



**HAL**  
open science

# Characterization of homogeneous regions for regional peaks-over-threshold modeling of heavy precipitation

Julie Carreau, Philippe Naveau, Luc Neppel

► **To cite this version:**

Julie Carreau, Philippe Naveau, Luc Neppel. Characterization of homogeneous regions for regional peaks-over-threshold modeling of heavy precipitation. 2016. ird-01331374

**HAL Id: ird-01331374**

**<https://ird.hal.science/ird-01331374>**

Preprint submitted on 13 Jun 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Characterization of homogeneous regions for regional peaks-over-threshold modeling of heavy precipitation

JULIE CARREAU<sup>a</sup>

PHILIPPE NAVEAU<sup>b</sup>

LUC NEPPEL<sup>a</sup>

Julie.Carreau@univ-montp2.fr, (phone) +33467149027, (fax) +33467144774

<sup>a</sup> HydroSciences Montpellier, CNRS/IRD/UM, Université de Montpellier - Case 17, 163 rue Auguste Broussonet 34090 Montpellier, France

<sup>b</sup> LSCE, IPSL-CNRS, Orme des Merisiers, Gif-sur-Yvette, France

## Abstract

In the French Mediterranean area where heavy precipitation events can yield devastating consequences, it is essential to obtain reliable estimates of the distribution of extreme precipitation at gauged and ungauged locations. Under mild assumptions, extremes defined as excesses over a high enough threshold can be modeled by the generalized Pareto (GP) distribution. The shape parameter of the GP which characterizes the behavior of extreme events is notoriously difficult to estimate. In regional analysis, the sample variability of the shape parameter estimate can be reduced by increasing the sample size. This is achieved by assuming that sites in a so-called homogeneous region are identically distributed apart from a scaling factor and therefore share the same shape parameter. A major difficulty is the proper definition of homogeneous regions. We build upon a recently proposed approach, based on the probability weighted moment (PWM) for the GP distribution, that can be cast into a regional framework for a single homogeneous region. Our main contribution is to extend its applicability to complex regions by characterizing each site with the second PWM of the scaled excesses. We show on synthetic data that this new characterization is successful at identifying the homogeneous regions of the generative model and leads to accurate GP parameter estimates. The proposed framework is applied to 332 daily precipitation stations in the French Mediterranean area which are splitted into homogeneous regions with shape parameter estimates ranging from 0 to 0.3. The uncertainty of the estimators is evaluated with an easy-to-implement spatial block bootstrap.

**keywords :** regional analysis, probability weighted moment, generalized Pareto distribution, spatial block bootstrap, extreme precipitation, French mediterranean area, clustering

# 1 Introduction

Flash floods, a sudden rise of the water level (in a few hours or less) together with a significant peak discharge, are the main natural hazard in the French Mediterranean area. They can potentially cause fatalities and important material damage [Borga et al., 2011, Braud et al., 2014]. Flash floods might be triggered by intense rainfall events occurring mainly in the fall [Delrieu et al., 2005]. Therefore, to design infrastructure to mitigate the impacts of these natural hazards, a reliable estimation of the distribution of extreme precipitation events is crucial, both at gauged and ungauged locations.

Extreme value theory [Coles, 2001] provides a sound asymptotic framework to model the distribution of extremes. In particular, the extremal-type theorem [Fisher and Tippett, 1928, Gnedenko, 1943] states that if the distribution of properly re-scaled maxima converges to a nondegenerate distribution, this distribution is the generalized Extreme Value (GEV) distribution. This gives rise to the block maxima approach which consists of fitting the GEV to the maxima extracted from sufficiently large blocks of observations, often taken as years. Moreover, provided that the maxima converge in distribution to the GEV, the distribution of the excesses above a threshold converges to the generalized Pareto (GP) distribution [Balkema and de Haan, 1974, Pickands, 1975]. The so-called peaks-over-threshold (PoT) approach proceeds by setting a sufficiently high threshold and estimating the GP parameters from the excesses above that threshold. The PoT approach is often preferred over the block maxima approach as more observations can be included in the analysis and this might reduce the estimator variance [Roth et al., 2012]. However, Ferreira and de Haan [2015] showed that the block maxima approach can be rather efficient. In both approaches, the shape parameter of the GEV or the GP characterizes the behavior of the upper tail of the distribution that governs the probability of extreme events.

While the block maxima and the PoT approaches require rather long sample and may be employed solely at gauged sites, regional analysis developed a robust framework to estimate the distribution of extreme events at ungauged sites or sites with short record length. In order to increase the sample size at a given target site, extreme events at neighboring sites are assumed to have the same distribution apart from a scaling factor. Regions for which this assumption is valid are termed *homogeneous*. These regions can be either contiguous and form a partition, or overlapping, defined as neighborhoods around each target site as in the region of influence approach [Burn, 1990]. For a given target site, that could be ungauged or with short record length, regional analysis involves the following steps [Hosking and Wallis, 2005]. First, a homogeneous region, to which the target site belongs, must be defined. Second, observations at each gauged site in the homogeneous region are normalized, i.e. divided by the site-specific scaling factor. Then, the regional distribution is fitted to the normalized observations from all the gauged sites in the region. Next, the scaling factor is interpolated (or estimated locally if enough observations are available) at the target site. Return levels are obtained as the product of the return levels of the regional distribution and the scaling factor of the target site.

Combining an extreme-value approach with regional analysis allows to interpolate at ungauged locations and to decrease the uncertainty of the estimation of the shape parameter which is central in assessing the risk of extreme precipitation events. Indeed, to satisfy the homogeneity assumption, the shape parameter has to be constant across the region and the normalized observations from all the sites contribute to the estimation. For instance, Kysely et al. [2011] and Carreau et al. [2013] employed the GEV in a regional analysis of annual maxima of precipitation while Madsen

1 and Rosbjerg [1997] and Roth et al. [2012] used the GP for threshold excesses. The latter combi-  
2 nation leads to the potentially largest increase in sample size. Besides regional analysis, other ways  
3 to exploit information of gauged sites that are similar in distribution in order to interpolate or to  
4 strengthen the distribution of extreme events have been proposed. For instance, a direct interpo-  
5 lation of the site parameter estimates or a regression model for the GEV of GP parameters either  
6 in a frequentist [Blanchet and Lehning, 2010, Ceresetti et al., 2012] or bayesian approach [Cooley  
7 et al., 2007, Renard, 2011] have been considered.

8 As noted in Renard [2011], the identification of homogeneous regions can be seen as a limita-  
9 tion of the regional approach. Indeed, it generally involves several steps some of which call for  
10 subjective decisions, see Burn and Goel [2000], Kysely et al. [2007, 2011] for instance. The rec-  
11 ommended approach, described in Hosking and Wallis [2005], advocates the use of physiographic  
12 variables such as geographical and climatological characteristics to identify the regions. After the  
13 initial identification, the regions are tested for homogeneity by resorting to L-Moment ratios. Het-  
14 erogeneous regions are re-defined until they pass the homogeneity test. An alternative approach  
15 to identify homogeneous regions, see for example Durocher et al. [2016] and the references therein,  
16 seeks to model the relationship between physiographic variables, available at all sites, and hydrolog-  
17 ical variables, available only at gauged sites. Additional shortcomings of regional analysis concern  
18 (i) the potential invalidity of the scale invariance assumption of the regional distribution which  
19 implies that the normalized observations from a given homogeneous region have a constant scale  
20 parameter, (ii) the lack of physical reason behind the definition of the scaling factor and (iii) the  
21 difficulty to evaluate the uncertainty of the estimators partly as a result of the spatial dependence  
22 of the observations [Gupta et al., 1994, Renard, 2011, Van de Vyver, 2012].

23 In this work, we build on the approach recently proposed in Naveau et al. [2014] to address  
24 some of the shortcomings of the regional approach. The Naveau et al. [2014] approach, which  
25 rely on probability weighted moments for the GP distribution [Diebolt et al., 2007], can easily  
26 be cast into the regional framework with a single homogeneous region. As is the case with L-  
27 Moments, probability weighted moments estimates are fast to compute and may serve as starting  
28 values to estimation procedures that require an optimization scheme (such as maximum-likelihood  
29 or Bayesian estimators). Moreover, the Naveau et al. [2014] approach does not need to enforce the  
30 scale invariance assumption of the normalized observations since it is automatically fulfilled thanks  
31 to the choice of scaling factor. The main contribution of this paper is to propose a characterization  
32 of each site based on probability weighted moments that stems straightforwardly from the Naveau  
33 et al. [2014] approach and extends its applicability to complex area with several homogeneous  
34 regions. Sites that are characterized with similar values belong to the same homogeneous region.  
35 Based on this notion of similarity, homogeneous regions can be identified either as contiguous  
36 regions by employing a clustering algorithm or as overlapping regions resulting from neighborhoods  
37 around target sites. We focus on the former option and rely on the K-Means algorithm to partition  
38 the sites into homogeneous regions. The k-nearest neighbor rule is used to assign ungauged sites  
39 to a homogeneous region (see Ripley [1996] for a detailed presentation of both K-means and the  
40 k-nearest neighbor rule). Partitions of homogeneous regions associated to risk levels linked to  
41 the shape parameters of the regional GP distribution may be useful for operational early warning  
42 system such as <http://vigilance.meteofrance.com/>. Lastly, the sampling distribution of the  
43 GP parameter estimates of the proposed regional framework, from which uncertainty estimates  
44 can be deduced, is obtained with an easy-to-implement spatial block bootstrap.

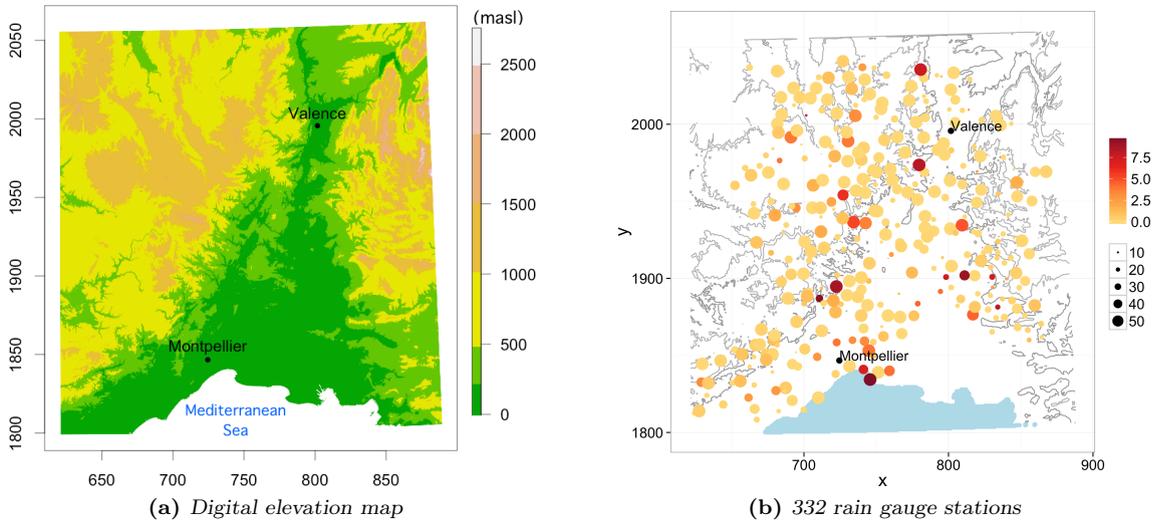
45 The article is organised as follows. The precipitation data which motivates this work is presented

1 in Section 2. We then describe the basic framework, in Section 3, to set up the methodology which  
 2 is the basis of the regional framework presented in Section 4. The latter contains a simulation  
 3 study, Section 4.4, and both the basic and the regional framework are applied to the precipitation  
 4 data, Section 3.4 and 4.5 respectively. In Section 5, results are discussed and some conclusions are  
 5 drawn.

## 6 2 Daily precipitation data

7 The French Mediterranean area is subject to intense rainfall events which may trigger floods and  
 8 landslide with dramatic human and material consequences [Delrieu et al., 2005, Braud et al.,  
 9 2014]. The occurrence of these intense rainfall events and their high spatial variability are due to  
 10 the combination of the Mediterranean climate with the complex orography of the region.

11 Daily precipitation at 332 stations over the period 01/01/1958 to 31/12/2014 (57 years) were  
 12 collected by Météo-France, the French weather service. Stations are located in the French Mediter-  
 13 ranean area whose orography can be seen from the digital elevation map in Fig. 1a. The 332  
 14 stations are depicted in Fig. 1b. The size of the plotting symbol is proportional to the length of the  
 15 observation period available (from 10 to 57 years). The color indicates the percentage of missing  
 16 values over the observation period (from 0% in light orange to 10 % in dark red). The following  
 17 landmarks are depicted (and will be in the subsequent figures related to the precipitation data) :  
 18 two contour level curves, 400 m and 800 m, of the digital elevation map in dark and light shades  
 19 of gray respectively and two cities (Valence and Montpellier).



**Figure 1:** Region of the French Mediterranean area : orography (left) and rain gauge stations (right). In the latter figure, the size of the symbol is proportional to the length of the observation period (10 to 57 years) and the color shade (light orange to dark red) indicates the percentage of missing values (0-10%).

## 20 3 Basic framework

21 We introduce the following notation. Let the  $M$  gauged sites in the region of interest be indexed  
 22 by  $i$ . In addition, let  $\mathbf{x}$  be a vector of covariates which is available at any site in the region, gauged

1 or ungauged.

## 2 **3.1 Kernel regression**

3 In all the approaches developed subsequently, whenever a quantity has to be interpolated, we use  
 4 kernel regression which is a non-parametric approach [Nadaraya, 1964, Watson, 1964]. The kernel  
 5 function  $K_h(\cdot)$  can be thought of as a symmetric density function for which the so-called *bandwidth*  
 6  $h$  acts as a scale parameter. The bandwidth controls the amount of smoothing in the interpolation.

7 More precisely, to interpolate a given quantity  $q(\cdot)$  with respects to covariates  $\mathbf{x}$ , we proceed in  
 8 two steps. First, for each site  $i$ ,  $q(\mathbf{x}_i)$  is estimated locally, i.e. based on the observations at site  $i$ .  
 9 Second, the following weighted average corresponds to the interpolated value for the covariates  $\mathbf{x}$  :

$$\tilde{q}(\mathbf{x}) = \frac{1}{\sum_i K_h(\mathbf{x} - \mathbf{x}_i)} \sum_{i=1}^M \hat{q}(\mathbf{x}_i) K_h(\mathbf{x} - \mathbf{x}_i) \quad (1)$$

10 where  $\hat{q}(\mathbf{x}_i)$  is a local estimate at site  $i$ .

11 We rely on the implementation in the `np` package of R [Hayfield and Racine, 2008]. It imple-  
 12 ments kernel regression with various types of kernels and several automated bandwidth selection  
 13 methods. We employed the Epanechnikov kernel which is optimal in the sense that it minimizes  
 14 the asymptotic mean integrated square error [Epanechnikov, 1969, Abadir and Lawford, 2004].  
 15 Bandwidth selection is performed before each spatial interpolation (see Li and Racine [2004] and  
 16 the documentation in the `np` package [Hayfield and Racine, 2008]).

## 17 **3.2 Generalized Pareto tail approximation**

18 Under mild assumptions, the generalized Pareto (GP) distribution can be used as an approximation  
 19 to the upper tail of the distribution of most random variables [Pickands, 1975]. In other words,  
 20 given a high enough threshold  $u$  suitably chosen, the GP distribution approximates the distribution  
 21 of the excesses over  $u$ . Let  $Y \sim G(\sigma, \xi)$  be a random variable representing the excesses that follows  
 22 a GP distribution with scale parameter  $\sigma > 0$  and shape parameter  $\xi \in \mathbb{R}$ . The survival function  
 23 of  $Y$  is provided in Eq. (2) and obey the following domain restrictions :  $y \geq 0$  when  $\xi \geq 0$   
 24 and  $y \in [0, -\sigma/\xi)$  when  $\xi < 0$ . The shape parameter describes the upper tail behavior : heavy  
 25 (Pareto-type) when  $\xi > 0$ , light (exponential) when  $\xi = 0$  or with a finite upper bound for  $\xi < 0$ .

$$\mathbb{P}(Y > y) = \bar{G}(y; \sigma, \xi) = 1 - G(y; \sigma, \xi) = \begin{cases} (1 + \xi \frac{y}{\sigma})^{-1/\xi} & \text{if } \xi \neq 0 \\ \exp(-\frac{y}{\sigma}) & \text{if } \xi = 0. \end{cases} \quad (2)$$

26 High quantiles associated to long return periods such as 100 years are often used by practioners  
 27 for risk assessment. Let  $l(T)$  be the quantile, also termed return level, with a return period of  
 28  $T$  years, i.e. the level that is exceeded on average once every  $T$  years. Thanks to the GP tail  
 29 approximation,  $l(T)$  can be estimated as a quantile of the GP distribution as follows :

$$l(T) = \begin{cases} u + \frac{\sigma}{\xi} ((T N_{exc})^\xi - 1) & \text{if } \xi \neq 0 \\ u + \sigma \log(T N_{exc}) & \text{if } \xi = 0 \end{cases} \quad (3)$$

30 provided that  $l(T)$  is greater than the threshold  $u$  and where  $N_{exc} = 365.25 \zeta_u$  is the average  
 31 number of excesses per year with  $\zeta_u$  the probability of exceeding the threshold  $u$ .

### 3.3 Probability weighted moment estimators

We develop expressions to estimate the parameters of the GP distribution based on the probability weighted moments. The introduction of a normalized variable  $Z$  enables a straightforward extension to the framework of regional analysis.

For  $r \geq 0$ , the probability weighted moments for the GP distribution are given by [Diebolt et al., 2007] :

$$\mathbb{E}[Y \overline{G}(Y; \sigma, \xi)^r] = \sigma \frac{1}{(1+r)(1+r-\xi)}. \quad (4)$$

Sample estimates can be computed using U-statistics, see Furrer and Naveau [2007]. In particular, if  $\xi < 1$ , the first probability weighted moment, obtained with  $r = 0$  in Eq. (4), is the expectation of  $Y$  and can be written as :

$$\mu = \mathbb{E}[Y] = \frac{\sigma}{1-\xi}. \quad (5)$$

We consider  $\mu$  as the scaling factor. In other words, let  $Z = Y/\mu$  be the normalized variable. Since  $\mathbb{P}(Z > z) = \mathbb{P}(Y > \mu z) = \mathbb{P}(Y > (\sigma z)/(1-\xi))$ , where the last equality follows by making use of Eq. (5), we have, by replacing  $y$  with  $(\sigma z)/(1-\xi)$  in Eq. (2), that  $Z \sim G(1-\xi, \xi)$ . Therefore, the normalized variable  $Z$  only depends on the shape parameter  $\xi$ .

Let  $\nu$  be the second probability weighted moment of  $Z$  obtained by plugging  $\sigma = 1 - \xi$  and  $r = 1$  in Eq. (4) :

$$\nu = \frac{1-\xi}{4-2\xi}. \quad (6)$$

To estimate the shape parameter of the GP distribution, we then replace  $\nu$  by its estimator and solve for  $\xi$  :

$$\hat{\xi} = \frac{1-4\hat{\nu}}{1-2\hat{\nu}}. \quad (7)$$

The scale parameter is estimated by solving Eq. (5) for  $\sigma$  and replacing  $\mu$  and  $\xi$  by their sample estimates  $\hat{\mu}$  and  $\hat{\xi}$  respectively :

$$\hat{\sigma} = \hat{\mu}(1-\hat{\xi}). \quad (8)$$

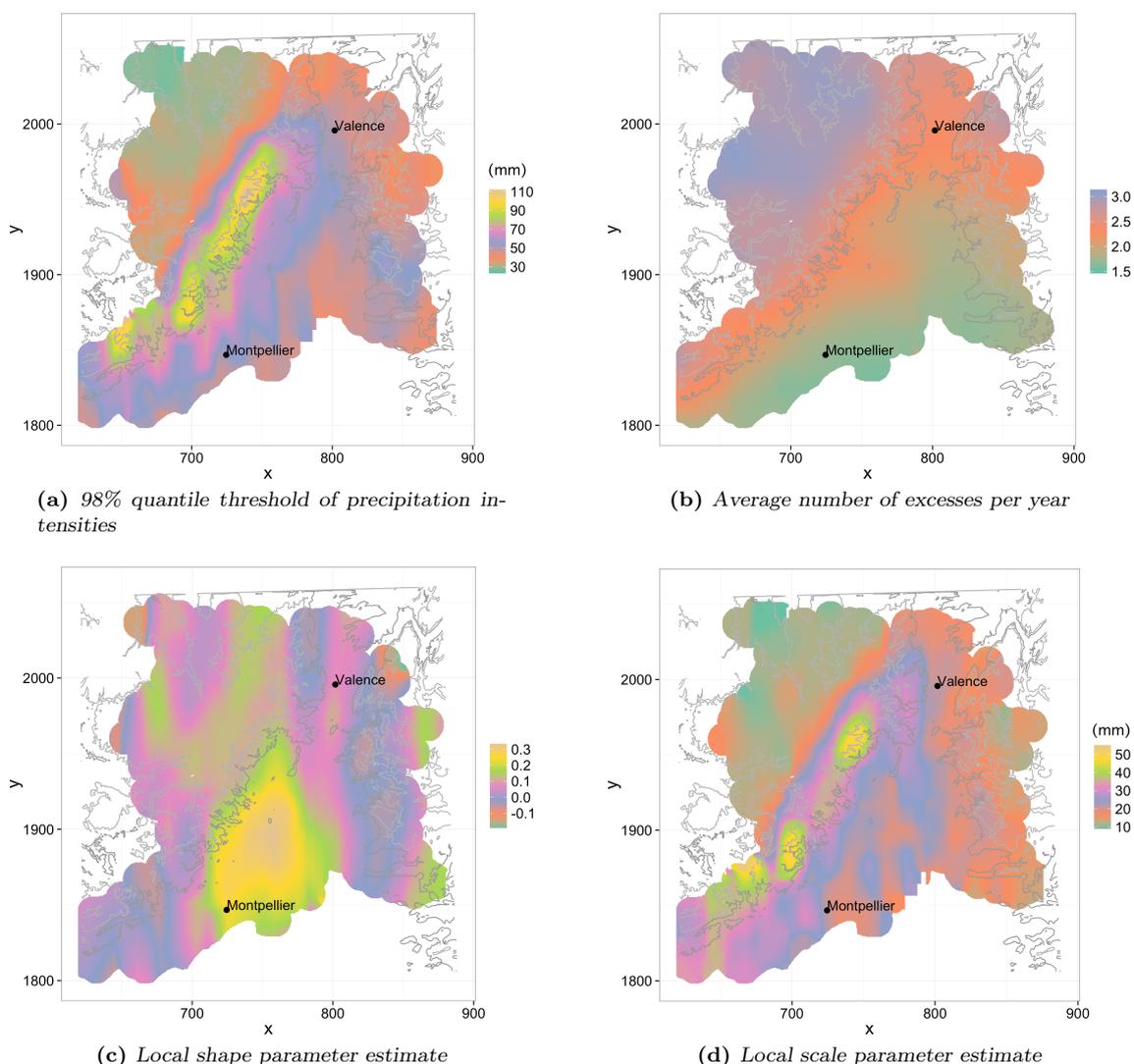
Eq. (7) and (8) are used to estimate the shape and scale parameters in the basic and the regional framework developed in Section 4. In the former, the sample with which the shape parameter is estimated is formed only from the normalized observations from the target site. In the latter, the sample can include normalized observations from all the sites in the homogeneous neighborhood of the target site.

### 3.4 Preliminary analysis of the daily precipitation data

We apply the basic framework to the French Mediterranean precipitation data described in Section 2. We set up a regular grid (approximately 500 m) covering the region where the stations lie on which the interpolation is carried out. For this application,  $M = 332$ , the number of gauged sites, and  $\mathbf{x}$  is taken as the x and y coordinates (extended Lambert II projections of latitude and longitude).

1 The threshold that defines the excesses for which the GP tail approximation is used is set to  
 2 the 98% quantile of the precipitation intensities, i.e. the observations greater than 0.1 mm (the  
 3 sensitivity of daily rain gauges). The threshold and the average number of excesses per year, see  
 4 Section 3.2, are computed at each station resulting in an overall number of excesses per station  
 5 ranging from 20 to 191. Local GP parameter estimates are obtained thanks to Eqs. (7-8).

6 All four local estimates (threshold, average number of excesses per year and shape and scale  
 7 parameters) are interpolated with kernel regression, see Section 3.1, onto the regular grid, see  
 8 Fig 2a-2d. The interpolated threshold and average number of excesses define the tail approximation  
 9 of the GP and will be used also when the regional framework introduced in Section 4 is applied to  
 10 the French Mediterranean precipitation data. The interpolated shape and scale parameters of the  
 11 GP will be compared with the estimates from the regional framework.



**Figure 2:** Application of the basic framework to the French Mediterranean precipitation data. The threshold and resulting average number of excesses per year together with the GP shape and scale parameters are estimated locally and then interpolated with kernel regression.

## 4 Regional framework

### 4.1 Single homogeneous region

We apply the expressions in Section 3.3 to the estimation of the GP parameters in the regional framework stemming from the Naveau et al. [2014] approach with a single homogeneous region, i.e. all the sites belong to the same homogeneous region.

In this work, a region is called homogeneous if the shape parameter is constant over the region and the scale parameter varies smoothly spatially as a function of a vector of covariates  $\mathbf{x}$ . In other words, for a given site  $i$ , the distribution of the excesses is given as :

$$Y_i \sim G(\sigma(\mathbf{x}_i), \xi), \quad (9)$$

where  $\xi$  can be thought of as a *regional* shape parameter. As a result, the scaling factor, that is the expectation of  $Y_i$ , also varies spatially since, from Eq. (5),  $\mu(\mathbf{x}_i) = \sigma(\mathbf{x}_i)/(1-\xi)$ .

Although the scaling factor varies spatially, the normalized variable  $Z_i = Y_i/\mu(\mathbf{x}_i)$  is identically distributed over the region. Indeed, for all  $i$ ,  $Z_i \sim G(1-\xi, \xi)$ , i.e. the normalized variable only depends on the regional shape parameter  $\xi$ , as in the basic framework in Section 3.3. By construction, the regional shape parameter is constant over the region. It follows that the scale parameter of the normalized variable is also constant. Therefore, the scale invariance assumption, mentioned in the introduction, is automatically fulfilled without any further assumptions on the scale parameter.

The observed excesses from all the sites in the region, once normalized by their expectation, can be used to estimate the regional shape parameter with Eq. (7). Hence, as in the classical regional approach, the sample variability of the estimator of the shape parameter is reduced thanks to an increased sample size. The scale parameter is estimated as before with Eq. (8).

The approach described in this section is related to the work in Naveau et al. [2014] but differs in two main respects. First, they considered  $\sigma(\mathbf{x}_i)$  as the scaling factor instead of  $\mu(\mathbf{x}_i)$  which implies that  $Z \sim G(1, \xi)$ . However, in such a case  $Z$  is not observable because  $\sigma(\mathbf{x}_i)$  is unknown. To circumvent this problem, Naveau et al. [2014] normalized the observations with  $\mu(\mathbf{x}_i)$  and account for the difference between  $\sigma(\mathbf{x}_i)$  and  $\mu(\mathbf{x}_i)$  in their estimators of the GP parameters. Second, in Naveau et al. [2014], both the second and the third probability weighted moments are employed in the estimators. In this work, we adopt right away  $\mu(\mathbf{x}_i)$  as the scaling factor and use only the second probability weighted moment of the normalized variable  $Z$  to estimate  $\xi$ . These choices lead to simpler expressions.

### 4.2 Characterization of homogeneous regions

As can be seen from the local estimation of the shape parameter in the basic framework in Fig. 2c, the assumption of constant shape parameter and thus, of a single homogeneous region, is not reasonable for the French Mediterranean precipitation application. Building on the expressions in Section 4.1, we introduce a characterization of each site with which homogeneous regions can be defined. We partition the sites into  $N_{reg}$  contiguous regions, i.e. each site belongs to one region. The so-called “region of influence” approach [Burn, 1990] could also be used as discussed in Section 5.

Let  $C_i \in \{1, \dots, N_{reg}\}$  be the homogeneous region label associated to site  $i$  and let  $\{\xi_1, \dots, \xi_{N_{reg}}\}$

1 be the regional shape parameter associated to each homogeneous region. For a site  $i$  belonging  
 2 to the region  $C_i$ ,  $Y_i \sim G(\sigma(\mathbf{x}_i), \xi_{C_i})$  and  $Z_i \sim G(1 - \xi_{C_i}, \xi_{C_i})$ . The regional shape parameter  $\xi_{C_i}$   
 3 and hence the tail behavior of  $Y_i$  varies according to the region  $C_i$ . The higher the regional shape  
 4 parameter is, the greater the risk of extreme precipitation events in the region.

5 Since the normalized variable only depends on the regional shape parameter, we propose to  
 6 characterize the site  $i$  with a statistic of  $Z_i$ . In this work, we choose to use  $\nu$ , the second probability  
 7 weighted moment of the normalized variable  $Z_i$ , see Eq. (6), to summarize the information on the  
 8 tail behavior of a given site  $i$ . We have that  $\nu(\mathbf{x}_i) = \nu(\mathbf{x}_j)$  if and only if  $C_i = C_j$ . For each site  $i$ ,  
 9 let  $\hat{\nu}(\mathbf{x}_i)$  be the estimation of  $\nu$ . This characteristic can be fed to a clustering algorithm to identify  
 10 the homogeneous regions. In this work, we resort to K-Means to perform the clustering [Ripley,  
 11 1996]. K-Means iteratively assigns a site  $i$  to the cluster whose cluster center is closer in terms of  
 12  $\hat{\nu}(\mathbf{x}_i)$  and then re-computes the cluster centers as the averages of the  $\hat{\nu}(\mathbf{x}_i)$  of the sites belonging to  
 13 each cluster. We set the initial cluster centers to  $N_{reg}$  empirical quantiles of  $\hat{\nu}(\mathbf{x}_i)$ ,  $1 \leq i \leq M$ , with  
 14 probabilities that spread regularly the  $[0, 1]$  interval. This ensures that K-Means always converges  
 15 to the same partition.

### 16 4.3 Estimation at ungauged sites

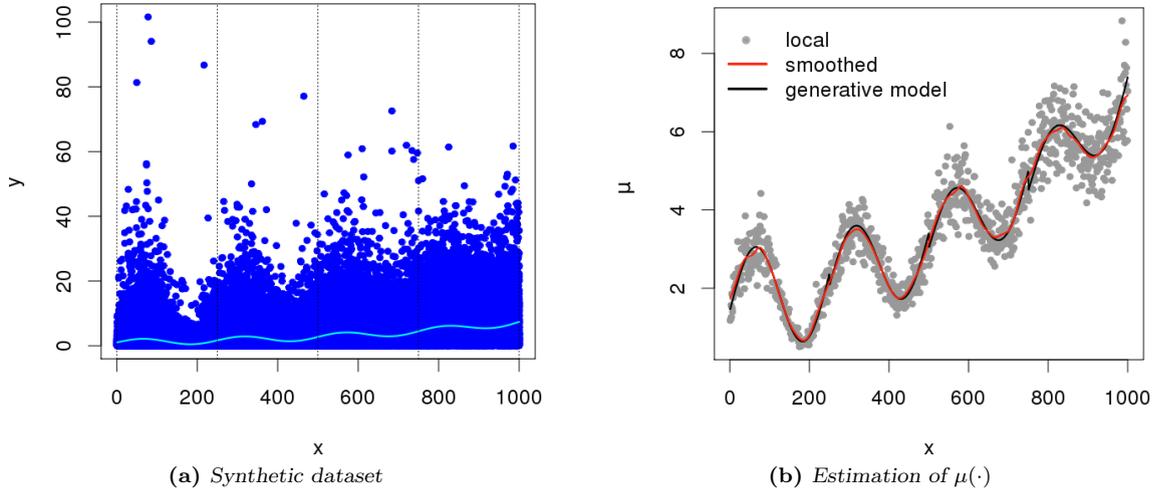
17 To estimate the GP parameters at an ungauged site  $i^*$ , we must first determine to which homoge-  
 18 neous region it belongs. This is a classification problem and we employ the k-nearest neighbor rule  
 19 with  $k = 5$ , a non-parametric classifier [Ripley, 1996]. This classifier determines the five nearest  
 20 neighbors of  $i^*$  by evaluating the Euclidean distances  $d(\mathbf{x}_{i^*}, \mathbf{x}_i)$  for all  $i \in \{1, \dots, M\}$  and assigns  
 21  $i^*$  to a class  $C_{i^*}$  by taking a majority vote among its five nearest neighbors. The regional shape  
 22 parameter at the site  $i^*$  is given by  $\xi_{C_{i^*}}$  and is estimated as explained in Section 4.1.

23 To estimate the scale parameter at site  $i^*$ , the scaling factor  $\mu(\mathbf{x}_{i^*})$  is interpolated as in the  
 24 classical regional approach. It is then combined with the regional shape parameter estimate in  
 25 Eq. (8).

### 26 4.4 Simulation study

27 We illustrate the regional framework described in this section on synthetic data whose generative  
 28 model satisfies the assumptions of the framework. The synthetic dataset is a variant, with four  
 29 homogeneous regions, of the dataset proposed in Naveau et al. [2014] which consists of a single  
 30 homogeneous region. The one-dimensional covariate  $x$  takes value in the interval  $[1, 1000]$  that is  
 31 splitted into four equal sub-intervals :  $\xi_1 = 0.3$  when  $x \in [1, 250]$ ,  $\xi_2 = 0.2$  when  $x \in [251, 500]$ ,  
 32  $\xi_3 = 0.1$  when  $x \in [501, 750]$  and  $\xi_4 = 0$  when  $x \in [751, 1000]$ . The scale parameter varies with  $x$   
 33 as a combination of a periodic and exponential signal. Fig. 3a shows a sample from the synthetic  
 34 dataset with  $M = 1000$  sites,  $x_i = i$  for  $i \in \{1, \dots, M\}$  and  $n_i = 100$  GP samples generated from  
 35 each  $Y_i$ . In Fig. 3a, the scale parameter is depicted as the cyan curve and the homogeneous regions  
 36 are indicated by the vertical bands.

37 A detailed description of the regional framework proposed in this paper is given in Algorithm 1.  
 38 The algorithm must be provided with the following inputs : the number of homogeneous regions  
 39  $N_{reg}$  and for each site  $1 \leq i \leq M$ , the  $n_i$  observed excesses  $\mathbf{y}_i = \{y_{i1}, \dots, y_{in_i}\}$  and the vector  
 40 of covariates  $\mathbf{x}_i$ . It is possible to provide additional vectors of covariates  $\mathbf{x}_{i^*}$  corresponding to  
 41 ungauged sites  $1 \leq i^* \leq M^*$ . The outputs of the algorithm are the  $N_{reg}$  regional shape parameter  
 42 estimates  $\hat{\xi}_j$ ,  $1 \leq j \leq N_{reg}$ , the scale parameter estimates  $\hat{\sigma}(\mathbf{x}_i)$  and the region labels  $C_i$  for each



**Figure 3:** Left panel : A GP random sample is simulated in the interval  $[1, 1000]$  for 1000 sites with covariates taking value  $x = 1, \dots, 1000$ . The scale parameter varies as a combination of periodic and exponential signal (cyan curve) and the regional shape parameter is piecewise constant decreasing from 0.3, 0.2, 0.1 to 0 in each of the vertical bands. The sample size at each site is  $n_i = 100$ . Right panel : At each  $x$ , a local estimate  $\hat{\mu}_i$  of  $\mu$  is computed (gray dots) and then smoothed with kernel regression to obtain  $\hat{\mu}(\cdot)$  (red curve). The generative model of  $\mu(\cdot)$  is represented by the black curve.

1 site  $1 \leq i \leq M$ . If additional target sites are included in the inputs, the algorithm returns their  
 2 region labels and their scale parameter estimates as well. In the synthetic data application, there  
 3 is no ungauged site estimation. We set the number of regions to the value of the generative model,  
 4 that is  $N_{reg} = 4$ .

5 The first step of the regional framework proposed in this paper consists in estimating the  
 6 scaling factor (Fig. 3b) and computing the normalized observations (Fig 4a). This corresponds  
 7 in Algorithm 1 to lines 1 and 2 respectively. Kernel regression is applied to local estimates  $\hat{\mu}_i =$   
 8  $1/n_i \sum_{k=1}^{n_i} y_{ik}$  to obtain a smooth estimate of  $\mu(\cdot)$  (line 1). Then, the excesses at each site  $i$  are  
 9 normalized with the estimated  $\hat{\mu}(\cdot)$  yielding  $z_{ik} = y_{ik}/\hat{\mu}(x_i)$  for  $1 \leq k \leq n_i$  (line 2). In Fig. 3b, the  
 10 scaling factor of the synthetic data example is represented. The gray dots are the local estimates  
 11  $\hat{\mu}_i$ , the red curve represents the kernel regression estimate  $\hat{\mu}(\cdot)$  and the generative function  $\mu(\cdot)$  is  
 12 shown in black. In this example, the  $\mu(\cdot)$  function of the generative model has discontinuities at  
 13 the borders of the homogeneous regions and these cannot be well captured by kernel regression.  
 14 Fig 4a illustrates the normalized sample.

15 In the second step of the proposed framework, the gauged and ungauged sites are assigned to  
 16 a homogeneous region (Fig. 4b). If a single homogeneous region is requested, all the  $M$  gauged  
 17 sites and the  $M^*$  ungauged sites are pooled together (line 3 of Algorithm 1). Otherwise, the sites  
 18 are partitioned into  $N_{reg}$  regions (line 5). The partitioning goes as follows. Similarly as for  $\mu(\cdot)$ ,  
 19  $\nu(\cdot)$  is estimated (line 6) by applying kernel regression to the local estimates  $\nu_i$  computed from  $z_{ik}$ ,  
 20  $1 \leq k \leq n_i$  with U-statistics [Furrer and Naveau, 2007]. Then, the  $M$  gauged sites are clustered  
 21 into  $N_{reg}$  regions with K-Means based on  $\hat{\nu}(x_i)$  (line 7). If ungauged sites are provided as well,  
 22 they are assigned to a homogeneous region  $C_{i^*}$  for each  $i^*$  with a k-nearest neighbor classifier with  
 23  $k = 5$  (line 8). For the synthetic data application, Fig. 4b shows the  $\nu_i$  as gray dots, the regressed  
 24  $\hat{\nu}(\cdot)$  as colored curves and the function from the generative model  $\nu(\cdot)$  in black (piecewise constant).  
 25 Each color of the  $\hat{\nu}(\cdot)$  curve indicates a cluster and hence an homogeneous region with constant

---

**Algorithm 1:** Regional framework for peaks-over-threshold based on the probability weighted moments with a variable number of homogeneous regions with constant shape parameter

---

**input** :  $N_{reg}$  the number of homogeneous regions ;  
 $\mathbf{y}_i = \{y_{i1}, \dots, y_{in_i}\}$ ,  $n_i$  observed excesses and  $\mathbf{x}_i$  a vector of covariates at sites  $1 \leq i \leq M$  ;  
 $\mathbf{x}_{i^*}$   $1 \leq i^* \leq M^*$  for ungauged sites (optional)  
**output**:  $\{\hat{\xi}_1, \dots, \hat{\xi}_{N_{reg}}\}$ ,  $\hat{\sigma}(\mathbf{x}_i)$  and  $C_i$  for  $1 \leq i \leq M$  ;  
 $C_{i^*}$  and  $\hat{\sigma}(\mathbf{x}_{i^*})$  for  $1 \leq i^* \leq M^*$

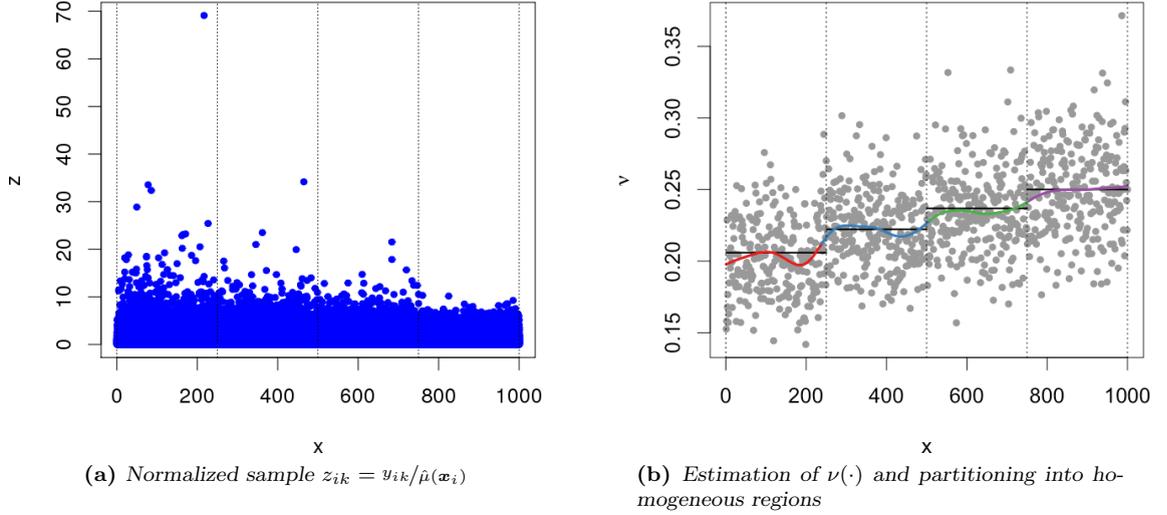
- 1 Estimate  $\mu(\cdot)$  by regressing  $\hat{\mu}_i = 1/n_i \sum_{k=1}^{n_i} y_{ik}$  over  $\mathbf{x}_i$  ;
- 2 Compute the normalized excesses  $z_{ik} = y_{ik}/\hat{\mu}(\mathbf{x}_i)$  for  $1 \leq k \leq n_i$  ;
- 3 **if**  $N_{reg} == 1$  **then** // single homogeneous region case
- 4 | Assign all the sites to a single region  $C_i = 1 \forall i$  and  $C_{i^*} = 1 \forall i^*$ ;
- 5 **else** // partitioning into  $N_{reg}$  regions
- 6 | Estimate  $\nu(\cdot)$  by regressing  $\hat{\nu}_i$ , the second probability weighted moment sample estimate of  $Z_i$ , over  $\mathbf{x}_i$  ;  
//  $\nu_i$  is estimated using U-statistics [Furrer and Naveau, 2007]
- 7 | Assign each site  $i$  to a region  $C_i$ ,  $1 \leq C_i \leq N_{reg}$  by clustering  $\hat{\nu}(\mathbf{x}_i)$  ;
- 8 | Assign each  $i^*$  to a region  $C_{i^*}$  with a classifier based on  $\mathbf{x}_{i^*}$  and  $\mathbf{x}_i$  ;
- 9 **end**
- 10 **for**  $j \leftarrow 1$  **to**  $N_{reg}$  **do**
- 11 | Estimate  $\nu_j$  from all  $z_{ik}$ ,  $1 \leq k \leq n_i$ , such that  $C_i = j$  ;
- 12 | Estimate  $\xi_j$ , the shape parameter of region  $j$  as  $\hat{\xi}_j = (1-4\hat{\nu}_j)/(1-2\hat{\nu}_j)$ , see Eq. (7) ;
- 13 **end**
- 14 Estimate  $\sigma(\mathbf{x}_i)$  thanks to  $\hat{\sigma}(\mathbf{x}_i) = \hat{\mu}(\mathbf{x}_i)(1 - \hat{\xi}_{C_i})$ , see Eq. (8) ;
- 15 Estimate  $\sigma(\mathbf{x}_{i^*})$  similarly ;

---

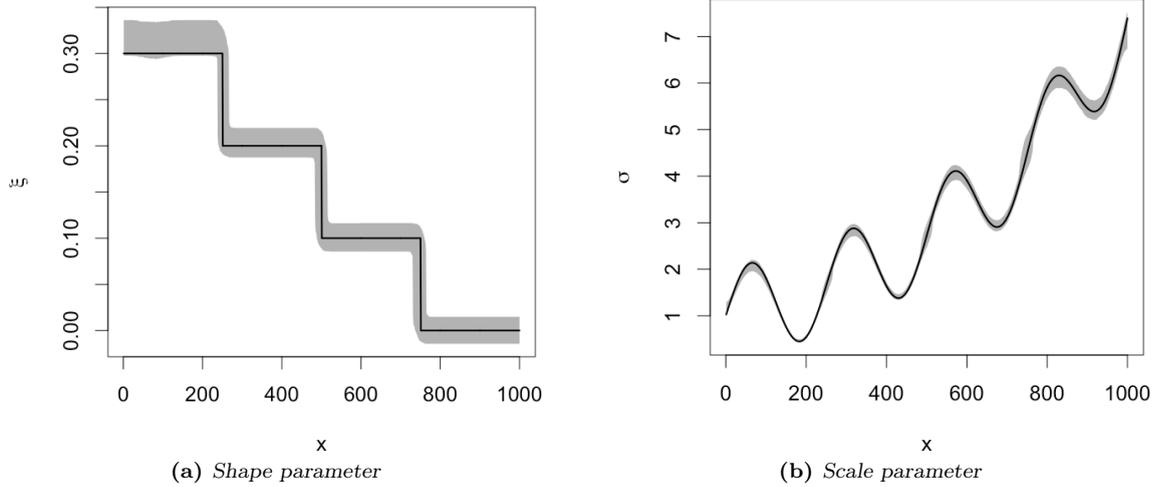
1 shape parameter.

2 In the last step of the regional framework described in Algorithm 1, the GP parameters are  
3 estimated at the gauged and ungauged sites (Fig. 5a and 5b, lines 10 to 15 of Algorithm 1). For  
4 each homogeneous region  $j \in \{1, \dots, N_{reg}\}$ , all the  $z_{ik}$ ,  $1 \leq k \leq n_i$ , belonging to that region,  
5 i.e. such that  $C_i = j$ , serve to estimate  $\nu_j$ , the regional second probability weighted moment,  
6 and  $\xi_j$ , the regional shape parameter with Eq. (7) (lines 11-12). Finally, the scale parameter  
7 is computed by combining the  $\hat{\mu}(\mathbf{x}_i)$  estimate (line 1) at the gauged sites  $1 \leq i \leq M$  with the  
8 regional shape parameter estimate  $\hat{\xi}_{C_i}$  of the associated homogeneous region in Eq. (8) (line 14).  
9 The same computation is carried out for ungauged sites if needed (line 15). Parametric bootstrap  
10 is employed to deduce 95% confidence intervals for the shape and scale estimators of the proposed  
11 regional framework (Fig. 5a and 5b respectively). A 1000 copies of the synthetic data set are  
12 generated and the estimation at all  $M = 1000$  sites is performed on each copy. In Fig. 5a and 5b,  
13 the 95 % confidence intervals are shown as gray bands and the parameters of the generative model  
14 are shown as black curves.

15 The main contribution of this work is the use, in the regional framework stemming from the  
16 Naveau et al. [2014] framework, of the statistic  $\hat{\nu}(\mathbf{x}_i)$  that is successful at identifying regions  
17 with constant shape parameter, as shown in Fig. 4b. As a consequence, the proposed framework  
18 yields a reliable estimation of the shape parameter in each region (see Fig. 5a), provided that  
19 the assumptions behind the framework are fulfilled. In addition, by relying on non-parametric  
20 regression instead of local estimation for both  $\mu$  and  $\nu$ , spatial information is introduced and  
21 the noise of the local estimates is considerably reduced (see Fig. 3b and 4b). Finally, the 95%  
22 confidence intervals shows the stability of the algorithm for the synthetic data example and good



**Figure 4:** Left panel : The synthetic dataset from Fig 3a is normalized with the smoothed  $\hat{\mu}(\cdot)$  estimate (red curve in Fig. 3b). Right panel : Local estimates  $\nu_i$  are obtained from the normalized sample (gray dots) and kernel regression is applied to obtain a smooth estimate  $\hat{\nu}(\cdot)$  (colored curves). Homogeneous regions are determined by applying K-Means on the smoothed estimates  $\hat{\nu}(x_i)$ . Each region, represented by a different colored  $\hat{\nu}(\cdot)$  curve, determines an area with constant shape parameter.



**Figure 5:** GP parameter estimates from the regional framework described in Algorithm 1 together with 95% confidence bands computed with parametric bootstrap. The black curve represents the generative model. The normalized sample (Fig 4a) is used to estimate the shape parameter within each homogeneous region. The scale parameter estimate is obtained by combining the shape parameter estimate with the smoothed  $\hat{\mu}(\cdot)$  estimate from Fig. 3b.

1 agreement between the estimated and the generative models. Only in the case of the sub-interval  
2 corresponding to the higher shape parameter ( $\xi_1 = 0.3$ ), the shape parameter of the generative  
3 model is at the lower end of the asymmetric confidence bands.

## 4 4.5 Regional analysis of the daily precipitation data

5 We apply the regional framework described above to the 332 French Mediterranean precipitation  
6 stations from Section 2 and compare the results with those from the basic framework in Section 3.4.  
7 The same covariates  $\mathbf{x}$  (the x and y coordinate in extended Lambert II projections of latitude and  
8 longitude) and the same regular grid of about 500 m is used for the interpolation. The grid boxes  
9 are provided as ungauged sites in Algorithm 1. The threshold that defines the excesses which are  
10 approximately GP distributed, along with the corresponding average number of excesses per year,  
11 are the same as those presented in the basic framework, see Fig. 2a-2b.

### 12 4.5.1 Partitioning into homogeneous regions

13 From the regional framework of Algorithm 1, a partition into homogeneous regions is obtained as  
14 each grid box is assigned to a region corresponding to a regional shape parameter. We present the  
15 partitioning for increasing numbers of regions (three to six), see Fig. 6a-6d. This partitioning can  
16 be compared to the estimated shape parameter from the basic framework, see Fig. 2c.

17 The homogeneous regions are remarkably continuous in space although geographical information  
18 is used only indirectly to define the regions through the regression of the local  $\nu_i$  estimates (line 6  
19 of Algorithm 1). Even in the 6-region partition in Fig. 6d, i.e with the larger number of regions,  
20 although the regional parameter estimates can be very similar for some regions, the regions remain  
21 approximately spatially coherent. In addition, for all the partitions, the regions are roughly aligned  
22 along the same direction, with a slight counterclockwise angle from 12 o'clock (a central vertical  
23 line).

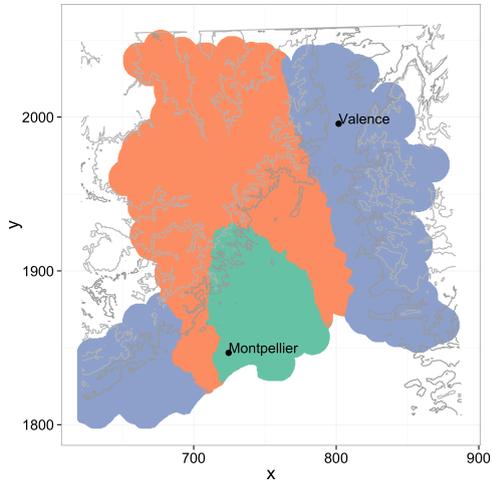
24 In most cases, the regions have clear borders. However, in a number of cases, especially when  
25 the number of regions increases, the borders are somewhat blurred. This happens when a region has  
26 only a few scattered representative stations in a given geographical area. In such a configuration,  
27 the k-nearest neighbor rule yields unstable region assignment.

28 As the number of regions increases, the partitioning provides more detailed patterns that are  
29 nested into the smoother patterns of partitions with less regions. As expected, with a larger  
30 number of regions, the partitioning reproduces more closely the patterns of the local estimates in  
31 Fig. 2c. On the contrary, with less regions, the shape parameter estimates seem to smooth the  
32 local estimates over the larger regions.

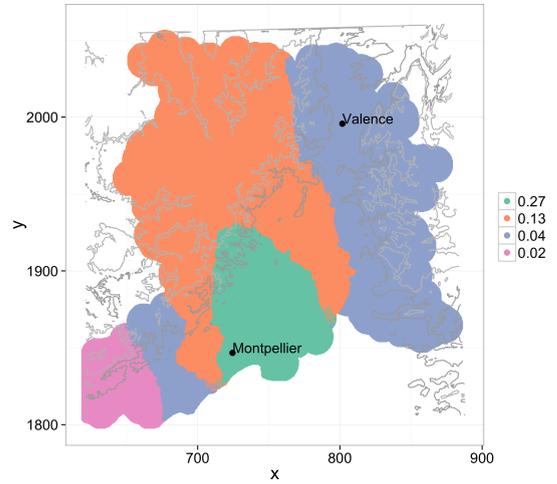
33 For all partitions, the high risk region, corresponding to the highest shape parameter value with  
34  $\hat{\xi}_j \approx 0.27 - 0.30$ , is located in the South and represented with the blue-green color. The region  
35 starts at the coast, goes up to the foothills of the Cevennes mountain range and is consistent with  
36 the local estimates in Fig. 2c and expert knowledge [Delrieu et al., 2005, Braud et al., 2014].

### 37 4.5.2 Scale parameter estimates

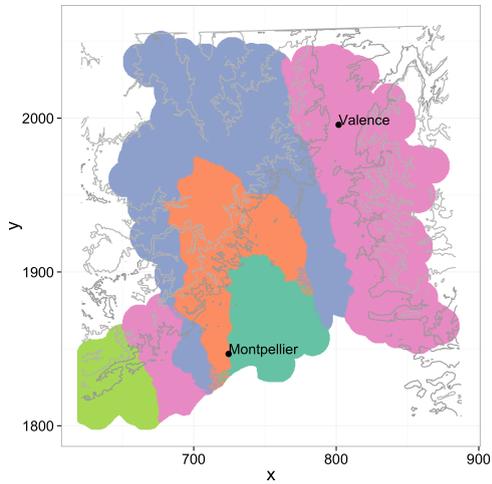
38 Fig. 7a and 7b present the maps of the differences between the estimated scale parameter from  
39 the basic framework, see Fig. 2d, and the estimated scale parameter from the regional framework  
40 described in Algorithm 1. The latter estimate corresponds to the partitioning into three and six



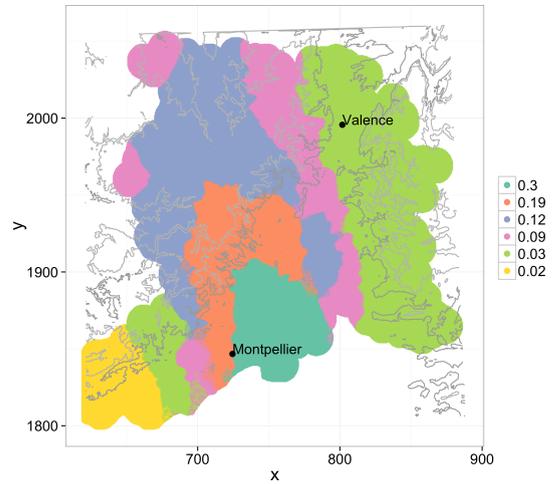
(a) Three homogeneous regions



(b) Four homogeneous regions



(c) Five homogeneous regions



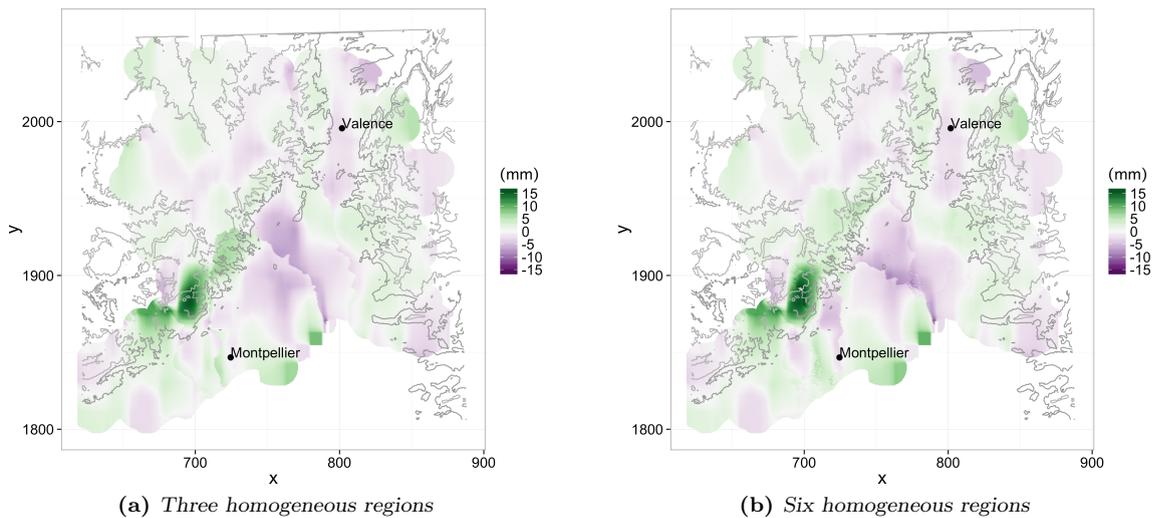
(d) Six homogeneous regions

**Figure 6:** Partitioning into homogeneous regions associated to estimated regional shape parameter values  $\{\hat{\xi}_1, \dots, \hat{\xi}_{N_{reg}}\}$  in the legends. The partitions, with  $N_{reg} = 3, 4, 5, 6$ , are defined based on the regional framework detailed in Algorithm 1.

1 regions shown in Fig. 6a and 6d. The results are similar for the partitioning into four and five  
 2 regions (not shown).

3 The estimation of the scale parameter with the proposed regional framework is little sensitive to  
 4 the selected number of regions and is very close to the estimate from the basic framework. Indeed,  
 5 for all the partitions considered (three to six regions), the differences in scale parameter estimates  
 6 for about 95% of the grid boxes is at most 5 mm in magnitude.

7 In most cases, the scale parameter estimates of the proposed regional framework do not show  
 8 major discontinuities at the region borders. The highest risk region discussed above is an exception :  
 9 the North-East border is quite visible in the maps of differences in scale parameter estimates Fig. 7a  
 10 and 7b (this can be compared to the corresponding partitions in Fig. 6a and 6d).



**Figure 7:** Differences in scale parameter estimates : basic framework estimates minus regional framework estimates. In the latter,  $\sigma(\mathbf{x}) = \mu(\mathbf{x}) (1 - \xi_j)$  where  $\xi_j$  is the regional shape parameter for the  $j^{\text{th}}$  homogeneous region and  $\mu(\mathbf{x})$  is the conditional expectation of the excesses for the covariates  $\mathbf{x}$ . The homogeneous regions corresponds to the partitioning in Fig. 6a and 6d.

### 11 4.5.3 Confidence intervals

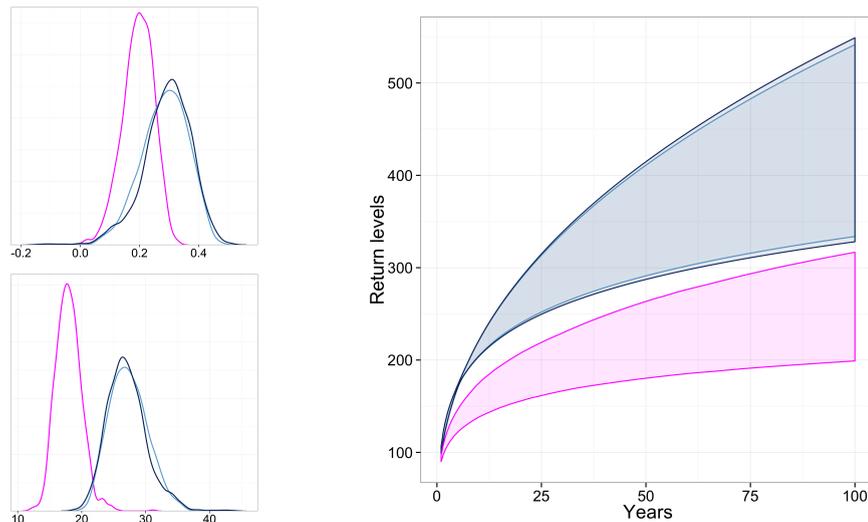
12 We selected two distinctive grid boxes as target sites to illustrate the sampling distributions of the  
 13 estimators in both the basic and the regional framework. The first grid box is located 20 km East  
 14 of the city of Montpellier in the high risk region with shape parameter taking value from 0.27 to  
 15 0.30, see Fig. 6. In contrast, the second grid box lies 20 km North of the city of Valence in a low  
 16 risk region with shape parameter value around zero.

17 The sampling distributions of the GP parameter estimates can be obtained with spatial block  
 18 bootstrap to preserve temporal and spatial dependence of the excesses. More precisely, blocks of  
 19 three days are randomly sampled from the original observations for all the stations simultaneously.  
 20 The size of the block was determined from the maximum number of consecutive excesses in the  
 21 precipitation data.

22 The region labels  $C_i$  for each gauged site are determined once and for all on the original  
 23 observations. This means that the configuration of the regions is kept fixed for each bootstrap  
 24 sample. In addition, the thresholds are estimated only once on the original observations. For

1 each bootstrap sample, for each site, the excesses are extracted and the scaling factor  $\mu(\cdot)$ , the  
 2 conditional estimation of the excesses, is estimated. The normalized observations  $z_{ik}$ ,  $1 \leq k \leq n_i$ ,  
 3 are then computed and serve to estimate the regional shape parameters using the original region  
 4 labels  $C_i$ . Last, the scale parameter estimates are obtained as usual. Return levels are computed  
 5 thanks to Eq. (3) for return periods ranging from 1 to 100 years. This is repeated a 1000 times.

6 Confidence bands at 95 % for the return levels of the basic framework (magenta) and the  
 7 regional framework with three (light blue) and six regions (dark blue) are presented in Fig. 8 and  
 8 9 for the high and low risk grid boxes respectively. In addition, to the left of each figure, the  
 9 smoothed empirical distributions (resulting from the application of the function `density` of R to  
 10 the sample estimates) of the shape (top) and scale parameter (bottom) estimates are shown.



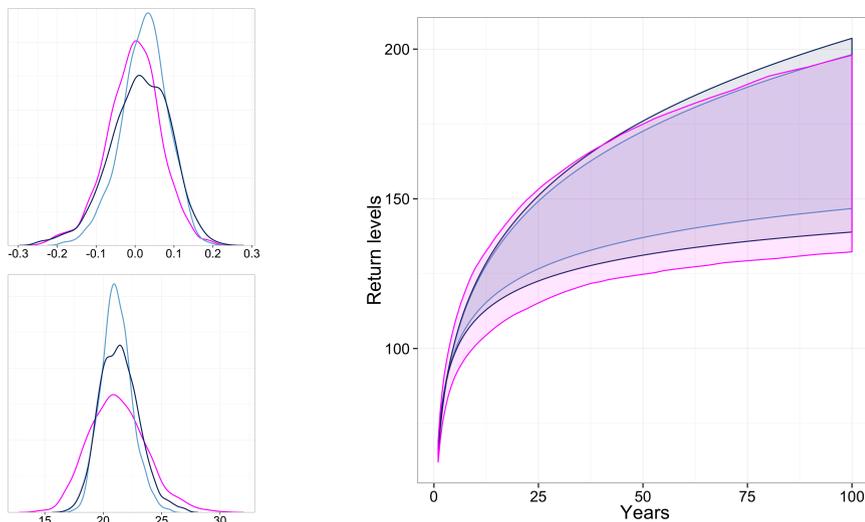
**Figure 8:** *Uncertainty estimation at the grid box 20 km East of the city of Montpellier (high risk region). Left panel : bootstrap distribution of the shape (top) and scale (bottom) parameters for the basic framework (magenta) and the regional framework with three regions (light blue) and six regions (dark blue). Right panel : 95 % confidence bands for the return level curves with the same color code.*

11 For the high risk grid box, Fig. 8, the proposed regional framework with either three or six  
 12 regions tend to yield higher shape and scale parameter estimates which result in significantly  
 13 higher return levels. In contrast, for the low risk grid box, Fig. 9, the sampling distribution of the  
 14 shape and scale parameter estimates is similar in both frameworks which explains the overlapping  
 15 confidence bands of the estimated return levels.

## 16 5 Discussion and Conclusion

17 In an area such as the French Mediterranean area where heavy precipitation events can trigger flash  
 18 floods with devastating consequences, it is essential to obtain reliable estimates of the distribution  
 19 of extreme precipitation at both gauged and ungauged locations. To this end, regional analysis can  
 20 be combined with the block maxima or the peaks-over-threshold approach in a robust framework.

21 In this paper, we built on the approach proposed in Naveau et al. [2014] to address some of the  
 22 shortcomings of regional analysis. First, we cast the Naveau et al. [2014] approach into a regional  
 23 framework for peaks-over-threshold with a single homogeneous region and simplified the expressions  
 24 to compute the GP parameter estimates. The scaling factor, in this approach, is the estimated



**Figure 9:** *Uncertainty estimation at the grid box 20 km North of the city of Valence (low risk region). Left panel : bootstrap distribution of the shape (top) and scale (bottom) parameters for the basic framework (magenta) and the regional framework with three regions (light blue) and six regions (dark blue). Right panel : 95 % confidence bands for the return level curves with the same color code.*

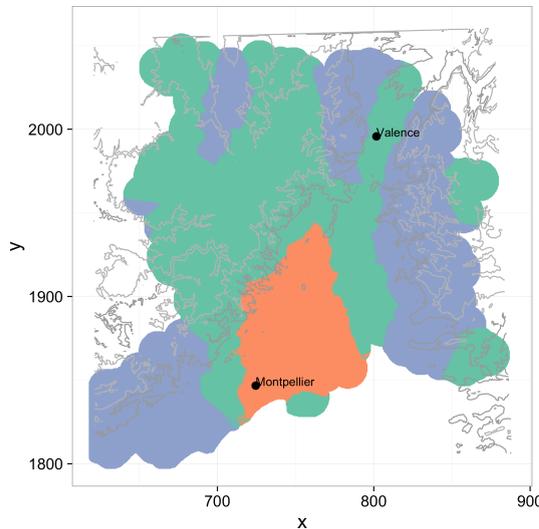
1 conditional expected value of the excesses, noted  $\hat{\mu}(\mathbf{x}_i)$ . Although the scaling factor does not have  
2 a clear physical meaning, it does have a clear statistical advantage. Indeed, the scale invariance  
3 property of the normalized variable is automatically fulfilled, without further assumptions. In  
4 addition, the scale parameter of the normalized variable solely depends on the shape parameter.  
5 In other words, in the regional framework proposed in this paper, the regional distribution has a  
6 single parameter. Madsen and Rosbjerg [1997] also considered a regional framework for peaks-over-  
7 threshold with the expected value of the excesses as the scaling factor. They resorted to L-moment  
8 ratios [Hosking and Wallis, 2005] to estimate the shape parameter based on the normalized excesses.

9 The main contribution of this work is, in the regional framework derived from Naveau et al.  
10 [2014], the characterization of homogeneous regions with the second probability weighted moment  
11 estimate of the normalized variable, noted  $\hat{\nu}(\mathbf{x}_i)$  for site  $i$ . In the simulation study on synthetic  
12 data, homogeneous regions, i.e. with constant shape parameter, were successfully identified. The  
13 correct identification of homogeneous regions lead to GP parameter estimates with low variance,  
14 as shown by the narrow 95% confidence bands computed with parametric bootstrap. Finally,  
15 the use of regressed (as opposed to local) statistics to estimate both  $\mu$  and  $\nu$  introduced spatial  
16 information and reduced the noise of the estimators. The application on daily precipitation data  
17 from the French Mediterranean area illustrated the regional framework on a complex region with  
18 several homogeneous sub-regions and shape parameter estimates ranging from approximately 0 to  
19 0.3. For the real data application, uncertainty was estimated with an easy-to-implement spatial  
20 block bootstrap. Another approach to properly estimate the uncertainty when observations are  
21 spatially dependent has been proposed in Van de Vyver [2012].

22 As recommended in Hosking and Wallis [2005], most authors (e.g. Kysely et al. [2011]) define  
23 regions based on physiographic variables rather than on site statistics. In particular, Hosking and  
24 Wallis [2005] argue against the use of the L-CV statistic (L-moment analog of the coefficient of  
25 variation). They claim that (1) not much information on homogeneous regions can be gained from  
26 the L-CV statistic as it won't take very different values from site to site, (2) outliers might bias the

1 identification of the regions and (3) the homogeneity of the regions must be tested with a statistic  
 2 that may contain redundant information with the statistic used to create the regions. We discuss  
 3 these three points in turn.

4 (1) The L-CV statistic is used to partition into three homogeneous regions the sites of the French  
 5 Mediterranean precipitation data in Fig. 10 in order to compare with the partitioning obtains with  
 6 the  $\nu$  statistic. The L-CV statistic is estimated at each gauged site with the R package `lmomRFA`  
 7 [Hosking, 2015]. To ensure a fair comparison, the same steps from the regional framework described  
 8 in Algorithm 1 for the  $\nu$  statistic are applied : the local L-CV statistics are smoothed with kernel  
 9 regression, K-Means clusters the sites into three regions and the k-nearest neighbor rule with  $k = 5$   
 10 assigns each grid point to a region. From Fig. 10, we can see that the L-CV statistic allows to  
 11 identify the high risk region in the South, including the city of Montpellier. However, the other  
 12 two regions are not well separated. Similar partitioning are obtained with more regions or with  
 13 other L-Moment ratios (not shown). This corroborates the claim above that not much information  
 14 on homogeneous regions can be gained from the L-CV or other L-Moment ratios.



**Figure 10:** Partitioning into three homogeneous regions based on the L-CV statistic estimated from kernel regression. K-Means performed the clustering and the k-nearest neighbor rule with  $k = 5$  assigned each grid point to a region, following the steps of the regional framework described in Algorithm 1 but with a different statistic.

15 On the other hand, in view of the results on both the synthetic and real precipitation data, we  
 16 claim that the  $\nu$  statistic is efficient to identify homogeneous regions. The obtained regions are  
 17 mostly continuous in space and are meaningful according to expert knowledge. The choice of the  $\nu$   
 18 statistic, in the proposed regional framework, is quite natural as it is a summary of the distribution  
 19 of the normalized variable  $Z$  that possesses a unique parameter which is the shape parameter. This  
 20 is precisely what characterizes homogeneous regions, their shape parameter value. In addition,  $\nu$   
 21 is the lower not trivial probability weighted moment of  $Z$  (the only probability weighted moment  
 22 below  $\nu$  is the expectation of  $Z$  which is one by construction). As each homogeneous region is  
 23 associated with a shape parameter value derived from the  $\nu$  statistic, it is associated with an  
 24 interpretable level of risk of extreme precipitation.

25 (2) The potential effect of outliers is greatly reduced in the regional framework from Algorithm 1  
 26 as the local estimate of  $\nu$  are regressed. As mentioned earlier, this introduces spatial information  
 27 and smooths out the noise. Large outliers might still affect the  $\nu$  estimates but probably to a lesser

1 degree.

2 (3) In the proposed regional framework, there is no automated method to select an appropriate  
3 number of homogeneous regions. Homogeneity tests could be employed, for instance, to determine  
4 if a region needs to be further divided to achieve homogeneity. However, as concluded in Viglione  
5 et al. [2007], these tests lack power. For this reason, we prefer to resort to expert knowledge to  
6 choose the number of homogeneous regions. This is the only subjective decision in the proposed  
7 framework.

8 Quite often, because its sampling variability is larger than its spatial variability, the shape pa-  
9 rameter is kept constant across the region of interest. This appears to be appropriate when studying  
10 extremes in a region with little orography such as Belgium [Van de Vyver, 2012]. Nevertheless,  
11 in a region with a sharp orography inducing a high spatial variability such as the French Mediter-  
12 ranean area studied in this paper, we believe that the assumption of a constant shape parameter  
13 is not realistic. The partitioning into homogeneous regions strikes a middle ground between the  
14 assumption of a constant and spatially varying shape parameter. Indeed, each region possesses  
15 several sites whose observations contributes to the estimation of the regional shape parameter. In  
16 order to guide the selection of the number of homogeneous regions, the sampling distribution could  
17 be used to determine how many distinct regional shape parameters can be identified. Preliminary  
18 analysis of the French Mediterranean precipitation data showed that at least two regions could be  
19 identified. However, this is beyond the scope of this paper and should be further analysed.

20 Further study is also required in order to determine in which cases the regional framework with  
21 the GP distribution proposed in this paper performs better than other interpolation methods such  
22 as the direct interpolation of the GP parameters described in the basic framework. For the block  
23 maxima approach with the GEV distribution, Carreau et al. [2013] have shown that the regional  
24 framework outperforms a direct interpolation of the parameters for sparsely gauged network. In  
25 addition, the comparison of return levels showed that there are significant differences between the  
26 direct interpolation of the GP parameters from the basic framework and the interpolation from  
27 the regional framework for the grid box in the high risk region, see Fig. 8, but not for the grid box  
28 in the low risk region, see Fig. 9. Although this should be validated, it might indicate that there  
29 is a gain in resorting to the regional framework in particular when the interpolation conditions are  
30 more difficult such as when the network is sparsely gauged or when the shape parameter can take  
31 high values.

32 Other perspectives for this work are as follows. Instead of employing the characterization of  
33 homogeneous regions to create contiguous regions, the  $\nu$  statistic could be employed in a region-  
34 of-influence type-of approach [Burn, 1990]. Indeed, a neighborhood around a site (gauged or  
35 ungauged) could be defined in terms of similarity in the  $\nu$  statistic. The size of the neighborhood  
36 could be determined based on expert knowledge, as for the number of regions in the contigu-  
37 ous case. Another perspective would be to apply the proposed regional framework to a sparsely  
38 gauged network. In such a case, covariates would have to be chosen with care. Most likely,  $x$   
39 and  $y$  coordinates would not be informative enough and covariates related to the orography could  
40 be of interest [Benichou and Le Breton, 1987]. In addition, non-parametric methods (k-nearest  
41 neighbor rule and kernel regression) worked well in the dense precipitation network to which the  
42 regional framework was applied. However, in sparser network, it might be more appropriate to  
43 seek parcimonious parametric models.

## 1 References

- 2 K. M. Abadir and S. Lawford. Optimal asymmetric kernels. *Economics Letters*, 83(1):61–68, 2004.
- 3 A. A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability*, 2(5):792–804,  
4 1974.
- 5 P. Benichou and O. Le Breton. Prise en compte de la topographie pour la cartographie des champs  
6 pluviométriques statistiques. une application de la methode Aurelhy: la cartographie nationale  
7 de champs de normales pluviométriques. *Météorologie*, 1987.
- 8 J. Blanchet and M. Lehning. Mapping snow depth return levels: smooth spatial modeling versus  
9 station interpolation. *Hydrology and Earth System Sciences*, 14(12):2527–2544, 2010.
- 10 M. Borga, E.N. Anagnostou, G. Blöschl, and J.-D. Creutin. Flash flood forecasting, warning  
11 and risk management: the HYDRATE project. *Environmental Science & Policy*, 14(7):834 –  
12 844, 2011. ISSN 1462-9011. doi: <http://dx.doi.org/10.1016/j.envsci.2011.05.017>. URL <http://www.sciencedirect.com/science/article/pii/S1462901111000943>. Adapting to Climate  
13 Change: Reducing Water-related Risks in Europe.
- 14
- 15 I. Braud, P.-A. Ayrat, C. Bouvier, F. Branger, G. Delrieu, J. Le Coz, G. Nord, J.-P. Van-  
16 dervaere, S. Anquetin, M. Adamovic, J. Andrieu, C. Batiot, B. Boudevillain, P. Brunet,  
17 J. Carreau, A. Confoland, J.-F. Didon-Lescot, J.-M. Domergue, J. Douvinet, G. Dramais,  
18 R. Freydier, S. Gérard, J. Huza, E. Leblois, O. Le Bourgeois, R. Le Boursicaud, P. Marc-  
19 hand, P. Martin, L. Nottale, N. Patris, B. Renard, J.-L. Seidel, J.-D. Taupin, O. Vannier,  
20 B. Vincendon, and A. Wijbrans. Multi-scale hydrometeorological observation and modelling for  
21 flash flood understanding. *Hydrology and Earth System Sciences*, 18(9):3733–3761, 2014. doi:  
22 10.5194/hess-18-3733-2014. URL <http://www.hydrol-earth-syst-sci.net/18/3733/2014/>.
- 23 D.H. Burn. Evaluation of regional flood frequency analysis with a region of influence approach.  
24 *Water Resources Research*, 26(10):2257–2265, 1990.
- 25 D.H. Burn and N. K. Goel. The formation of groups for regional flood frequency analysis.  
26 *Hydrological Sciences Journal*, 45(1):97–112, 2000. doi: 10.1080/02626660009492308. URL  
27 <http://dx.doi.org/10.1080/02626660009492308>.
- 28 J. Carreau, L. Neppel, P. Arnaud, and P. Cantet. Extreme rainfall analysis at ungauged sites in the  
29 South of France: comparison of three approaches. *Journal de la Société Française de Statistique*,  
30 154(2):119–138, 2013.
- 31 D. Ceresetti, E. Ursu, J. Carreau, S. Anquetin, J.-D. Creutin, L. Gardes, S. Girard, and G. Molinie.  
32 Evaluation of classical spatial-analysis schemes of extreme rainfall. *Natural hazards and earth  
33 system sciences*, 12:3229–3240, 2012.
- 34 S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.  
35 Springer, 2001.
- 36 D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return  
37 levels. *Journal of the American Statistical Association*, 102(479):824–840, 2007.

- 1 G. Delrieu, J. Nicol, E. Yates, P.-E. Kirstetter, J.-D. Creutin, S. Anquetin, C. Obled, G.-M.  
2 Saulnier, V. Ducrocq, E. Gaume, O. Payrastre, H. Andrieu, P.-A. Ayral, C. Bouvier, L. Neppel,  
3 M. Livet, M. Lang, J. P. du Châtelet, A. Walpersdorf, and W. Wobrock. The catastrophic flash-  
4 flood event of 8-9 september 2002 in the Gard region, France: A first case study for the Cévennes-  
5 Vivarais Mediterranean Hydrometeorological Observatory. *Journal of Hydrometeorology*, 6(1):  
6 34–52, 2005.
- 7 J. Diebolt, A. Guillou, and I. Rached. Approximation of the distribution of excesses through a  
8 generalized probability-weighted moments method. *Journal of Statistical Planning and Inference*,  
9 137(3):841–857, 2007.
- 10 M. Durocher, F. Chebana, and T. B. M. J. Ouarda. Delineation of homogenous regions us-  
11 ing hydrological variables predicted by projection pursuit regression. *Hydrology and Earth*  
12 *System Sciences Discussions*, 2016:1–23, 2016. doi: 10.5194/hess-2016-123. URL [http:](http://www.hydrol-earth-syst-sci-discuss.net/hess-2016-123/)  
13 [//www.hydrol-earth-syst-sci-discuss.net/hess-2016-123/](http://www.hydrol-earth-syst-sci-discuss.net/hess-2016-123/).
- 14 V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of*  
15 *Probability & Its Applications*, 14(1):153–158, 1969.
- 16 A. Ferreira and L. de Haan. On the block maxima method in extreme value theory: PWM  
17 estimators. *The Annals of Statistics*, 43(1):276–298, 2015.
- 18 R. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest and  
19 smallest member of a sample. In *Cambridge Philosophical Society*, volume 24, pages 180–190,  
20 1928.
- 21 R. Furrer and P. Naveau. Probability weighted moments properties for small samples. *Statistics*  
22 *& probability letters*, 77(2):190–195, 2007.
- 23 B. Gnedenko. Sur la distribution limite du terme maximum d’une serie aleatoire. *Annals of*  
24 *mathematics*, pages 423–453, 1943.
- 25 V. K. Gupta, O. J. Mesa, and D. R. Dawdy. Multiscaling theory of flood peaks: Regional quantile  
26 analysis. *Water Resources Research*, 30(12):3405–3421, 1994. ISSN 1944-7973. doi: 10.1029/  
27 94WR01791. URL <http://dx.doi.org/10.1029/94WR01791>.
- 28 T. Hayfield and J.S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical*  
29 *Software*, 27(5), 2008. URL <http://www.jstatsoft.org/v27/i05/>.
- 30 J. R. M. Hosking and J. R. Wallis. *Regional frequency analysis: an approach based on L-moments*.  
31 Cambridge University Press, 2005.
- 32 J.R.M. Hosking. *Regional frequency analysis using L-moments*, 2015. URL [http://CRAN.](http://CRAN.R-project.org/package=lmomRFA)  
33 [R-project.org/package=lmomRFA](http://CRAN.R-project.org/package=lmomRFA). R package, version 3.0-1.
- 34 J. Kysely, J. Picek, and R. Huth. Formation of homogeneous regions for regional frequency analysis  
35 of extreme precipitation events in the czech republic. *Studia Geophysica et Geodaetica*, 51(2):  
36 327–344, 2007.
- 37 J. Kysely, L. Gaál, and J. Picek. Comparison of regional and at-site approaches to modelling  
38 probabilities of heavy precipitation. *International Journal of Climatology*, 31(10):1457–1472,  
39 2011.

- 1 Q. Li and J. Racine. Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14  
2 (2):485–512, 2004.
- 3 H. Madsen and D. Rosbjerg. The partial duration series method in regional index-flood modeling.  
4 *Water Resources Research*, 33(4):737–746, 1997.
- 5 E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142,  
6 1964.
- 7 P. Naveau, A. Toreti, I. Smith, and E. Xoplaki. A fast nonparametric spatio-temporal regression  
8 scheme for generalized Pareto distributed heavy precipitation. *Water Resources Research*, 50(5):  
9 4011–4017, 2014.
- 10 J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131,  
11 1975.
- 12 B. Renard. A bayesian hierarchical approach to regional frequency analysis. *Water Resources*  
13 *Research*, 47(11), 2011.
- 14 B. D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 1996.
- 15 M. Roth, T.A. Buishand, G. Jongbloed, A.M.G. Klein Tank, and J.H. van Zanten. A regional  
16 peaks-over-threshold model in a nonstationary climate. *Water Resources Research*, 48(11), 2012.
- 17 H. Van de Vyver. Spatial regression models for extreme precipitation in belgium. *Water Resources*  
18 *Research*, 48(9), 2012.
- 19 A. Viglione, F. Laio, and P. Claps. A comparison of homogeneity tests for regional frequency  
20 analysis. *Water Resources Research*, 43(3), 2007.
- 21 G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*,  
22 pages 359–372, 1964.