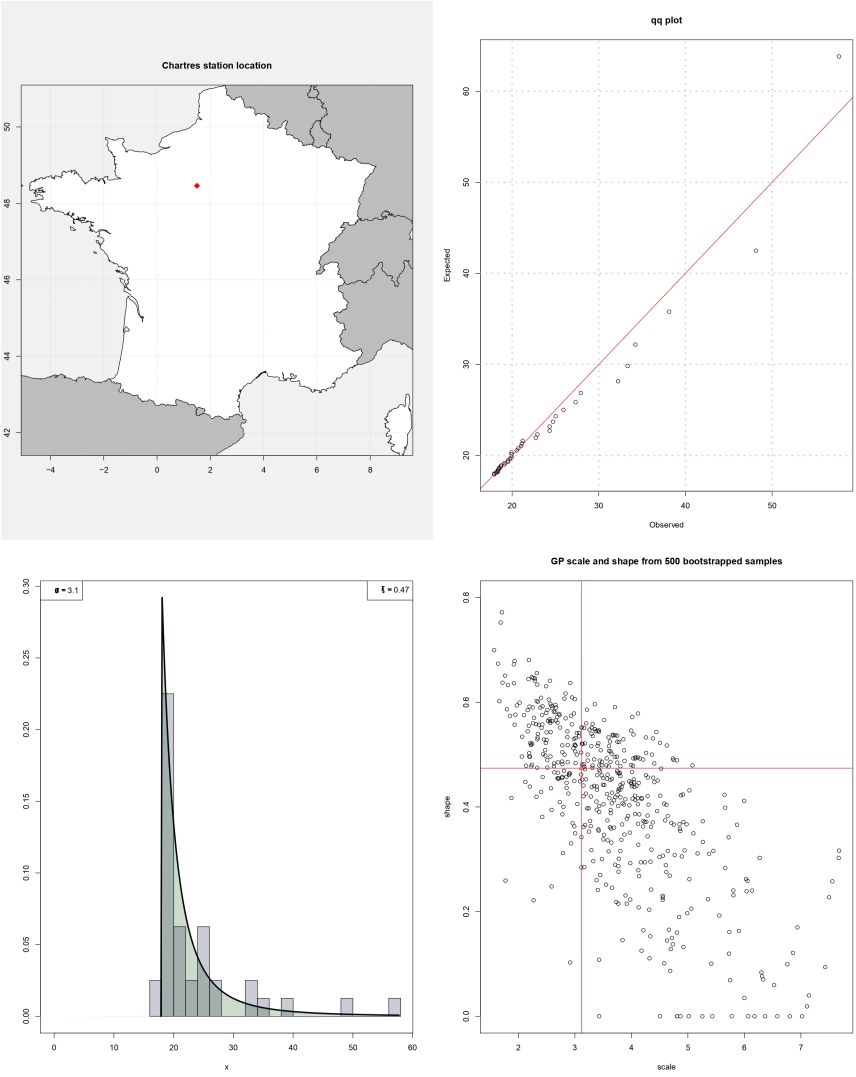
Moving away from the idea of a mixture, Naveau *et al.* (2016) recently proposed a construction that allows a smooth transition between GPD type tails and the middle part (bulk) of the distribution. Here, this class of models is referred as the Extended Generalized Pareto Distribution (EGPD) family. It bypasses the thresholds selection step and it is in compliance with EVT, not only for heavy rainfall, but also for low precipitation amounts. In particular, low precipitation amounts are classically modeled by a gamma distribution.

Mathematically, a member of the EGPD model has to be expressed as

$$F(x) = G\left\{H_\xi(x/\sigma)\right\}, \text{ for all } x > 0, \tag{1}$$

or, in terms of densities, as $f(x) = \frac{1}{\sigma} g\left\{H_\xi(x/\sigma)\right\} \cdot h_\xi(x/\sigma)$, where h_ξ and H_ξ represent the pdf and the cdf of the GPD, while g (*resp.* G) denotes a continuous pdf (*resp.* cdf) on the unit interval. To insure that the upper tail behavior of F is driven by a GPD with parameter ξ, the survival function $\bar{G} = 1-G$ has to satisfy that the limit $a := \lim_{u \downarrow 0} \frac{\bar{G}(1-u)}{u}$ with $u = \overline{H}_\xi(\frac{x}{\sigma})$ is finite and positive as x tends to

Figure 2: GPD analysis of the Chartres station (coordinates: 48.46 Lat, 1.5 Lon)



infinity (u tends to 0). In this case, the upper tail behavior of $\overline{F}(x)$ is equivalent to the original GPD tail used to build $F(x)$, *i.e.*, $\frac{\overline{F}(x)}{\overline{H}_\xi(\frac{x}{\sigma})} = \frac{\bar{G}(1-u)}{u} \sim a > 0$ as x tends to $+\infty$.

To force low rainfall (modeled as $-X$) to follow a GPD for small values near zero, we need that the limit $c := \lim_{u \downarrow 0} \frac{G(u)}{u^s}$ is positive and finite for some positive real s. In this case, $F(x) \sim c \left\{H_\xi(\frac{x}{\sigma})\right\}^s \sim \frac{c}{\sigma^s} x^s$ as x tends to zero.

In Naveau *et al.* (2016), four parametric models for the G function satisfying the required constraints were proposed and compared. But, besides inferential convenience, there is not a theoretical reason to choose a particular parametric G. In this context, the main goal of the present work is to determine if a nonparametric family for G in Equation (1) can be proposed, and more importantly, if this model can be quickly and efficiently inferred. This will bring flexibility to this family and it will be a versatile tool for hydrologists. To reach this target, we take advantage of Bernstein polynomials approximation by relying on the link between Bernstein polynomials and the beta distributions.

The paper is organized as follows. In Section 2, a short background on Bernstein polynomials and its relationship with density estimation is provided. Our proposed semiparametric EGPD model is also described in Section 2 . Section 3 discusses in depth the estimation procedure. Sections 4 and 5 are dedicated to case studies on simulated and rainfall datasets, respectively. Conclusions and perspectives are presented in Section 6.

2 Semiparametric EGPD model class

Naveau *et al.* (2016) explained that the key component of the EGPD class is the continuous function G on $[0, 1]$ (called transition function). The inverse of G can be viewed as a type of anamorphosis on $[0; 1]$. It is applied to define X as $X = \sigma H^{-1}(G^{-1}(U))$ where U follows an uniform distribution. This building block G links the bulk of the distribution with both the upper and lower tails. As our goal is to propose a flexible form for G, a first idea could be to work with Gaussian mixtures. But, as the support of G is the unit interval, this option will lead to overly complex truncations and boundaries effects. Another alternative could be mixtures of beta densities with a few components, *e.g.*, Ji *et al.* (2005). However, parametric mixture models become problematic when the number of components increases, because too many parameters have to be estimated. Another issue is that observed rainfall is measured in millimeters, but G describes data on $[0, 1]$. This implies that the estimation of G is based on pseudo-observations $H_\xi(X_i/\sigma)$, thus requiring as a first step the estimation of $H_\xi(x/\sigma)$.

This implies that any nonparametric estimation of G has to be simple in order to keep at bay computational issues. In this context, kernel density estimators, polynomials approximation, or projections techniques, rather than mixture models, should be favored. A natural way to approximate functions on the unit interval is to use Bernstein polynomials.

2.1 Bernstein polynomials and density estimation

In 1912, Bernstein introduced the polynomials named after him in a proof of the Weierstrass Approximation Theorem. He showed that any continuous function G on the interval $[0,1]$ can be approximated up to some degree of accuracy by Bernstein polynomials. Hence, if G denotes a continuous cdf on $[0,1]$, it can be approximated by the Bernstein estimator of degree $m > 0$ defined by

$$P_m(t, G) = \sum_{k=0}^{m} G\left(\frac{k}{m}\right) b_{k,m}(t),$$

where $b_{k,m}(t) = \binom{m}{k} t^k (1-t)^{m-k}$ for $t \in [0,1]$. These so-called Bernstein bases have attractive properties, *e.g.*, non-negativity, partition of unity and symmetry (see Farouki, 2012, for details). To use the approximation $P_m(t, G)$ in a statistical context, one needs to compute $G\left(\frac{k}{m}\right)$ from a sample drawn from G. The idea of Vitale (1975) was to estimate these coefficients by replacing the unknown G by its empirical cumulative distribution function (ecdf), say $G_n(t)$. This strategy led Vitale to propose the following estimator of the pdf $g(t)$

$$\hat{g}_{m,n}(t) = m \sum_{k=0}^{m-1} \left\{ G_n\left(\frac{k+1}{m}\right) - G_n\left(\frac{k}{m}\right) \right\} b_{k,m-1}(t) \tag{2}$$

that is a valid density (positive and $\int_0^1 \hat{g}_{m,n}(t)dt = 1$) because of $G_n(0) = 0$ and $G_n(1) = 1$. Babu *et al.* (2002) showed that the degree m should be chosen such that $m \in \{2, \ldots, [n/\log(n)]\}$ for large samples of size n. More recently, Leblanc (2010, 2012a,b) studied the boundary properties of both density and distribution estimators, (see also Petrone, 1999; Ghosal, 2001; Kakizawa, 2004; Bouezmarni and Rolin, 2007, for extensions in a Bayesian and/or multivariate context).

As mentioned by Vitale (1975), all approximations based on $b_{k,m}(t)$ can be rewritten in terms of linear combinations of beta densities. In particular, we rewrite $\hat{g}_{m,n}(t)$ as a sum of beta densities with the following notation

$$\hat{g}_{m,n}(t) = \sum_{k=1}^{m} \omega_{k,m} \beta_{k,m-k+1}(t), \tag{3}$$

where $t \in [0,1]$, $\omega_{k,m} = G_n\left(\frac{k}{m}\right) - G_n\left(\frac{k-1}{m}\right)$, and $\beta_{a,b}(t) = t^{a-1}(1-t)^{b-1}/B(a,b)$ corresponds to the classical beta pdf with parameters a and b, respectively, and $B(a,b)$ denotes the beta function. Here, we stress that, given m, the approximation $\hat{g}_{m,n}(t)$ is not a classical mixture of densities with unknown parameters. The beta coefficients a and b are known ($a = k$ and $b = m - k + 1$) and the weights are straightforward to compute if G_n is given. In other words, although $\hat{g}_{m,n}(t)$ is a mixture of beta pdfs, it has to be interpreted as an expansion on the beta "bases". This implies that, given G_n, the only unknown is m. It can be large and interpreted as a type of bandwidth. Finding m corresponds to resolving a bias-variance tradeoff (see *e.g.*, Vitale, 1975; Leblanc, 2012b).

2.2 EGPD model class based on Bernstein-beta density

Now, the EGPD family defined by Equation (1) and the Bernstein approximation captured by Equation (3) can be combined via

$$F_{m,n,\boldsymbol{\theta}}(x) = \hat{G}_{m,n}\left\{H_\xi(x/\sigma)\right\}, \tag{4}$$

where $\hat{G}_{m,n}(t)$ represents the cdf associated with $\hat{g}_{m,n}(t)$ – see Equation (3) – and $\boldsymbol{\theta} = (\sigma, \xi)^t$ corresponds to the GPD parameters. At this stage, we need to determine the constraints on $\hat{G}_{m,n}$ in order that F belongs to the EGDP class.

Lemma 1 *If among all the coefficients $\omega_{k,m} = G_n\left(\frac{k}{m}\right) - G_n\left(\frac{k-1}{m}\right)$ with $k = 1, \ldots, m$, the last one is positive, i.e., $\omega_{m,m} > 0$, then we have*

1. $\lim_{x\to 0} \frac{F_{m,n,\boldsymbol{\theta}}(x)}{x^s} = \frac{m}{s}\sigma^{-s}\binom{m-1}{s-1}\omega_{s,m} > 0$, *where s denotes the position of the first non-null weight in $\boldsymbol{\omega} = (\omega_{1,m}, \ldots, \omega_{m,m})^t$;*

2. $\lim_{x\to\infty} \frac{\overline{F}_{m,n,\boldsymbol{\theta}}(x)}{\overline{H}_\xi(x/\sigma)} = m\omega_{m,m} > 0$.

In addition, let Y be any non-negative continuous random variable such that the conditional limit $\mathbb{P}(Y > x + u | Y > u)$ goes, as u gets large, to $(1 + \tilde{\xi}\frac{x}{\tilde{\sigma}})^{-1/\tilde{\xi}}$ for some parameters $\tilde{\sigma}$ and $\tilde{\xi} > 0$.

3. *If Y can be rewritten as $Y = \sigma H_\xi^{-1}\left\{G^{-1}(U)\right\}$ with the survival of the cdf G satisfying $\lim_{w\downarrow 0}\frac{\overline{G}(1-w)}{w} \in (0, \infty)$, then $\tilde{\xi} = \xi$.*

[Proof of Lemma 1] The proof is postponed to the appendix.

Remark 2 *From Item 1. of Lemma 1 above, we conclude that the lower tail behavior of the model described by Equation* (4) *is controlled by the rank of the first non-null weight in $\boldsymbol{\omega}$. From Item 2. of Lemma 1, we see that the assumption $\omega_{m,m} > 0$ is required to prove that the upper tail behavior in our model is described by a GPD. Item 3. tells us that imposing $\lim_{w\downarrow 0}\frac{\overline{G}(1-w)}{w} \in (0, \infty)$ ensures that our* EGPD *tail behavior driven by ξ is equivalent to the one obtained using a genuine* EVT *argument. This explains why our Bernstein approximation of G has to satisfy this condition too. As*

$$\lim_{w\to 0}\frac{\overline{\hat{G}}_{m,n}(1-w)}{w} = m\omega_{m,m},$$

combining Item 2 and Item 3 implies that the constraint $\omega_{m,m} > 0$ is sufficient to make sure that our Bernstein approximation, $\hat{G}_{m,n}$, does not impact the expected upper GPD tail behavior (under the condition that the observations can be rewritten as $\sigma H_\xi^{-1}\left\{G^{-1}(U)\right\}$).

In practice, we have to deal with a few issues. First, we do not observe random draws from G, but only rainfall measurements from F. Hence, we do not have a direct way to compute the ecdf G_n, and consequently the weights $\omega_{k,m} = G_n\left(\frac{k}{m}\right) - G_n\left(\frac{k-1}{m}\right)$. In addition, we have the constraint $\omega_{m,m} > 0$ that may not be always satisfied. Last but not least, the number of components m in $\hat{g}_{m,n}(t)$ has to be chosen and the parameters $\boldsymbol{\theta} = (\sigma, \xi)^t$ have to be inferred. On the positive side, we can notice that, if random draws from G were available and m given, then the weights $\omega_{k,m}$ can be instantaneously computed, *i.e.*, there is no need of running a time consuming optimization scheme in such a case. In this context, our inferential strategy is to use a recursive argument. Basically, we infer the hidden values drawn from G, compute and adjust the weights, and then estimate $\boldsymbol{\theta} = (\sigma, \xi)^t$. This algorithm can be repeated until the values of some criterion are stable. The next section will detail our strategy.

3 Methodology for fitting a semiparametric EGPD model

3.1 Initial values for $\boldsymbol{\theta} = (\sigma, \xi)^t$

Suppose that $(X_1, \ldots, X_n)$ represents our observed rainfall sample with cdf F. For precipitation data, Naveau *et al.* (2016) noticed that the special case of $G(u) = u^\kappa$ with $\kappa > 0$ provided a decent fit for hourly and daily precipitation in France. Evin *et al.* (2018) used this same G to model precipitation in Switzerland. Consequently, this parametric model for G appears to be a good starting point to give initial estimates for $\boldsymbol{\theta} = (\sigma, \xi)^t$. Let us call them $\hat{\boldsymbol{\theta}}_0 = (\hat{\sigma}_0, \hat{\xi}_0)^t$.

3.2 Bernstein weights approximation given $\mathbf{x} = (x_1, \ldots, x_n)^t, \sigma, \xi, m$

From Equation (1), it is possible to show that the random variable $Z = H_\xi\left(X/\sigma\right)$ follows the cdf G and that the random variable $V = \sigma H_\xi\left(G(Z)\right)$ is simply the Generalized Pareto $H_\xi(./\sigma)$. In this context, it makes sense to introduce the random sample

$$\hat{Z}_i = H_{\hat{\xi}_0}\left(\frac{X_i}{\hat{\sigma}_0}\right), \text{ for } i = 1, \ldots, n. \tag{5}$$

These reconstructed random variables should mimic the hidden sample driven by G. Consequently, the weights $\omega_{k,m}$ at this stage could be estimated by

$$\hat{\omega}_{k,m} = \hat{G}_n\left(\frac{k}{m}\right) - \hat{G}_n\left(\frac{k-1}{m}\right)$$

where $\hat{G}_n$ represents the ecdf obtained from $(\hat{Z}_1, \ldots, \hat{Z}_n)$. To complete our weights inference, we need to force the last weight $\hat{\omega}_{m,m}$ to be non null. If $\hat{\omega}_{m,m}$ is non null, then we are done. Otherwise, we recall that the cdf $\hat{G}_{m,n}(t)$ associated with $\hat{g}_{m,n}(t)$ – see Equation (3) – can be viewed as a mixture of Beta cdfs, and consequently, given the estimated weights $(\hat{\omega}_{1,m} \ldots, \hat{\omega}_{m-1,m}, 0)$, the value

$\hat{G}_{m,n}(t)$ can be computed for any $t \in [0,1]$, and in particular for $t = 1 - 1/m$. The last weight $\omega_{m,m} = G_{m,n}(1) - G_{m,n}(1-1/m) = 1 - G_{m,n}(1-1/m)$ can be estimated by $1 - \hat{G}_{m,n}(1-1/m)$. This quantity is positive because, at least, one of the weights in front of the Beta cdfs in the mixture is non null for $t = 1-1/m$. With this new non-null estimate of $\omega_{m,m}$, we need to renormalize all weights to insure that their sum is equal to one, see Algorithm 1 below.

Algorithm 1 Weights approximation given $\mathbf{x} = (x_1, \ldots, x_n)^t, \sigma, \xi, m$

procedure INPUTS:$(\mathbf{x}, \sigma, \xi, m)$
  $n = \text{length}(\mathbf{x})$
  $z_i = H_\xi(x_i/\sigma)$
  for each integer k in $1 : m$ **do**
    $\omega_{k,m} = G_n(\frac{k}{m}) - G_n(\frac{k-1}{m})$ with G_n ecdf obtained from $(z_1, \ldots, z_n)^t$
  end for
  if $(\omega_{m,m} = 0)$ **then**
    $\omega_{m,m} = 1 - \hat{G}_{m,n}(\frac{m-1}{m})$ with $\hat{G}_{m,n}(t)$ derived from (3)
  end if
  $\boldsymbol{\omega} = \boldsymbol{\omega} / \sum \boldsymbol{\omega}$ ▷ normalization to add to 1
  Return $\boldsymbol{\omega}$
end procedure

3.3 Sequential estimation of $\boldsymbol{\theta} = (\sigma, \xi)^t$ via Probability Weighted Moments

For this step, we assume that m and initial parameters of the GP part are given, say $\boldsymbol{\theta_0} = (\sigma_0, \xi_0)^t$. From Algorithm 1, the weights describing G can be estimated. So, the next step is to update the two parameters $\boldsymbol{\theta} = (\sigma, \xi)^t$. In Section 3.1, we notice that the random variable $V = \sigma H_\xi\left(G(Z)\right)$ always follows the Generalized Pareto $H_\xi(./\sigma)$. In hydrology, there is a long history of using the so-called Probability Weighted Moments (PWM) to infer the parameters of a GPD (see, e.g., Hosking and Wallis, 1987; Ribereau *et al.*, 2011; Naveau *et al.*, 2016; Carreau *et al.*, 2017). The PWM estimation method is easy to understand, fast, robust and efficient (if $\xi < 0.5$, a reasonable assumption for our rainfall data). To emphasise the speed of this inference method, we recall that the estimates of σ and ξ are *explicit* in terms of the sampled PWMs, m_0 and m_1 defined as follows,

$$\hat{\xi} = \frac{m_0 - 4m_1}{m_0 - 2m_1} \text{ and } \hat{\sigma} = m_0(1-\xi), \text{ where } m_0 = \frac{1}{n}\sum_{i=1}^{n} V_i \text{ and } m_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{n-i}{n-1} V_i, \tag{6}$$

whenever $(V_1, \ldots, V_n)^T$ represents a GPD distributed random sample with parameters σ and ξ.

For this reason, we define from (5) the following random variables

$$\hat{V}_i = \hat{\sigma}_0 H_{\hat{\xi}_0}\left(\hat{G}_{m,n}(\hat{Z}_i)\right), \text{ for } i = 1, \ldots, n. \tag{7}$$

and we can apply (6) to the sample $(\hat{V}_1, \ldots, \hat{V}_n)^T$ and get new estimates of σ and ξ. If these estimates are close to $\boldsymbol{\theta_0} = (\sigma_0, \xi_0)^t$, then the algorithm converges towards stable values. If it is not the case, $\boldsymbol{\theta_0} = (\sigma_0, \xi_0)^t$ have to be replaced by the new ones and the procedure starts over. Algorithm 2 in the next section summarizes the details of such a loop.

3.4 Main algorithm

Algorithm 2 Estimation of $(\omega_{1,m}, \ldots, \omega_{m,m})^t$, σ and ξ in Equation (4) for a given m

procedure INPUTS($\mathbf{x}$ and m)
  $\boldsymbol{\theta}_0 = (\sigma_0, \xi_0)^t$ ▷ Give initial values, e.g., by fitting a EGDP to $\mathbf{x}$ with $G(u) = u^\kappa$;
  $cond = true, eps = 0.001$ and $\boldsymbol{\theta}_{init} = \boldsymbol{\theta}_0$
  while $cond = true$ **do**
    $\boldsymbol{\omega}_{new} =$ ALGORITHM 1$(\mathbf{x}, \sigma_{init}, \xi_{init}, m)$ ▷ getting weights
    Compute $\hat{Z}_i = H_{\hat{\xi}_{init}}\left(\frac{X_i}{\hat{\sigma}_{init}}\right)$
    Compute $\hat{G}_{m,n}(\hat{Z}_i)$ from (3) with the weights $\boldsymbol{\omega}_{new}$ ▷ getting Bernstein values
    Compute $\hat{V}_i = \hat{\sigma}_{init} H_{\hat{\xi}_{init}}\left(\hat{G}_{m,n}(\hat{Z}_i)\right)$
    and $m_0 = \frac{1}{n}\sum_{i=1}^n \hat{V}_i$ and $m_1 = \frac{1}{n}\sum_{i=1}^n \frac{n-i}{n-1}\hat{V}_i$, ▷ PWMs estimates
    Estimate $\hat{\xi}_{new} = \frac{m_0 - 4m_1}{m_0 - 2m_1}$ and $\hat{\sigma}_{new} = m_0(1 - \hat{\xi}_{new})$ ▷ estimate $\boldsymbol{\theta}$ from PWMs
    if $abs(\hat{\xi}_{new} - \hat{\xi}_{init}) < eps$ **then**
      $cond = false$
    end if
    $\boldsymbol{\theta}_{init} = \boldsymbol{\theta}_{new}$
  end while
  Return $\boldsymbol{\theta}_{new}, \boldsymbol{\omega}_{new}$
end procedure

Remark 3 *All numerical computation within each "***While** *$cond = true$" step are explicit and consequently extremely fast. This contrasts with a likelihood approach for which a maximisation would have been required. Still, at each time step, the likelihood can be easily obtained by plugging $(\boldsymbol{\theta}_{new}, \boldsymbol{\omega}_{new})$ in*

$\hat{f}_{m,n,\boldsymbol{\theta}}(x) = \frac{1}{\sigma}\hat{g}_{m,n}\left\{H_\xi(x/\sigma)\right\} \cdot h_\xi(x/\sigma)$ *with* $\hat{g}_{m,n}(t) = \sum_{k=1}^{m} \omega_{k,m}\beta_{k,m-k+1}(t)$. *In particular, it would be possible to replace our stopping criterion* $abs(\hat{\xi}_{new} - \hat{\xi}_{init}) < eps$ *by the increment between the "init" and "new" log-likelihoods. We prefer monitoring the change in the shape parameter because it is expected to become constant in a EVT setting and a large variability in* ξ *will be a worrisome sign of our algorithm.*

3.5 Selection of the Bernstein polynomial degree m

The number of components in the mixture, or alternatively the degree of the polynomial, is a very important feature of the model as it directly influences the smoothness of the estimator. Babu *et al.* (2002) showed that m should be of order $o\left\{[\frac{n}{\log n}]\right\}$ for consistent convergence results. Also, their numerical study indicated that the setting $m = [\frac{n}{\log n}]$ works well. But, they only covered small sample sizes (up to 125 observations), so when working with larger samples, the degree $m = [\frac{n}{\log n}]$ could be too large.

To avoid this issue, we prefer to rely on a data driven approach and use the Least Square Cross Validation (LSCV) scheme which is based on the minimization of the Mean Integrated Squared Error (MISE) (see Kakizawa, 2004; Leblanc, 2010). The notation $\hat{f}^{(-i)}_{m,n,\theta}(x)$ and $\hat{g}^{(-i)}_{m,n}(t)$ indicate the same estimators based on all data but X_i, respectively. The optimal polynomial degree m (see Bouezmarni and Rolin, 2007, in the case of Bernstein estimators) is the integer that minimizes

$$\begin{aligned}\text{MISE}(m) =& \mathbb{E}\left[\int_0^\infty \left\{\hat{f}_{m,n,\theta}(x) - f(x)\right\}^2 \mathrm{d}x\right] \\ =& \mathbb{E}\left\{\int_0^\infty \hat{f}^2_{m,n,\theta}(x)\mathrm{d}x\right\} - 2\mathbb{E}\left\{\int_0^\infty \hat{f}_{m,n,\theta}(x)f(x)\mathrm{d}x\right\} + \int_0^\infty f^2(x)\mathrm{d}x.\end{aligned} \tag{8}$$

The last term in Equation (8) does not depend on m, thus it can be dropped to seek for the minimizer of MISE(m). We then search for the degree that minimizes the quantity $\text{MISE}(m) - \int_0^\infty f^2(x)\mathrm{d}x = \mathbb{E}\left\{\int_0^\infty \hat{f}^2_{m,n,\theta}(x)\mathrm{d}x\right\} - 2\mathbb{E}\left\{\int_0^\infty \hat{f}_{m,n,\theta}(x)f(x)\mathrm{d}x\right\}$, which depends on the unknown f. A common practice is to replace this last quantity by an estimator built from a data-driven procedure based on

$$\text{LSCV}(m) = \int_0^\infty \hat{f}^2_{m,n,\theta}(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \hat{f}^{(-i)}_{m,n,\theta}(X_i), \tag{9}$$

and then to infer the optimal positive integer that minimises this criterion, i.e. $m_{\text{LSCV}} = \min_{m\in\mathbb{Z}^+} \text{LSCV}(m)$.

As $\hat{f}_{m,n,\boldsymbol{\theta}}(x) = \frac{1}{\sigma}\hat{g}_{m,n}\left\{H_\xi(x/\sigma)\right\} \cdot h_\xi(x/\sigma)$, we have

$$\hat{f}_{m,n,\boldsymbol{\theta}}(x) = \frac{1}{\sigma}\hat{g}_{m,n}(u) \cdot (1-u)^{1+\xi} \text{ with } u = H(\tfrac{x}{\sigma}) = 1 - \left(1 + \xi\tfrac{x}{\sigma}\right)^{-\frac{1}{\xi}},$$

and the cross validation criterion for our specific GP-Berstein model becomes

$$\begin{aligned}\text{LSCV}(m) &= \frac{1}{\sigma}\left(\int_0^1 \hat{g}_{m,n}^2(u)\cdot(1-u)^{1+\xi}\mathrm{d}u - \frac{2}{n}\sum_{i=1}^{n}\hat{g}_{m,n}^{(-i)}\left\{H_\xi(X_i/\sigma)\right\}\cdot h_\xi(X_i/\sigma)\right),\\ &= \frac{1}{\sigma}\left(\int_0^1 \hat{g}_{m,n}^2(u)\cdot(1-u)^{1+\xi}\mathrm{d}u - \frac{2}{n}\sum_{i=1}^{n}\hat{g}_{m,n}^{(-i)}(Z_i)\cdot(1-Z_i)^{1+\xi}\right),\end{aligned}$$

with $Z = H_\xi\left(X/\sigma\right)$. This last equality indicates that the optimisation does not depend on σ, but could be impacted by ξ. For $\xi = -1$, i.e. when the GPD corresponds to an uniform random variable, LSCV(m) is equal to the expression for the classical Bernstein polynomial approximation

$$\int_0^1 \hat{g}_{m,n}^2(u)\mathrm{d}u - \frac{2}{n}\sum_{i=1}^{n}\hat{g}_{m,n}^{(-i)}(Z_i).$$

For this latter quantity, we recall that Leblanc (2010) on page 469 expressed it as

$$\int_0^1 \hat{g}_{m,n}^2(u)\mathrm{d}u - \frac{2}{n}\sum_{i=1}^{n}\hat{g}_{m,n}^{(-i)}(Z_i) = \frac{m^2}{2m-1}\mathbf{G}_m^T A_{m,m}\mathbf{G}_m - \frac{2}{n-1}\left(\sum_{i=1}^{n}\hat{g}_{m,n}(Z_i) - \sum_{i=1}^{n}\beta_{k_i+1,m-k_i}(Z_i)\right), \tag{10}$$

where the vector $\mathbf{G}_m^T$ equals $(G_n(1/m), G_n(2/m) - G_n(1/m), \ldots, 1 - G_n(1 - 1/m))$, $A_{m,m} = (a_{k,l})_{1\leq k,l\leq m}$, is the matrix with elements

$$a_{k,l} = \frac{\binom{m-1}{k}\binom{m-1}{l}}{\binom{2(m-1)}{k+l}},$$

and $k_i = [mZ_i]$ corresponds to the sequence of integers such that $Z_i \in (k_i/m, (k_i+1)/m]$. Computationally, the advantage of the right-hand side of Equation (10) over its left-hand side is enormous because there is no need to estimate $\hat{g}_{m,n}^{(-i)}$ for each $i =, 1 \ldots, n$. Applying similar mathematical derivations found in Kakizawa (2004) and Leblanc (2010), it is possible to show that $\sigma \times \text{LSCV}_\theta(m)$ is equal to

$$\frac{m^2}{2m+\xi}\mathbf{G}_m^T A_{m,m}\mathbf{G}_m - \frac{2}{n-1}\left(\sum_{i=1}^{n}\hat{g}_{m,n}(Z_i)(1-Z_i)^{1+\xi} - \frac{1}{n}\sum_{i=1}^{n}\beta_{k_i+1,m-k_i}(Z_i)(1-Z_i)^{1+\xi}\right), \tag{11}$$

where $A_{m,m}$ and $\mathbf{G}_m$ are identical to their definitions in (10). This new expression makes the optimization of the criterion computationally feasible.

4 Simulation study

In this section, we test our approach with the three following models. The first one belongs to the class of EGPD model and it is defined by taking $G(u) = u^2$,

$\sigma = 1$ and $\xi = 0.2$ in (1). To move away from the class of EGPD model, we also consider two kinds of gamma-GPD mixtures. The first one is a non contaminated mixture (called mixture in the following), whose tail is thus exactly GPD. It is described as follows

$$\mathbb{P}(X \leq x) = \begin{cases} F_{a,b}(x), \ \forall\, x \leq s, \\ F_{a,b}(u) + (1 - F(u)) H_\xi \left(\dfrac{x-u}{\tilde{\sigma}} \right) \ \forall\, x > s, \end{cases} \tag{12}$$

with $s > 0$ a given threshold and $F_{a,b}$ the cdf of a gamma distribution of shape parameter $a > 0$ and scale parameter $b > 0$. The second contaminated mixture model, for which we expect a real improvement by using our semiparametric estimation procedure rather than traditional EVT modeling, is described by

$$\mathbb{P}(X \leq x) = p\, F_{a,b}(x) + (1-p)\, H_\xi \left(\frac{x}{\tilde{\sigma}} \right), \ \forall\, x > 0, \tag{13}$$

with $p \in [0,1]$ a given mixture parameter. From EVT, we know that for any large threshold u and any $x \geq u$

$$\begin{aligned} \overline{F}(x) &= \mathbb{P}(X > u)\mathbb{P}(X > x | X > u), \\ &\approx (1 - q_u) \left(1 + \tilde{\xi} \tfrac{x-u}{\sigma_u} \right)^{-1/\xi}, \end{aligned}$$

with q_u the probability that the threshold u is not exceeded. As the upper tail of this mixture model is driven by the GPD component, we thus get, by identification, $\tilde{\xi} = \xi$.

Concerning the sample size and the number of replica, we investigate two sample sizes with $n = 700$, the order of magnitude of the number of observations in our rainfall application in Section 5, and $n = 1500$ for exploring large sample behaviors. The number of replica is set to 500 for all simulations. The range of the polynomial degree m is always equal to the set $\{1, 2, \ldots, 100\}$.

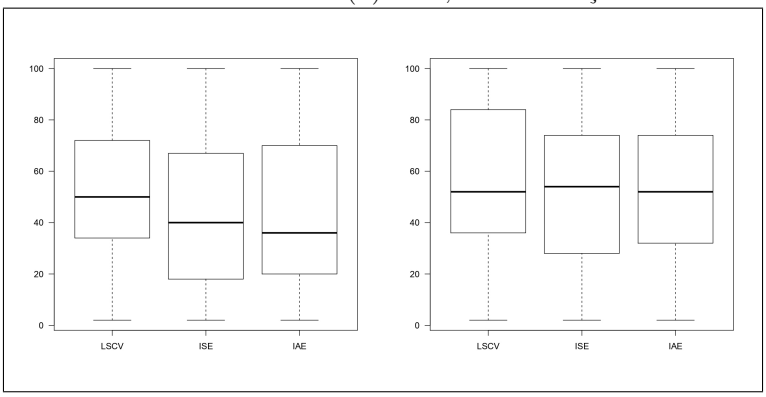
Concerning the model parameters, we set $\tilde{\sigma} = 1$, $\xi = 0.2$ and $G(u) = u^2$ for all models, and fix $s \in \{q_{0.7}, q_{0.9}\}$, $a = 2$, $b = 3$ for the mixture defined by (12) where $q_{0.7}$ and $q_{0.9}$ denote respectively the 70% and the 90% quantiles of the gamma distribution, and fix $p \in \{0.7, 0.9\}$, $a = 2$ and $b = 3$ for the contaminated mixture defined by (13). Note that $\tilde{\sigma} = 1$ in models (12) and (13) does not imply that $\sigma = 1$ in our EGPD model (1).

4.1 Finding the optimal polynomial degree m

In Figure 3, our LSCV criterion defined by (11) is computed for the chosen EGPD model. The left and right panels corresponds to $n = 700$ and $n = 1500$, respectively. The y-axis represents the different values of $m \in \{2, 4, ..., 100\}$, and the x-axis shows the boxplot of 500 optimal values of LSCV. Note that here and through the whole paper, the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range. It is compared to the Integrated Absolute Error (IAE, $\int_0^\infty |\hat{f}(x) - f(x)| \mathrm{d}x$) and the Integrated

Figure 3: Boxplots of the optimal Bernstein polynomial degree m of the 500 samples of size $n = 700$ (left panel) and size $n = 1500$ (right panel) described in Section 4.1 obtained with the LSCV, ISE, IAE approaches

Model EGPD with $G(u) = u^2$, $\sigma = 1$ and $\xi = 0.2$



Squared Error (ISE, $\int_0^\infty \left\{\hat{f}(x) - f(x)\right\}^2 \mathrm{d}x$) approaches, both can be computed if the true $f(.)$ is known. Overall, the three boxplots are similar for the three measures, independently of the sample size. This shows the robustness of the method for choosing m. Table 1 shows the median value of m obtained using the LSCV for the three models under study and for the two sample sizes. In the following, for each model, m is fixed to the value stated in Table 1. Note that the choice of m suggested in Babu *et al.* (2002), *i.e.*, $\bar{m}(n) = [\frac{n}{\log n}]$, yields $\bar{m}(700) = 106$ and $\bar{m}(1500) = 205$, which clearly overfits in our framework.

Table 1: Median value of m obtained using the LSCV for the three models under study.

Model	$n = 700$	$n = 1500$
EGPD	50	52
Mixture $s = q_{0.7}$	43	60
Mixture $s = q_{0.9}$	50	58
Contaminated mixture $p = 0.7$	60	66
Contaminated mixture $p = 0.9$	64	46

Now, we want to compare the fit under two setups. In the first one, we make inference with the knowledge that the true model is based on G of the form $G(u) = G_\kappa(u) = u^\kappa$. In that setup, we denote the true model by f_κ. In the second setup, we approximate the nonparametric function G with Bernstein polynomials, see equations (3) and (4). It leads to a semiparametric estimation. Intuitively, the inference should be better in the first setup. Still, the upper panels of Figure 4 indicate that our semiparametric EGPD fit (EGPD$_{m,n}$) based on Bernstein polynomials (dotted boxplots) performs well for the inference of

scale and shape parameters (left) and even better for the inference of quantiles (right). In the lower panels of Figure 4, we represent the functional boxplots obtained from the 500 estimates of the densities f (on the left) and g (on the right). Functional boxplots were introduced in Sun and Genton (2011). These type of boxplots use the band depth introduced by López-Pintado and Romo (2009) to order the functions. The band depth measures the centrality of the curves: the greater the band depth, the more the curve is central. The median curve is represented with a black solid line. Half of the curves (the more central ones) are contained in the bag colored in pink. The fence, which separates the outlier curves from the other ones, is delimited by blue lines. Outlier curves are represented with red dashed lines. The loop is defined as the curves contained in the fence, but outside the bag. We refer to Sun and Genton (2011) for more details on functional boxplots. Concerning the shape of the densities f, our Bernstein polynomials approach provides comparable estimates with the one obtained under the true model (left lower panel of Figure 4). The shape of the densities g are rather different under both different models. We also remark that there are many outlier curves (in red) for the estimates of g with the true model, which probably explains why the quantiles are better estimated under the semiparametric model.

4.2 Sensitivity studies under gamma-GPD mixtures

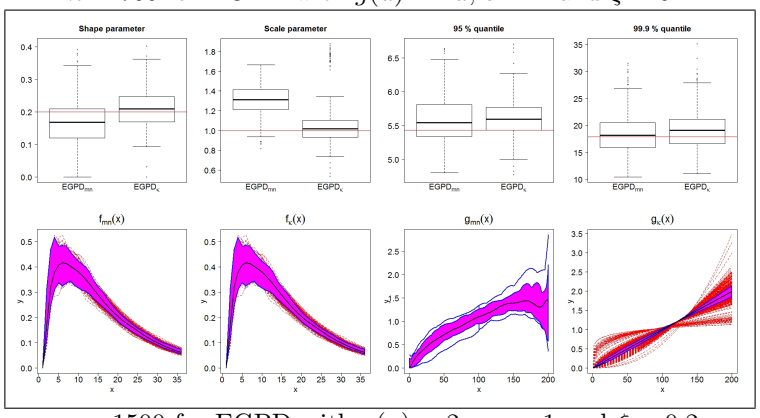
To assess the quality of our approach when the data are drawn from a model outside of the EGPD class, we now focus on the two models defined by (12) and (13).

We compare our $\text{EGPD}_{m,n}$ model fit with the ones obtained from the simplest parametric EGPD with $G(u) = u^{\kappa}$ (denoted EGPD_{κ}) and also from a classical GPD model. For the latter, excesses are defined with respect to two classical threshold values: (a) the $95\% -$ empirical quantile ($\text{GPD}_{q_{95}}$) and (b) the $98\% -$ empirical quantile ($\text{GPD}_{q_{98}}$).

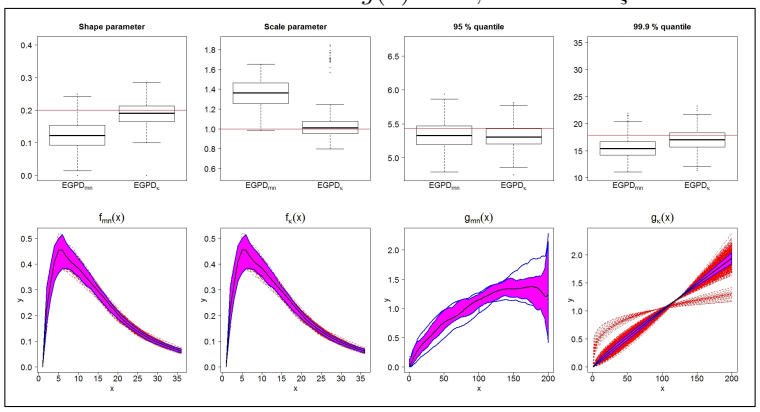
In this comparison exercise, our main goal is to assess the performance of each approach to infer large quantiles and also very large quantiles to assess their capacity to extrapolate beyond the largest observation. Table 2 summarizes the RMSE obtained for the 80%, 90%, 95%, 97.5% and 99.5% quantiles. The $\text{EGPD}_{m,n}$ and EGPD_{κ} models clearly outperform the classical GPD analysis, either based on excesses above the 95% quantile or the 98% quantile, which could seem particularly surprising for the estimation of large quantiles. This can be explained by two main reasons. First, the threshold choice (here the 95% and 98% quantiles) is a difficult task. Then, thresholding the data leads to a drastic information reduction. Indeed, for $n = 1500$, we only keep 75 or 30 data according to the chosen threshold, while for $n = 700$, we keep 35 or 14 data. Concerning the estimations based on the EGPD_{κ} model or on the $\text{EGPD}_{m,n}$ model, we remark that increasing the sample size n does not always lead to a decrease of the RMSE values. Note that it is probably due to a residual bias in both models induced by the modeling of the bulk either by a power function or by truncated Bernstein series.

Figure 4: Comparing our Bernstein EGPD fit ($\text{EGPD}_{m,n}$) with a parametric EGPD fit when the true model is $g(u) = 2u$ (EGPD_{κ}). The boxplots for the shape and scale parameters (left upper panel) and for the 95% and 99.9% quantiles (right upper panel) are obtained from 500 samples. The red horizontal lines in the upper panels indicate the true parameter and quantile values. Functional boxplots are obtained from the densities f (left lower panel) and g (right lower panel) estimated from the 500 samples. The semiparametric model is estimated with $m = \bar{m}_{\text{LSCV}} = 40$

$n = 700$ for EGPD with $g(u) = 2u$, $\sigma = 1$ and $\xi = 0.2$



$n = 1500$ for EGPD with $g(u) = 2u$, $\sigma = 1$ and $\xi = 0.2$



RMSE values increase with the order of the quantiles which is not surprising as larger quantiles are more difficult to estimate. We can also notice that our semiparametric model $\text{EGPD}_{m,n}$ is superior to the simpler model EGPD_{κ} for each quantile for the mixture model when $s = q_{0.7}$ and for the contaminated mixture when $p = 0.7$ for both sample sizes. Then, choosing $s = q_{0.9}$ for the mixture or $p = 0.9$ for the contaminated mixture, respectively, we decrease the weight of the bulk in the whole distribution. This explains why both semiparametric and parametric approaches behave similarly with respect to the RMSE values in these last cases, while $\text{EGPD}_{m,n}$ is not systematically outperforming EGPD_{κ}.

5 Application study - daily rainfall in France

In this section, we apply our Bernstein GPD model to the daily Fall positive rainfall (i.e., above 2 mm) recorded at 180 French weather stations during the period 1976-2015. This analysis can be compared to the classical GP study presented in the introduction, see Figure 1.

To visualize the impact of the choice of m, the Bernstein polynomial expansion order, the four panels in Figures 5 and 6 display the estimates of ξ and σ for four different values of $m = 5, 15, 30$ and 50. Overall, the choice of m does not appear to change much the inferred values of ξ and σ. This robustness with respect to the choice of m simplifies the comparison with a classical Pareto analysis on excesses above a threshold. Contrasting the right panel of Figure 1 with Figure 5 shows that our EGPD estimates of ξ, regardless of m, have less spatial variability than the classical GPD approach based on a 95% threshold. The spatially coherent structure where the Rhône valley from Lyon to the Mediterranean coast indicates higher values of ξ is climatologically expected for the South of France (e.g., see Carreau *et al.*, 2017, for a EVT analysis of this region). This spatial structure can also be viewed in the estimates of σ, see Figure 6.

Focusing now on the Chartres weather station, we recall, see Section 1, that the shape parameter estimated by using a classical GP analysis with 40 exceedances above the 95% threshold (around 18 mm) was equal to 0.47. This shape estimate was consider very large in comparison to its climatological similar neighboring stations. Applying our Bernstein GP model to the full sample of 786 positive rainfall values provides a more coherent estimate of ξ. For $m = 5, 15, 30$ and 15, the shape parameter is estimated to be equal to $0.16, 0.18, 0.22$ and 0.22, respectively. These inferred $\hat{\xi}$ are in the typical range of values expected around the Paris region. To visually access the quality of the fit, the QQ-plots for $m = 5, 15, 30$ are shown in the left panels of Figure[1] 7. Compared to the one obtained with a classical GP analysis solely based on exceedances, see Figure 2, the fit of the five largest rainfall values appear to be even better. In addition, we have modeled the full spectrum of positive precipitation and avoid

[1]The same type of QQ-plot is obtained for $m = 50$, but we kept only six panels instead of eight to make this figure readable.

Table 2: Root Mean Square Error (RMSE) between true and estimated quantiles. Each RMSE is computed by fitting four different models (columns) to 500 samples of size $n = 700$ (left panel) and size $n = 1500$ (right panel) drawn from the gamma-Fréchet mixture defined by (12) with $s = q_{0.7}$, $a = 1$, $b = 2.5$ and $\alpha = 2$, the gamma-Fréchet mixture with $s = q_{0.9}$, $a = 1$, $b = 2.5$ and $\alpha = 2$, the contaminated model defined by (13) with $p = 0.7$, $a = 2$, $b = 3$, $G(u) = u^2$, $\sigma = 1$ and $\xi = 0.2$, and the contaminated model with $p = 0.9$, $a = 2$, $b = 3$, $G(u) = u^2$, $\sigma = 1$ and $\xi = 0.2$

Mixture model with $s = q_{0.7}$, $a = 2$, $b = 3$, $G(u) = u^2$, $\sigma = 1$ and $\xi = 0.2$

q	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$
80%	0.332	0.353	22.089	26.274	0.438	0.527	16.058	17.598
90%	1.075	2.118	34.846	37.597	1.200	2.521	22.875	23.267
95%	1.860	3.396	52.664	50.481	2.113	4.092	31.220	28.945
97.5%	2.307	3.884	77.728	65.301	2.704	4.871	41.492	34.625
99.5%	9.084	12.347	184.736	110.082	4.618	8.071	75.857	47.816

Mixture model with $s = q_{0.9}$, $a = 2$, $b = 3$, $G(u) = u^2$, $\sigma = 1$ and $\xi = 0.2$

q	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$
80%	0.124	0.334	15.311	21.225	0.106	0.253	12.139	14.923
90%	0.535	1.001	25.936	32.522	0.493	0.826	19.337	22.131
95%	1.730	1.376	36.749	41.945	1.948	1.637	24.559	26.374
97.5%	3.449	3.435	51.951	52.642	4.034	3.929	30.973	30.563
99.5%	5.440	7.127	115.802	84.064	6.162	7.574	52.194	39.947

Contaminated mixture with $p = .7$, $a = 2$, $b = 3$, $G(u) = u^2$, $\sigma = 1$ and $\xi = 0.2$

q	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$
80%	0.369	0.435	12.245	22.909	0.177	0.202	14.397	18.757
90%	0.644	1.352	15.040	34.452	0.866	1.745	18.319	25.862
95%	1.255	2.794	17.371	49.392	1.718	3.427	21.902	33.686
97.5%	1.839	3.556	19.370	69.355	2.175	4.363	25.276	42.617
99.5%	7.997	11.247	22.679	148.663	5.987	8.193	32.108	69.390

Contaminated mixture with $p = .9$, $a = 2$, $b = 3$, $G(u) = u^2$, $\sigma = 1$ and $\xi = 0.2$

q	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$	$\mathbf{EGPD}_{m,n}$	$\mathbf{EGPD}_{\kappa}$	$\mathbf{GP}_{q95}$	$\mathbf{GP}_{q98}$
80%	0.056	0.163	12.976	18.562	0.070	0.131	10.543	17.664
90%	0.226	0.498	20.673	29.323	0.264	0.415	16.687	29.565
95%	0.577	0.467	28.759	41.023	0.500	0.455	22.768	44.051
97.5%	2.650	2.512	36.786	53.369	2.694	2.739	28.186	61.671
99.5%	5.842	6.528	64.685	95.282	5.904	6.951	46.443	133.571

Figure 5: Shape parameters from the Bernstein GPD analysis applied to daily Fall French rainfall (1976-2015). Each panel corresponds to a different Bernstein polynomial order with $m = 5, 15, 30, 50$ for the upper left, upper right, lower left and lower right panels respectively. Overall, the estimation of ξ is spatially robust with respect to the choice of m.

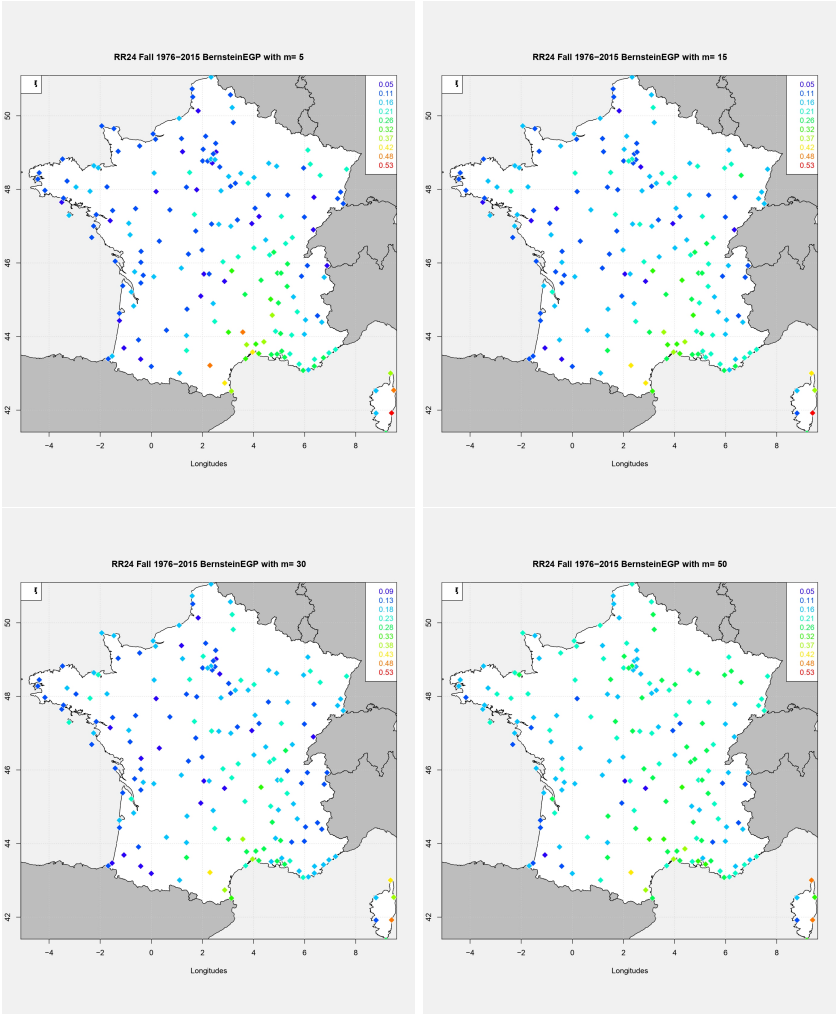
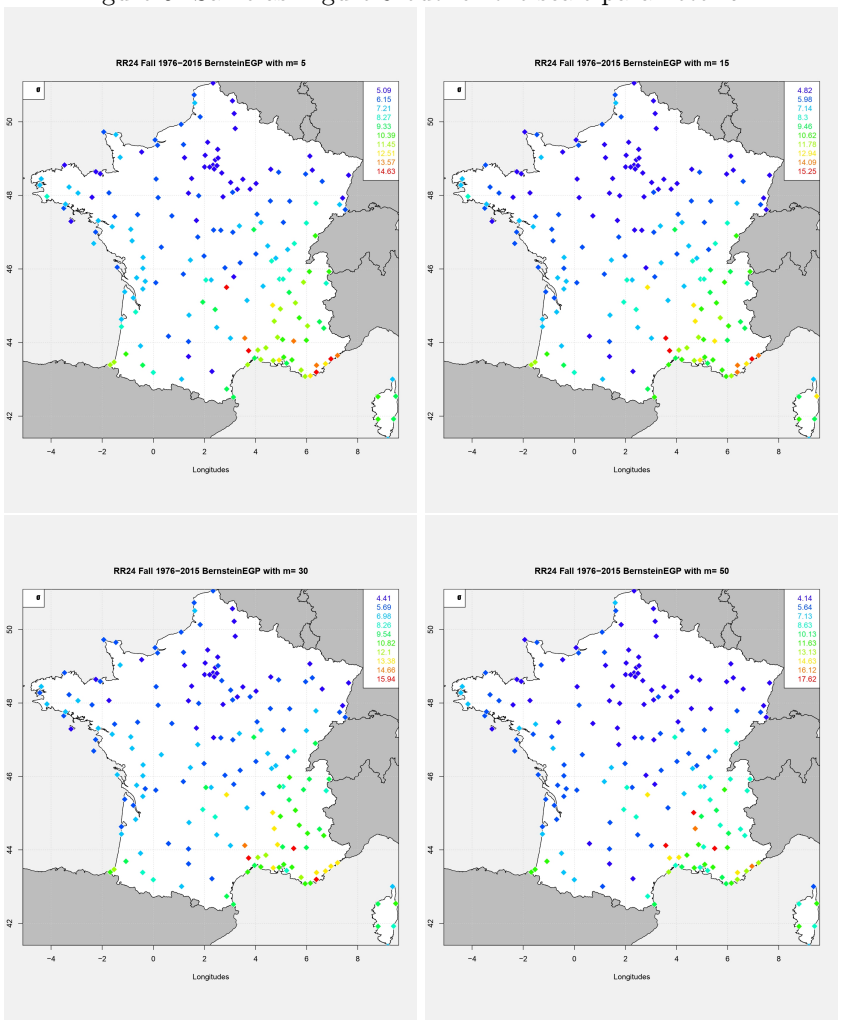


Figure 6: Same as Figure 5 but for the scale parameter σ.



the difficult step of threshold selection. Instead, we have to select the value of m that has the advantage of being robust with respect to the value of ξ and the QQ-plot fit. The left panels of Figure 2 display the estimated Bernstein densities $\hat{g}_{m,n}(u)$ for $n = 786$ and $m = 5, 15, 30$, respectively. Recalling that the identity case $g(u) = 1$ corresponds to a GP density, an increase in the value of m produces an added flexibility to capture variations either in small, moderate or large precipitation. This flexibility is particularly interesting if the geophysical processes driving small rainfall amounts differ from the ones responsible of heavy rainfall.

To conclude this example, one has to keep in mind that this analysis is preliminary. By independently fitting our EGPD model to each station, we did not take into account the spatial structure. It would be of interest to adapt our model to make a regional analysis that will spatially share spatial information, marginally and in a multivariate way (see, e.g. Cooley *et al.*, 2007; Davison *et al.*, 2012; Carreau *et al.*, 2017). In addition, the point estimates in Figures 6 and 5 were obtained by using all positive rainfall in order to work with large samples, e.g. 786 values above 2 mm for the Chartres station. Although daily rainfall become quickly uncorrelated with time, this temporal dependence (see, e.g. Fawcett and Walshaw, 2007) could affect the confidence intervals.

6 Conclusions and perspectives

This work addresses the statistical modeling of the entire range of precipitation amounts. The main benefit of our proposed approach is that low and large rainfall are in compliance with EVT and moderate precipitation is captured by a semiparametric model based on Bernstein polynomials. In other words, we have flexibility when we need flexibility (in the pdf bulk) and constraints when we need them (in the tails).

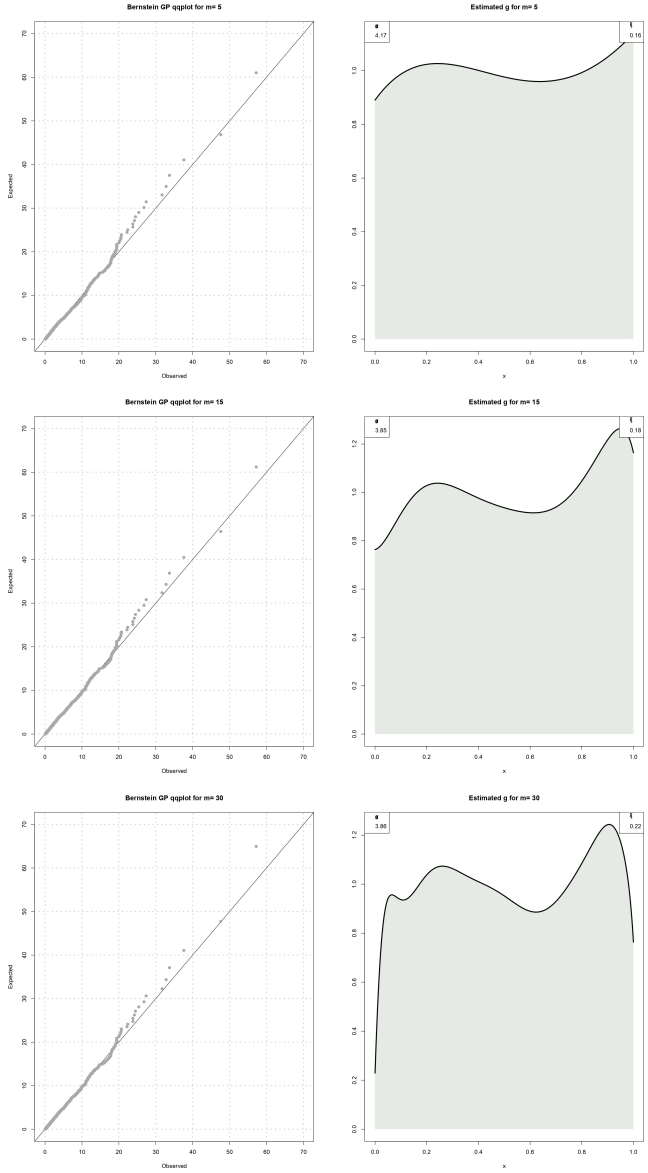
The performance of our semiparametric EGPD model has been evaluated with two simulation studies. For these examples, our inference method seems to accurately estimate moderate to high quantiles, and to outperform the classical GPD approach.

We also consider a real application composed with a large network of 180 precipitation time series over France. This analysis has shown the inherent variability of the shape parameter with the classical GPD approach and has highlighted the flexibility of our semiparametric approach with respect to rainfall data.

With our semiparametric model, an appropriate Bernstein polynomial degree m has to be selected. A criterion based on cross validation, LSCV, is used to this end. The criterion is computed with an algebraic formula, which only requires one fit of the model, thus reducing drastically the computational time.

On the real application, we have noticed that the choice of m does not appear to strongly affect the inferred ξ. This robustness with respect to m is rather reassuring and contrasts with the other classical option: selecting a threshold for the GPD model can, often, change a lot the estimated ξ.

Figure 7: Chartres weather station: estimated qq plots (left panels) and estimated Bernstein densities $\hat{g}_{m,n}(u)$ (right panels) for $m = 5$ (upper panels), $m = 15$ (middle panels) and $m = 30$ (lower panels)



Concerning future work, our semiparametric EGPD model can be viewed as a "building block" for more complex statistical models (such as rainfall weather generators). In particular, the coupling with precipitation occurrences models could be very fruitful for assessment studies, like sensitivity of floods, erosion or crops models. On the same token, the modeling of the entire-range of precipitation amounts at multiple sites would be a welcome addition. More specifically, Evin *et al.* (2016) showed that a regional model can considerably improve the estimation of the GPD shape parameter.

Acknowledgement

The analysis presented in the case studies (simulation and rainfall data) of this work was performed in R Software (code provided upon request). P. Naveau would like to thank the LJK and IGE labs, as well as the Inria project/team AIRSEA, for hosting him for a month in Grenoble. This work has been partially supported by the Labex Persyval-Lab (ANR-11- LABX-0025-01) funded by the French program "Investissements d'avenir" through the exploratory project STAREX. program *Investissement d'avenir*. Part of this work was supported by the French national program LEFE/INSU-FRAISE, LEFE/INSU-Multirisk ERC-A2C2 and EUPHEME projects. This work contributes to the CDP-Trajectories project, supported by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02) The authors acknowledge Meteo France for the rainfall dataset, that is available upon request. In addition, we wish to thank Valérie Monbet for her insightful suggestions and discussions.

References

Apipattanavis S, Podestã G, Rajagopalan B, Katz RW, 2007. A semiparametric multivariate and multisite weather generator. *Water Resources Research* **43**(11), w11401.

Babu G, Canty A, Chaubey Y, 2002. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference* **105**(2): 377–392.

Bernstein S, 1912. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Communications de la Société mathématique de Kharkow* **13**(1): 1–2.

Bouezmarni T, Rolin J, 2007. Bernstein estimator for unbounded density function. *Journal of Nonparametric Statistics* **19**(3): 145–161.

Carreau J, Bengio Y, 2008. A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes* **12**(1): 53–76.

Carreau J, Naveau P, Neppel L, 2017. Partitioning into hazard subregions for regional peaks-over- threshold modeling of heavy precipitation. *Water Resources Research* **53**: 4407–4426.

Coles S, 2001. *An introduction to statistical modeling of extreme values*. Springer, 224 pp.

Cooley D, Nychka D, Naveau P, 2007. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association* **102**(479): 824–840.

Davison AC, Padoan SA, Ribatet M, 2012. Statistical modeling of spatial extremes, with discussion. *Statist. Sci.* **27**(2): 161–186.

Evin G, Blanchet J, Paquet E, Garavaglia F, Penot D, 2016. A regional model for extreme rainfall based on weather patterns subsampling. *Journal of Hydrology* **541**: 1185–1198.

Evin G, Favre AC, Hingray B, 2018. Stochastic generation of multi-site daily precipitation focusing on extreme events. *Hydrology and Earth System Sciences* **22**(1): 655–672.

Farouki R, 2012. The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design* **29**(6): 379–419.

Fawcett L, Walshaw D, 2007. Improved estimation for temporally clustered extremes. *Environmetrics* **18**(2): 173–188.

Furrer EM, Katz RW, 2008. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research* **44**(12), w12439.

Garavaglia F, Gailhard J, Paquet E, Lang M, Garcon R, Bernardara P, 2010. Introducing a rainfall compound distribution model based on weather patterns sub-sampling. *Hydrology and Earth System Sciences* **14**(6): 951–964.

Ghosal S, 2001. Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics* **29**(5): 1264–1280.

Hegerl G, Zwiers F, 2011. Use of models in detection and attribution of climate change. *Wiley interdisciplinary reviews: climate change* **2(4)**: 570–591.

Hosking JR, Wallis JR, 1987. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics* **29**(3): 339–349.

Ji Y, Wu C, Liu P, Wang J, Coombes K, 2005. Applications of beta-mixture models in bioinformatics. *Bioinformatics* **21**(9): 2118–2122.

Kakizawa Y, 2004. Bernstein polynomial probability density estimation. *Journal of Nonparametric Statistics* **16**(5): 709–729.

Katz R, 1977. Precipitation as a chain-dependent process. *Journal of Applied Meteorology* **16**(7): 671–676.

Katz R, Parlange M, Naveau P, 2002. Statistics of extremes in hydrology. *Advances in Water Resources* **25**(8): 1287–1304.

Leblanc A, 2010. A bias-reduced approach to density estimation using Bernstein polynomials. *Journal of Nonparametric Statistics* **22**(4): 459–475.

Leblanc A, 2012a. On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathematics* **64**(5): 919–943.

Leblanc A, 2012b. On the boundary properties of Bernstein polynomial estimators of density and distribution functions. *Journal of Statistical Planning and Inference* **142**(10): 2762–2778.

López-Pintado S, Romo J, 2009. On the concept of depth for functional data. *Journal of the American Statistical Association* **104**(486): 718–734.

MacDonald A, Scarrott C, Lee D, Darlow B, Reale M, Russell G, 2011. A flexible extreme value mixture model. *Computational Statistics & Data Analysis* **55**: 2137–2157.

Nadarajah S, 2005. Extremes of daily rainfall in west central Florida. *Climatic Change* **69**(2-3): 325–342.

Naveau P, Huser R, Ribereau P, Hannart A, 2016. Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research* **52**(4): 2753–2769.

Petrone S, 1999. Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics* **27**(1): 105–126.

Ribereau P, Naveau P, Guillou A, 2011. A note of caution when interpreting parameters of the distribution of excesses. *Advances in Water Resources* **34**(10): 1215 – 1221.

Richardson C, 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research* **17**(1): 182–190.

Stern R, Coe R, 1984. A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)* **147**(1): 1.

Sun Y, Genton MG, 2011. Functional boxplots. *Journal of Computational and Graphical Statistics* **20**(2): 316–334.

Vitale R, 1975. A Bernstein polynomial approach to density function estimation. *Statistical inference and related topics* **2**: 87–99.

Vrac M, Naveau P, 2007. Stochastic downscaling of precipitation: from dry events to heavy rainfalls. *Water Resources Research* **43**(7): 1–13.

Vrac M, Stein M, Hayhoe K, 2007. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research* **34**(3): 169–184.

Wilks D, 1989. Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resources Research* **25**(6): 1429–1439.

Wilks D, 1999. Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology* **93**(3): 153–169.

Wilks D, 2011. *Statistical methods in the atmospheric sciences*. Academic Press, 676 pp.

Woolhiser D, Pegram G, 1979. Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *Journal of Applied Meteorology* **18**(1): 34–42.

Zucchini W, Adamson P, 1984. *The occurrence and severity of droughts in South Africa*. South Africa Water Research Commission.

[Proof of Lemma 1]

1. We write:

$$\frac{\hat{G}_{m,n}\left\{H_\xi(x/\sigma)\right\}}{x^s} = \frac{\hat{G}_{m,n}\left\{v\frac{H_\xi(v)}{v}\right\}}{\hat{G}_{m,n}(v)}\frac{\hat{G}_{m,n}(v)}{v^s}\sigma^{-s}$$

   where $v = x/\sigma$. Note that $\lim_{v\to 0}\frac{H_\xi(v)}{v} = 1$. Thus, from the polynomial assumption on our model – $\hat{G}_{m,n}(t) = \sum_{k=0}^{m} G_n(\frac{k}{m})b_{k,m}(t)$, $t \in [0,1]$ – we deduce:

$$\lim_{x\to 0}\frac{\hat{G}_{m,n}\left\{H_\xi(x/\sigma)\right\}}{x^s} = \lim_{v\to 0}\frac{\hat{G}_{m,n}\left\{v\frac{H_\xi(v)}{v}\right\}}{\hat{G}_{m,n}(v)}\frac{\hat{G}_{m,n}(v)}{v^s}\sigma^{-s} = \sigma^{-s}\lim_{v\to 0}\frac{\hat{G}_{m,n}(v)}{v^s}.$$

   Then, from l'Hôpital's rule, and from Equation (2),

$$\lim_{v\to 0}\frac{\hat{G}_{m,n}(v)}{v^s} = \lim_{x\to 0}\frac{\hat{g}_{m,n}(v)}{sv^{s-1}}$$

   is equivalent to $\frac{m\binom{m-1}{s-1}\omega_{s,m}v^{s-1}}{sv^{s-1}} = \frac{m\binom{m-1}{s-1}\omega_{s,m}}{s}$ with s the position of the first non-null weight in $\boldsymbol{\omega}$.

2. Let $u = \overline{H}_\xi(x/\sigma)$. We have

$$\lim_{x\to\infty}\frac{\overline{\hat{G}}_{m,n}\left\{H_\xi(x/\sigma)\right\}}{\overline{H}_\xi(x/\sigma)} = \lim_{u\to 0}\frac{\overline{\hat{G}}_{m,n}(1-u)}{u}.$$

Then, by applying l'Hôpital's rule, we get

$$\lim_{u\to 0}\frac{\overline{\hat{G}}_{m,n}(1-u)}{u} = \lim_{v\to 0}\hat{g}_{m,n}(1-u) = \hat{g}_{m,n}(1) = m\omega_{m,m}.$$

Therefore, to force the upper tail behavior in our model, we assume $\omega_{m,m} > 0$.

3. As the random variable Y can be written (in distribution) as $Y = \sigma H_\xi^{-1}\left\{G^{-1}(U)\right\}$, where U follows an uniform distribution on $[0,1]$, *i.e.,* $\mathbb{P}(U > w) = 1 - w$ for any $w \in [0,1]$, it follows that

$$\begin{aligned}
\mathbb{P}(Y > x+u | Y > u) &= \frac{\mathbb{P}(\sigma H_\xi^{-1}\left\{G^{-1}(U)\right\} > x+u)}{\mathbb{P}(\sigma H_\xi^{-1}\left\{G^{-1}(U)\right\} > u)},\\
&= \frac{\mathbb{P}(U > G\left[H_\xi\left\{(x+u)/\sigma\right\}\right])}{\mathbb{P}(U > G\left\{H_\xi(u/\sigma)\right\})},\\
&= \frac{1 - G\left[H_\xi\left\{(x+u)/\sigma\right\}\right]}{1 - G\left\{H_\xi(u/\sigma)\right\}},\\
&= \frac{\overline{G}(1-w)}{\overline{G}(1-w^*)}, \text{ with } w = \overline{H}_\xi\left\{(x+u)/\sigma\right\} \text{ and } w^* = \overline{H}_\xi(u/\sigma),\\
&= \frac{\overline{G}(1-w)}{w}\frac{w^*}{\overline{G}(1-w^*)}\frac{w}{w^*}.
\end{aligned}$$

We assume that Y follows the classical EVT theory, *i.e.,* $\mathbb{P}(Y > x+u|Y > u)$, goes to $(1+\tilde{\xi}\frac{x}{\tilde{\sigma}})^{-1/\tilde{\xi}}$, as u gets large. Our constraint on G implies that $\frac{\overline{G}(1-w)}{w}\frac{w^*}{\overline{G}(1-w^*)} \to 1$ for large u. So, the righthand side behaves as $\frac{w}{w^*} = (1+\xi\frac{x}{\sigma_u})^{-1/\xi}$ with $\sigma_u = \sigma + \xi u$. This implies that $\tilde{\xi} = \xi$ and $\tilde{\sigma} = \sigma_u$ for large u.