Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation

J. Carreau, P. Naveau, L. Neppel

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2017WR020758

Key Points:
- Regional peaks-over-threshold formalized as a conditional mixture model
- Inference strategy based on probability weighted moments and nonparametric estimators
- Selection of the number of subregions with a cross-validation procedure

Supporting Information:
- Supporting Information S1

Correspondence to:
J. Carreau,
julie.carreau@ird.fr

Citation:
Carreau, J., P. Naveau, and L. Neppel (2017), Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation, *Water Resour. Res.*, *53*, 4407–4426, doi:10.1002/2017WR020758.

Received 17 MAR 2017
Accepted 11 MAY 2017
Accepted article online 17 MAY 2017
Published online 31 MAY 2017

Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation

J. Carreau[1], P. Naveau[2], and L. Neppel[1]

[1]HSM, CNRS/IRD/UM, Université de Montpellier, Montpellier, France, [2]LSCE, IPSL-CNRS, Orme des Merisiers, Gif-sur-Yvette, France

Abstract The French Mediterranean is subject to intense precipitation events occurring mostly in autumn. These can potentially cause flash floods, the main natural danger in the area. The distribution of these events follows specific spatial patterns, i.e., some sites are more likely to be affected than others. The peaks-over-threshold approach consists in modeling extremes, such as heavy precipitation, by the generalized Pareto (GP) distribution. The shape parameter of the GP controls the probability of extreme events and can be related to the hazard level of a given site. When interpolating across a region, the shape parameter should reproduce the observed spatial patterns of the probability of heavy precipitation. However, the shape parameter estimators have high uncertainty which might hide the underlying spatial variability. As a compromise, we choose to let the shape parameter vary in a moderate fashion. More precisely, we assume that the region of interest can be partitioned into subregions with constant hazard level. We formalize the model as a conditional mixture of GP distributions. We develop a two-step inference strategy based on probability weighted moments and put forward a cross-validation procedure to select the number of subregions. A synthetic data study reveals that the inference strategy is consistent and not very sensitive to the selected number of subregions. An application on daily precipitation data from the French Mediterranean shows that the conditional mixture of GPs outperforms two interpolation approaches (with constant or smoothly varying shape parameter).

1. Introduction

In the French Mediterranean, heavy precipitation events occurring mainly in autumn, often called *Cevenol events* tend to gather in very specific areas. Factors explaining the spatial distribution of these events are the presence of mountains and the trajectories usually taken by contrasting air masses, humid, and warm coming from the Mediterranean sea and cold from the North. Heavy precipitation might trigger flash floods, the main natural danger in the Mediterranean area, that can potentially cause fatalities and important material damage [*Delrieu et al*., 2005; *Borga et al*., 2011; *Braud et al*., 2014].

Extreme value theory [*Coles*, 2001] provides a sound asymptotic framework commonly used to model the distribution of extremes such as heavy precipitation. The Extremal-types theorem [*Fisher and Tippett*, 1928; *Gnedenko*, 1943] states that the behavior of the upper tail of the distribution, which governs the probability of extreme events, can be of three distinct types. Two classical strategies for statistical inference are as follows. In the block maxima approach, the generalized Extreme Value (GEV) distribution is fitted to maxima over sufficiently large blocks of observations, often taken as years. On the other hand, the peaks-over-threshold approach consists in approximating the distribution of the excesses over a large enough threshold by the generalized Pareto (GP) distribution [*Balkema and de Haan*, 1974; *Pickands*, 1975]. Both the GEV and the GP distribution have a shape parameter which determines the type of extremal behavior: Fréchet (heavy tail), Gumbel (exponential or light tail), or Weibull (finite bounded tail) type depending on the range of values of the shape parameter (positive, zero or negative, respectively).

Although risk assessment is conventionally performed by hydrologists based on return levels such as the 100 year return level, the probability of extreme events and therefore hazard levels are strongly influenced by the shape parameter. Indeed, return levels of period T years (the amount which is expected to be exceeded on average once in T years) can be expressed as quantiles of the GEV or the GP distribution. In both cases, as T gets large, the shape parameter becomes the determining factor of the value of return

levels. Therefore, to assess hazard levels, considering return levels with a long period is strongly linked to the value of the shape parameter.

Several regression approaches have been developed to interpolate the distribution of extremes in a region. As the estimation of the shape parameter is known to be difficult, it is often assumed to be constant in space, thereby assuming a constant hazard level in the region [*Sang and Gelfand*, 2009; *Renard*, 2011; *Naveau et al*., 2014]. However, the shape parameter has also been modeled as functions of covariates with various degrees of freedom [*Blanchet and Lehning*, 2010; *Carreau and Girard*, 2011]. In *Cooley et al*. [2007], hierarchical Bayesian models of increasing complexity for the shape parameter are compared: a single value for the entire region (constant shape parameter in space), two values for two predefined regions (the selected model) and as functions of covariates with a Gaussian process.

Regional frequency analysis can also be applied as an interpolation approach of the distribution of extremes [*Hosking and Wallis*, 2005; *Carreau et al*., 2013]. It relies on the concept of homogeneous regions that can be defined either contiguously in space or as neighborhoods around target sites [*Burn*, 1990]. Extreme observations (annual maxima or excesses above a high threshold) from all the sites in a given homogeneous region are scaled by a site-specific factor and pooled together to estimate the so-called regional distribution. Both with the block maxima or the peaks-over-threshold strategies, this implies that homogeneous regions have constant shape parameter and thus the corresponding hazard levels can be either piecewise constant (in the case of contiguous regions) or smoothly varying (with the neighborhood approach also called *region of influence*).

In this work, we propose a regional peaks-over-threshold model to interpolate the distribution of extremes, defined as excesses above a high threshold, for regions in which the hazard level is assumed to vary moderately. More precisely, we assume that the hazard level is piecewise constant and that the region can be partitioned into subregions with constant shape parameter. The scale parameter is assumed to vary smoothly spatially as a function of covariates. The regional peaks-over-threshold model can be formalized as a conditional mixture of GP distributions, see section 2.2. We develop in section 2.2.1 a two-step inference and interpolation procedure to estimate the partition and the GP parameters, analogously to the Expectation-Maximization (EM) algorithm [*Dempster et al*., 1977]. The inference is fast since it relies on probability weighted moment estimators [*Diebolt et al*., 2007] and could be used to initialize EM or any other inference strategy involving optimization. The estimation of the GP parameters in a given subregion is a reformulation, with simplified expressions (requiring two instead of three probability weighted moments), of the inference strategy proposed in *Naveau et al*. [2014] (see Appendix A for details).

The number of subregions, i.e., the size of the partition or equivalently the number of components in the mixture, is selected with a cross-validation procedure, see section 2.2.2. This is a standard way to assess out-of-sample performance by creating test sets, i.e., data not used for estimation, which makes an efficient use of the available data [see *Bishop*, 2011, section 1.3]. By comparing several partition sizes with out-of-sample performance, the cross-validation procedure can find a balance between sufficient adaptability of the model obtained with enough subregions and too much variability as a result of too many shape parameters to infer. This is a typical bias-variance trade-off which is often addressed in regional frequency analysis with statistical tests to assess homogeneity [*Viglione et al*., 2007].

The proposed regional peaks-over-threshold model can be cast into regional frequency analysis combined with peaks-over-threshold [*Madsen and Rosbjerg*, 1997a; *Roth et al*., 2012; *Evin et al*., 2016]. However, none of these studies has considered the case of a partition into subregions with constant shape parameter (i.e., contiguous homogeneous regions). In addition, the procedure developed in this work to identify the partition takes an entirely different angle. In particular, the partition is formed based on a probability weighted moment that depends only on the shape parameter rather than from physiographic variables such as geographical and climatological characteristics [*Hosking and Wallis*, 2005]. Besides, although the inference strategy is completely different, the proposed conditional mixture of GPs can also be thought of as an extension of the two subregion model proposed in *Cooley et al*. [2007] in which the number of subregions is not fixed a priori.

The performance of the proposed regional peaks-over-threshold model is evaluated on 18 different synthetic data sets in section 3 in terms of its ability to select the appropriate number of hazard subregions and to estimate the GP parameters when the generative model is known. In section 4, the proposed model

is compared on daily precipitation from the French Mediterranean to two other interpolation approaches with different assumptions on the variability of the shape parameter (either constant on the entire region or smoothly varying). To this end, the performance of each interpolation approach is evaluated via cross validation. In addition, thanks to spatial block bootstrap, 95% confidence bands of return level curves are obtained for each approach at two stations that were kept aside for validation.

2. Regional Peaks-Over-Threshold Approach

2.1. Peaks-Over-Threshold Approach

The peaks-over-threshold approach is based on a theorem which states that, under mild assumptions, the generalized Pareto (GP) distribution can be used as an approximation to the upper tail of the distribution of most random variables [*Pickands*, 1975]. In other words, given a high enough threshold u suitably chosen, the GP distribution approximates the distribution of the excesses over u. Let $Y \sim G(\sigma, \xi)$ be a random variable representing the excesses that follows a GP distribution with scale parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$. The survival function of Y is provided in equation (1) and obey the following domain restrictions: $y \geq 0$ when $\xi \geq 0$ and $y \in [0, -\sigma/\xi)$ when $\xi < 0$.

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

The shape parameter describes the upper tail behavior and can be thought of as a way to associate a hazard level to Y. Indeed, the larger the shape parameter is, the higher the probability of extreme events. If $\xi > 0$, Y is said to have a heavy or Pareto-type upper tail and $\mathbb{P}(Y > y) \approx 1/y^{1/\xi}$ for $y \uparrow \infty$. The upper tail is said to be light or exponential when $\xi = 0$ and $\mathbb{P}(Y > y) \approx 1/\exp(-y)$ for $y \uparrow \infty$. The upper tail is bounded for $\xi < 0$.

2.1.1. Return Levels

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

$$I(T) = \begin{cases} u + \frac{\sigma}{\xi}((T N_{exc})^{\xi} - 1) & \text{if } \xi \neq 0 \\ u + \sigma \log(T N_{exc}) & \text{if } \xi = 0 \end{cases} \tag{2}$$

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

Underestimation of the shape parameter leads to underestimation of return levels with greater discrepancies for longer return periods. Indeed, it can be seen from equation (2) that the larger T is, the greater the influence of the value of the shape parameter on the return level $I(T)$.

2.1.2. Probability Weighted Moment Estimators

To estimate the GP parameters, we rely on a method based on probability weighted moments (PWM). For $r \geq 0$, the PWMs of $Y \sim G(\sigma, \xi)$ are given by [*Diebolt et al.*, 2007]:

$$\mathbb{E}[Y \bar{G}(Y; \sigma, \xi)^r] = \frac{\sigma}{(1+r)(1+r-\xi)}. \tag{3}$$

Provided that $\xi < 1$, let $\mu = \mathbb{E}[Y] = \sigma/(1-\xi)$ be the first probability weighted moment of Y (plug $r = 0$ in equation (3)). Let us define $Z = Y/\mu$. We have that $Z \sim G(1-\xi, \xi)$ since, by making use of equation (1):

$$\mathbb{P}(Z > z) = \mathbb{P}(Y > \mu z) = \mathbb{P}\left(Y > \frac{\sigma z}{1-\xi}\right) = \bar{G}(1-\xi, \xi). \tag{4}$$

Note that the first PWM of Z is equal to one, i.e., $\mathbb{E}[Z] = 1$. Let ν denote the second PWM of Z:

$$\nu = \frac{1-\xi}{4-2\xi}. \tag{5}$$

We express the GP parameters ξ and σ as functions of the two aforementioned PWMs, ν and μ:

$$\xi = \frac{1-4\nu}{1-2\nu} \quad \text{and} \quad \sigma = \mu(1-\xi). \tag{6}$$

Sample estimates of PWMs can be computed using U-statistics [*Furrer and Naveau*, 2007].

2.2. Conditional Mixture of GPs

We assume that the region of interest can be partitioned into N_{reg} subregions with different hazard levels. In terms of GP distribution, this translates into each subregion having distinct but constant shape parameters. The scale parameter is assumed to vary smoothly as a function of the covariates.

For a given site, let Y be the random variable representing the excesses above a high enough threshold and let $\boldsymbol{x}$ be a vector of covariates associated to the site. In addition, let C be a discrete random variable taking values in $\{1, \ldots, N_{reg}\}$ representing the label of the subregion to which the site belongs. Then Y can be thought of as following a conditional mixture whose distribution is given by:

$$\mathbb{P}(Y \leq y|\boldsymbol{x}) = \sum_{j=1}^{N_{reg}} \mathbb{P}(C=j|\boldsymbol{x})\mathbb{P}(Y \leq y|C=j, \boldsymbol{x}), \tag{7}$$

where $\mathbb{P}(C=j|\boldsymbol{x})$ is the probability of belonging to the j^{th} subregion given the covariates $\boldsymbol{x}$ and $\mathbb{P}(Y|C=j, \boldsymbol{x})$ is the conditional distribution of Y given that it belongs to the j^{th} subregion.

Thanks to the GP tail approximation and the assumptions on the GP parameters, we have that:

$$\mathbb{P}(Y \leq y|C=j, \boldsymbol{x}) = G(y; \sigma(\boldsymbol{x}), \xi_j), \tag{8}$$

where G is the distribution function given in equation (1) and with the scale parameter $\sigma(\cdot)$ a smooth function of $\boldsymbol{x}$ and ξ_j the shape parameter of the j^{th} subregion.

2.2.1. Inference and Interpolation

We develop a two-step inference strategy adapted to the fact that the subregion label is not observed, i.e., C is a hidden (or latent) variable. Both steps, called E and M steps by analogy with the Expectation-Maximization (EM) algorithm [*Dempster et al.*, 1977], relies on PWM estimators (section 2.1.2). The first step or E-step aims to assign hazard levels to the sites by estimating C, i.e., it aims to estimate the partition of the subregions. In the second step or M-step, $\mathbb{P}(C=j|\boldsymbol{x})$ along with the GP scale parameter $\sigma(\cdot)$ and the shape parameters $\{\xi_1, \ldots, \xi_{N_{reg}}\}$ are estimated.

Let M be the number of gauged sites in the region of interest. For a given gauged site i, let $\{y_{i1}, \ldots, y_{in_i}\}$ be the n_i observed excesses and let $\boldsymbol{x}_i$ be the corresponding vector of covariates. In addition, let $C_i \in \{1, \ldots, N_{reg}\}$ be the unobserved subregion label.

2.2.1.1. E-Step: Partitioning into Hazard Subregions

Let $Y_i|C_i, \boldsymbol{x}_i \sim G(\sigma(\boldsymbol{x}_i), \xi_{C_i})$. Similar expressions as in section 2.1.2 can be developed with σ replaced by $\sigma(\boldsymbol{x}_i)$ and ξ by ξ_{C_i}. In particular, given C_i, equation (4) yields:

$$Z_i = \frac{Y_i|C_i, \boldsymbol{x}_i}{\mu(\boldsymbol{x}_i)} \sim G(1-\xi_{C_i}, \xi_{C_i}), \tag{9}$$

where the conditional average of the excesses is given by $\mu(\boldsymbol{x}_i) = \sigma(\boldsymbol{x}_i)/(1-\xi_{C_i})$.

To estimate $\mu(\cdot)$, we employ kernel regression, a nonparametric approach [*Nadaraya*, 1964; *Watson*, 1964]. The implementation details are provided in section 2.2.3. Let $\hat{\mu}(\cdot)$ be the kernel regression estimator of $\mu(\cdot)$. The estimated scaled excesses are then given by $\hat{z}_{ik} = y_{ik}/\hat{\mu}(\boldsymbol{x}_i)$ for $1 \leq i \leq M$ and $1 \leq k \leq n_i$.

For any two sites $1 \leq i, j \leq M$, let ν_i and ν_j be the second PWMs of Z_i and Z_j, respectively (see equation (5)). Since ν_i and ν_j only depend on ξ_{C_i} and ξ_{C_j} respectively, we have that

$$\nu_i = \nu_j \iff C_i = C_j. \tag{10}$$

The partitioning into hazard subregions can be thought of as an unsupervised classification problem. Let $\hat{\nu}_i$ be estimators computed using U-statistics [*Furrer and Naveau*, 2007] of the second PWM of Z_i, $1 \leq i \leq M$ based on $\{\hat{z}_{i1}, \ldots, \hat{z}_{in_i}\}$, the sample of estimated scaled excesses at site i. We resort to K-Means [*Ripley*,

1996] applied to $\hat{\nu}_i$ to obtain a partition of the hazard subregions. Any other statistic of Z_i, such as higher PWMs, could be considered to form the partition.

2.2.1.2. M-Step: Parameter Estimation

The estimation of the conditional probability of belonging to a subregion follows from the partitioning obtained in the *E-Step*. Since K-Means yields *hard* assignment rules, i.e., a site is or is not in a subregion, we have that $\hat{\mathbb{P}}(C=j|\boldsymbol{x}_i)=1_{\{\hat{C}_i=j\}}$. In the conditional mixture setup, the expressions in equation (6) to estimate the GP parameters become:

$$\xi_j=\frac{1-4\nu_j}{1-2\nu_j}\,1\le j\le N_{reg} \tag{11}$$

$$\sigma(\boldsymbol{x}_i)=\mu(\boldsymbol{x}_i)(1-\xi_j)\ \text{if}\ C_i=j. \tag{12}$$

For a given subregion j, the scaled excesses from all the sites in the subregion can be pooled together to estimate ν_j. Let $\hat{C}_i$ be the estimated subregion labels obtained at the *E-Step*, i.e., the cluster labels determined by K-Means. For each $1\le j\le N_{reg}$, ν_j is estimated from the pooled sample $\{\hat{z}_{ik}, 1\le k\le n_i : \hat{C}_i=j\}$, i.e., the set of estimated scaled excesses from all the sites that are assigned to subregion j. The estimated shape parameter $\hat{\xi}_j$ for each subregion is then obtained through equation (11) and the estimated scale parameter $\hat{\sigma}(\boldsymbol{x}_i)$ is provided by equation (12). See Appendix A for details of the related inference strategy developed in *Naveau et al.* [2014].

2.2.1.3. Interpolation to Ungauged Sites

Let i^* be a target site, poorly gauged or ungauged where we wish to estimate the parameters of the GP distribution according to the model of equation (7). The first step (or *E-Step*) is to assign a hazard level to i^*, i.e., to estimate the subregion label $C_{i^*}\in\{1,\ldots,N_{reg}\}$. This is a supervised classification problem for which we use the k-nearest neighbor rule, a nonparametric classifier [*Ripley*, 1996], based on the covariates $\boldsymbol{x}_{i^*}$ and $\boldsymbol{x}_i$, $1\le i\le M$. See section 2.2.3 for the implementation details.

In the *M-Step*, as the k-nearest neighbor rule also yields *hard* assignment rules, $\hat{\mathbb{P}}(C=j|\boldsymbol{x}_{i^*})=1_{\{\hat{C}_{i^*}=j\}}$. The estimated shape parameter at i^* is given by $\hat{\xi}_{\hat{C}_{i^*}}$. Then $\hat{\sigma}(\boldsymbol{x}_{i^*})$ is obtained via equation (12) in which the estimated $\hat{\mu}(\cdot)$ is applied to the covariates $\boldsymbol{x}_{i^*}$.

2.2.2. Selection of the Number of Hazard Subregions

The number of hazard subregions is the number of mixture components and it controls the complexity level of the conditional mixture of equation (7). Indeed, the number of components is equal to the number of free parameters in the mixture which corresponds to the shape parameters of the GP for each subregion. Therefore, increasing the number of subregions provides the conditional mixture with a greater ability to adapt to the variability of the hazard level in the region. On the other hand, added adaptability translates into higher variance of the conditional mixture since more parameters must be estimated from the data. This is particularly an issue in our setup since the shape parameter is notoriously difficult to estimate and is subject to considerable uncertainty [*Coles*, 2001].

We address this bias-variance trade-off (sufficient adaptability while keeping variance under control) by selecting the number of subregions such that the conditional mixture yields the highest performance on out-of-sample data. In practice, we implement cross validation as follows. A small set of sites is held out to assess performance while the remainder of the sites are used for parameter estimation. The procedure is repeated for all possible choices for the held-out set of sites and the performance is computed from all the held-out sets of sites [see *Bishop*, 2011, section 1.3].

The performance is measured in terms of loss functions which provide a measure of how poorly models perform. Thus, the best model is the one yielding the smallest lost. In order to stress the contribution of the shape parameter, the lost functions are computed on the estimated scaled excesses $\hat{z}_{ik}$. Due to the *hard* assignment rules used to form the partition, the fitted model for the scaled variable at a given held-out site i reduces to:

$$\hat{\mathbb{P}}(Z_i\le z|\boldsymbol{x}_i)=\hat{\mathbb{P}}(Y_i\le\hat{\mu}(\boldsymbol{x}_i)z|\boldsymbol{x}_i)=\sum_{j=1}^{N_{reg}}1_{\{\hat{C}_i=j\}}G(z;1-\hat{\xi}_j,\hat{\xi}_j)=G(z;1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i}). \tag{13}$$

We consider three lost functions to evaluate the fitted conditional mixture. The first loss function is the negative log-likelihood of the estimated scaled excesses which is given by:

$$-\sum_{i=1}^{M}\sum_{k=1}^{n_i}\ln g(\hat{z}_{ik};1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i}), \tag{14}$$

where g is the density function of the GP distribution.

The second loss function is the sum-of-squares quantile error:

$$\sum_{i=1}^{M}\sum_{k=1}^{n_i}\left(\hat{z}_{i(k)}-G^{-1}((k-0.5)/n_i;1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i})\right)^2 \tag{15}$$

where $\hat{z}_{i(1)}\leq\ldots\leq\hat{z}_{i(n_i)}$ are the ordered estimated scaled excesses and G^{-1} is the quantile function of the GP distribution computed on empirical frequencies.

The third loss function is the Anderson-Darling statistic that is "sensitive to discrepancies at the tails of the distribution" [*Anderson and Darling*, 1954]:

$$\sum_{i=1}^{M}\left\{-n_i-\frac{1}{n_i}\sum_{k=1}^{n_i}(2k-1)\left[\ln G(\hat{z}_{i(k)};1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i})+\ln\left(1-G(\hat{z}_{i(n_i-k+1)};1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i})\right)\right]\right\}. \tag{16}$$

2.2.3. Implementation Details for Fixed N_{reg}

The algorithm for the regional peaks-over-threshold approach is described in details in Algorithm 1 and is implemented in a package available upon request in the R environment [*R Core Team*, 2016]. Each step of the algorithm is illustrated on synthetic data generated according to a model that satisfies the assumptions of the approach. More precisely, let $x\in[0,1000]$ be a one-dimensional covariate, then:

$$Y|x\sim G(y;\sin(2\pi x/250)+\exp(x/500),\xi_C). \tag{17}$$

For the illustration, we use the sample in Figure 1a with four hazard subregions, i.e., $C\in\{1,2,3,4\}$ such that:

$$\begin{aligned}\xi_1&=0.3 \quad \text{if} \quad x\leq 250\\ \xi_2&=0.2 \quad \text{if} \quad 250<x\leq 500\\ \xi_3&=0.1 \quad \text{if} \quad 500<x\leq 750\\ \xi_4&=0 \quad \text{if} \quad x>750.\end{aligned} \tag{18}$$

The shape parameter values are chosen so as to span approximately the range of estimated values for the precipitation data, see section 4. The scale parameter is taken as a combination of periodic and exponential signal. Therefore, the conditional mean also follows a combination of periodic and exponential signal but with discontinuities at the borders of the subregions since $\mu(x_i)=\sigma(x_i)/(1-\xi_{C_i})$. We can thus assess the ability of kernel regression at estimating a nonlinear function with a small number of discontinuities.

To perform the so-called *E-Step* of the inference scheme in section 2.2.1, the conditional mean $\mu(\cdot)$ must first be estimated (see Figure 1b) and then the estimated scaled excesses can be computed $\hat{z}_{ik}=y_{ik}/\hat{\mu}(\boldsymbol{x}_i)$, $1\leq k\leq n_i$ and $1\leq i\leq M$ (see Figure 2a).

Let $K_h(\cdot)$ be a kernel function that can be thought of as a symmetric density function for which the so-called *bandwidth* h, which controls the amount of smoothing in the regression, acts as a scale parameter. The kernel regression estimator of $\mu(\boldsymbol{x})$ is given by:

$$\hat{\mu}(\boldsymbol{x})=\frac{1}{\sum_i K_h(\boldsymbol{x}-\boldsymbol{x}_i)}\sum_{i=1}^{M}\left(\frac{1}{n_i}\sum_{k=1}^{n_i}y_{ik}\right)K_h(\boldsymbol{x}-\boldsymbol{x}_i). \tag{19}$$

We rely on the `np` package of `R` [*Hayfield and Racine*, 2008] that implements kernel regression with various types of kernels and several automated bandwidth selection methods. We employed the Epanechnikov kernel which is optimal in the sense that it minimizes the asymptotic mean integrated square error [*Epanechnikov*, 1969; *Abadir and Lawford*, 2004]. Bandwidth selection is performed with cross validation [*Li and Racine*, 2004].

For each site $1\leq i\leq M$, v_i, is estimated from $\{\hat{z}_{ik},1\leq k\leq n_i\}$ with U-statistics [*Furrer and Naveau*, 2007] and then smoothed with kernel regression similarly as for $\mu(\cdot)$ to reduce the sampling variability. K-Means

Algorithm 1: Regional peaks-over-threshold model with fixed N_{reg}

input : N_{reg} the number of subregions ;

$\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{in_i}\}$ observed excesses, $\boldsymbol{x}_i$ vector of covariates for each site $1 \leq i \leq M$;

$\boldsymbol{x}_{i^*}$ $1 \leq i^* \leq M^*$ for ungauged sites (optional)

output $\{\hat{\xi}_1, \ldots, \hat{\xi}_{N_{reg}}\}$ shape parameter estimates for each subregion ;

:

$\hat{\sigma}(\boldsymbol{x}_i)$, $\hat{C}_i$ $1 \leq i \leq M$ the scale parameter and the subregion label estimates for each site ;

$\hat{C}_{i^*}$ and $\hat{\sigma}(\boldsymbol{x}_{i^*})$ 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 $\boldsymbol{x}_i$;

2 Compute the scaled excesses $\hat{z}_{ik} = y_{ik}/\hat{\mu}(\boldsymbol{x}_i)$ for $1 \leq k \leq n_i$;

3 **if** $N_{reg} == 1$ **then** // `single subregion case`

4 Assign all the sites to a single subregion $\hat{C}_i = 1\ \forall i$ and $\hat{C}_{i^*} = 1\ \forall i^*$;

5 **else**//`partitioning into` N_{reg} `subregions`

6 Estimate v_i by regressing the second PWM sample estimate of Z_i over $\boldsymbol{x}_i$;

7 Estimate C_i i.e., assign each site i to a subregion j, $1 \leq j \leq N_{reg}$ by clustering $\hat{v}_i$;

8 Estimate C_{i^*} i.e assign each i^* to a subregion j with a classifier based on $\boldsymbol{x}_{i^*}$ and $\boldsymbol{x}_i$;

9 **end**

10 **for** $j \leftarrow 1$ **to** N_{reg} **do**

11 Estimate v_j from all $\hat{z}_{ik}$, $1 \leq k \leq n_i$, such that $\hat{C}_i = j$;

12 Estimate ξ_j, the shape parameter of subregion j as $\hat{\xi}_j = (1-4\hat{v}_j)/(1-2\hat{v}_j)$;

13 Estimate $\sigma(\boldsymbol{x}_i)$ thanks to $\hat{\sigma}(\boldsymbol{x}_i) = \hat{\mu}(\boldsymbol{x}_i)(1-\hat{\xi}_{\hat{C}_i})$;

14 Estimate $\sigma(\boldsymbol{x}_{i^*})$ similarly ;

15 **end**

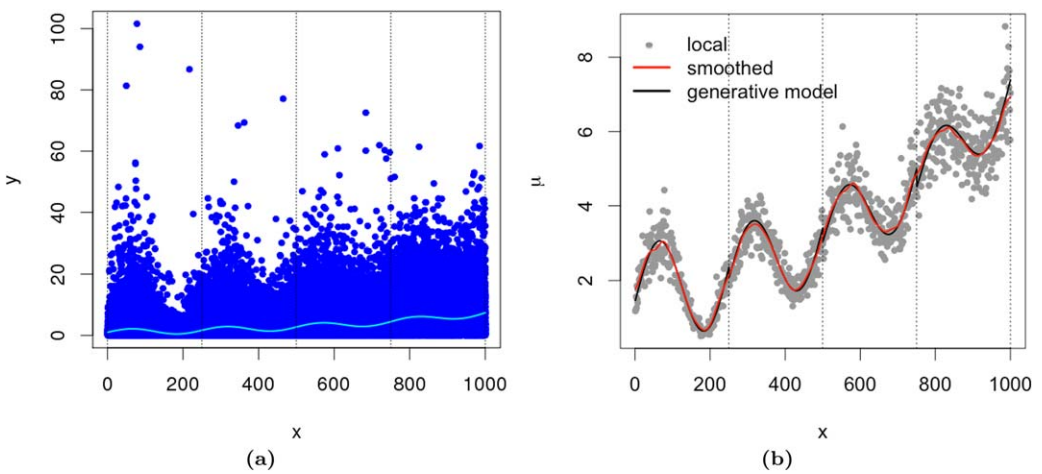

Figure 1. (a) Synthetic data set made of $n_i = 100$ random GP values Y with $x = i$ and $1 \leq i \leq 1000$ from the generative model in equations (17) and (18) whose scale parameter is represented by the cyan curve. (b) At each x, the sample average of the excesses is computed (gray dots) and then smoothed with kernel regression to obtain $\hat{\mu}(x_i)$ (red curve). The black curve represents the conditional mean of the generative model $\mu(x) = \sigma(x)/(1-\xi_C)$. See line 1 in Algorithm 1. In both figures, the hazard subregions of the generative model are defined by the vertical bands.

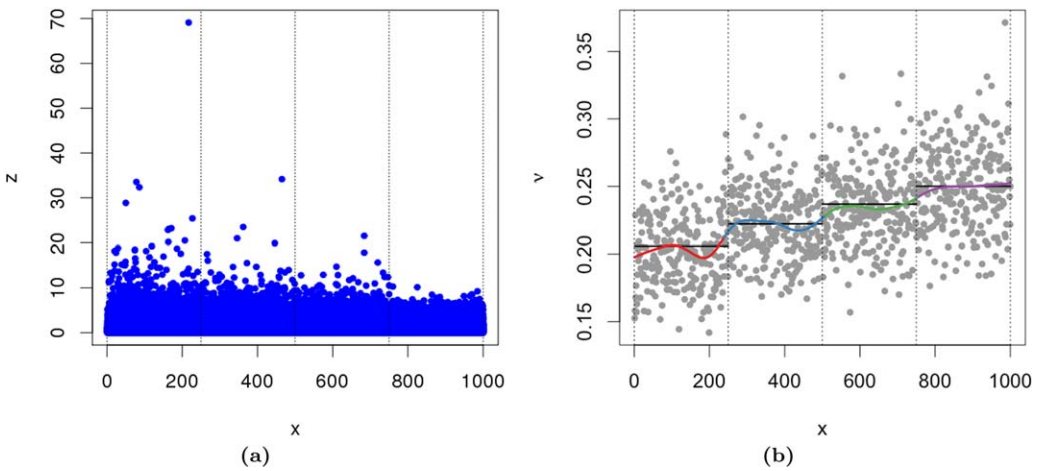

Figure 2. (a) Estimated scaled excesses $\hat{z}_{ik}=y_{ik}/\hat{\mu}(x_i)$, $1 \le k \le n_i$ and $1 \le i \le M$ of the data set in Figure 1a with $\hat{\mu}(x_i)$ as the red curve in Figure 1b. See line 2 in Algorithm 1. (b) The colored curves represent $\hat{v}_i$, the second PWM of Z_i estimate, obtained by smoothing with kernel regression the estimates computed from $\{\hat{z}_{ik}, 1 \le k \le n_i\}$ at each site (gray dots). Hazard subregions, identified by colors, are determined by applying K-Means to $\hat{v}_i$ with $N_{reg}=4$ clusters. The horizontal black lines represent the v values of the generative model. See lines 3–9 in Algorithm 1. In both figures, the hazard subregions of the generative model are defined by the vertical bands.

(implementation of the stats package in R) is then applied to the smoothed $\hat{v}_i$ estimate in order to partition the sites into subregions corresponding to clusters (see Figure 2b where the number of clusters is equal to the number of subregions of the generative model in equation (18)).

To ensure that K-Means always converges to the same partition, we set initial cluster centers to N_{reg} empirical quantiles of $\hat{v}_i$, $1 \le i \le M$, with probabilities that spread regularly the $[0,1]$ interval. The sites are iteratively assigned to the cluster whose center is closer in terms of $\hat{v}_i$ and then the cluster centers are updated as the averages of the $\hat{v}_i$ of the sites belonging to each cluster.

The k-nearest neighbor rule assigns a subregion label to ungauged sites based on the partition established by K-Means. The k nearest neighbors of an ungauged site i^* are determined from the Euclidean distances $d(\boldsymbol{x}_{i^*}, \boldsymbol{x}_i)$ for all $i \in \{1,\ldots,M\}$. The subregion C_{i^*} is the result of a majority vote among the k nearest neighbors. We set the number of neighbor to $k=5$, although it could be optimized just like the bandwidth of kernel regression.

In the so-called *M-Step* in section 2.2.1, v_j is estimated for each subregion $1 \le j \le N_{reg}$ with the pooled sample $\{\hat{z}_{ik}, 1 \le k \le n_i : \hat{C}_i=j\}$. The GP parameters are then estimated thanks to equations (11) and (12). The parameters of the generative model are shown as black curves in Figure 3 while the gray bands represent 95% confidence intervals obtained with parametric bootstrap (1000 replications).

3. Synthetic Data Study

We simulate synthetic data sets that vary in terms of generative models, number of sites, and sample sizes with the aim to evaluate the regional peaks-over-threshold model.

The two generative models considered share the functional form of equation (17) for the scale parameter but differ in terms of partitions into hazard subregions. The first partition has two subregions with very distinct hazard level values:

$$\begin{aligned} \xi_1 &= 0.3 \quad \text{if} \quad x \le 500 \\ \xi_2 &= 0 \quad \text{if} \quad x > 500. \end{aligned} \tag{20}$$

The second partition has four subregions with closer hazard level values as given in equation (18).

We only include differences in partitions in the generative models in order to assess the main contributions of the regional peaks-over-threshold model, namely the *E-Step* to determine the partition and the cross-validation procedure to select the number of subregions. Note that the estimation of the partition

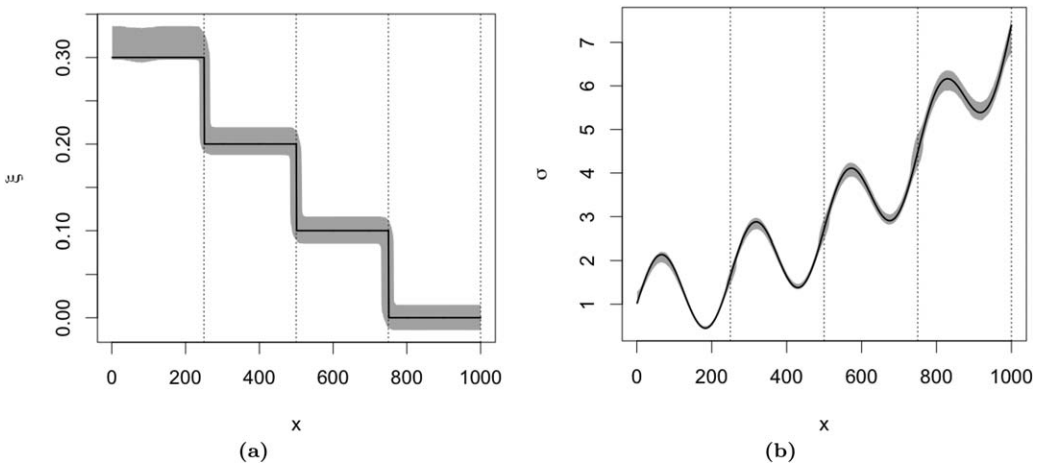

Figure 3. 95% confidence bands in gray obtained with parametric bootstrap (1000 replications) for the GP parameter estimates (*M-Step* in section 2.2.1). The parameters of the generative model are shown as black curves. (a) Shape parameter estimates (equation (11) with ν_j estimated from the pooled sample $\{\hat{z}_{ik}, 1 \leq k \leq n_i : \hat{C}_i = j\}$). (b) Scale parameter estimates (equation (12)). See lines 12 to 15 of Algorithm 1. In both figures, the hazard subregions of the generative model are defined by the vertical bands.

influences both the shape and scale parameter estimates (see equations (11) and (12)). On the other hand, as mentioned in section 2.2.3, the functional form for $\sigma(\cdot)$ is reasonably challenging for kernel regression.

There are, in all, 18 different types of synthetic data sets with either two or four subregions, made of M sites for which x is sampled randomly in $[0, 1000]$ with $M = 100$, 200, or 400 and the sample size is either 25, 50, or 100 for each site. Each synthetic data set is replicated 1000 times to assess uncertainty.

3.1. Selection of the Number of Hazard Subregions

We select the number of hazard subregions as the partition size that yields the highest out-of-sample performance, as described in section 2.2.2. We consider partitions with either 1, 2, 4, or 8 hazard subregions (a classical geometric progression). This choice of partition sizes is made to keep computation time tractable.

The cross-validation procedure is implemented as follows. The size of the held-out sets is established so that, whichever synthetic data set, there are 100 held-out sets. More precisely, with $M = 100, 200, 400$, the held-out sets contain 1, 2, and 4 sites, respectively.

For each of the three loss functions (see equations (14–16)), we computed the percentage of times, out of 1000 replications, each number of subregions is selected. The results for the negative log-likelihood loss function are presented in Figure 4 for each of the 18 types of synthetic data sets. For the sum-of-squares quantile error and the Anderson-Darling statistic, the results are provided in the supporting information.

3.2. Estimation of the GP Parameters

Once the number of hazard subregions is selected via cross validation, the conditional mixture of GPs is estimated anew on the whole data set. The GP parameters are then interpolated to fictitious "ungauged" sites, i.e., on a test set. The test set is made of 1001 sites for which x takes integer values $\{0, 1, \ldots, 1000\}$. Thanks to parametric bootstrap (1000 replicates), 95% confidence bands are computed.

Figures 5 and 6 presents, for the two-region and four-region generative models, respectively, the 95% confidence bands obtained for the smallest data set (100 sites and a sample size of 25) and the largest data set (400 sites and a sample size of 100), when the negative log-likelihood loss function is employed to select the partition size. The results for the other two loss functions are provided in the supporting information.

4. Daily Precipitation Data Application

We focus on the area in the French Mediterranean shown in Figure 7a. The Rhône river valley (in shades of dark green in Figure 7a) runs from north to south and encompasses the cities of Valence and Montpellier.

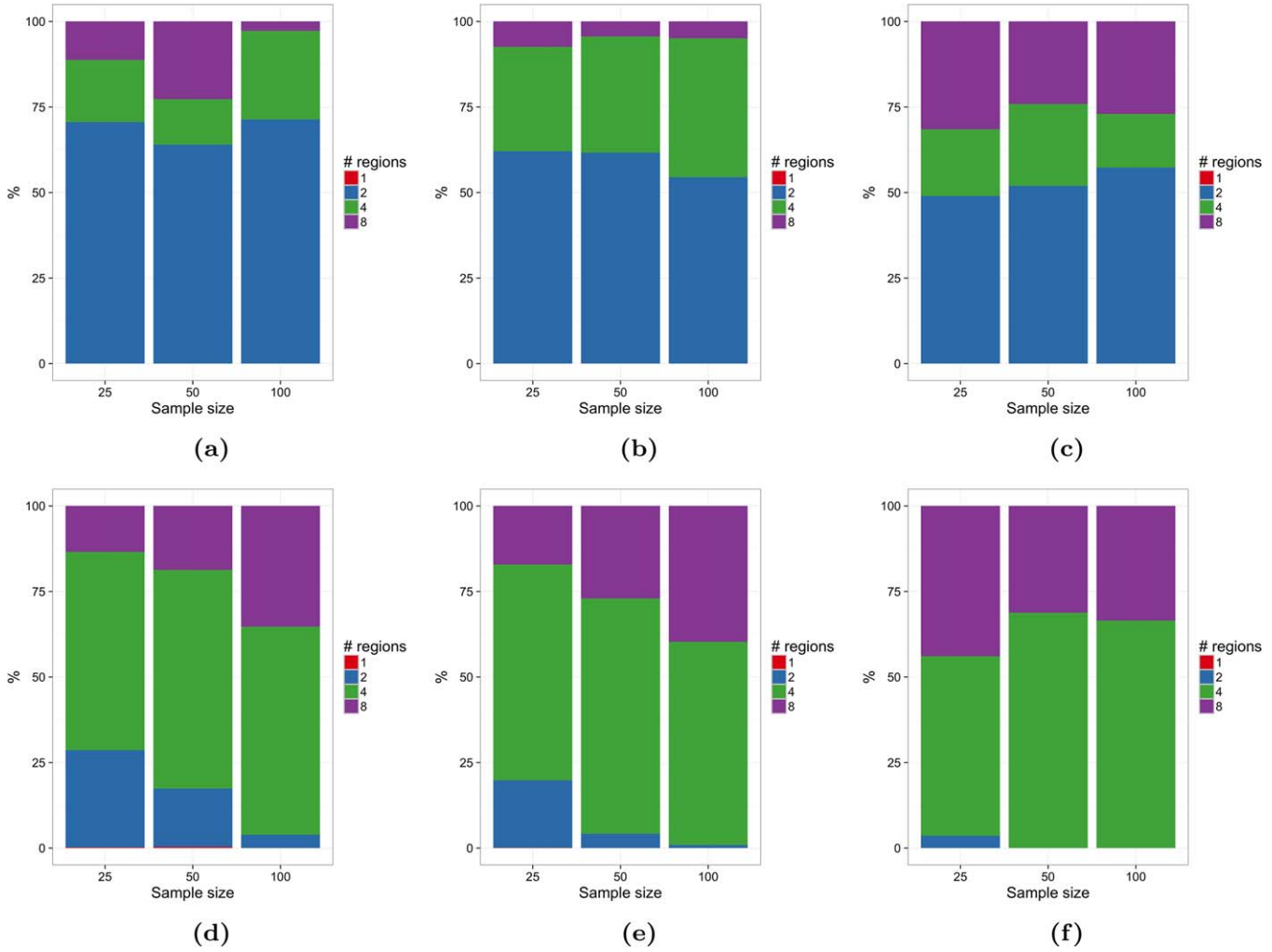

Figure 4. Selection of each partition size (# of subregions) in % over 1000 replications for (a)–(c) the two subregion generative model and (d)–(f) the four subregion generative model. In Figures 4a and 4d, the number of sites is $M = 100$, in Figures 4b and 4e $M = 200$ and in Figures 4c and 4f $M = 400$. The negative log-likelihood loss function is used to assess performance, see section 2.2.2.

On the left bank of the river, sits the prealps (highest point about 2700 m) while on the right bank, sits the Cevennes mountain range (highest point about 1700 m). The latter is well-known for the intense rainfall events called *épisodes cévenols* that occur mainly in autumn [*Delrieu et al*., 2005; *Braud et al*., 2014].

We selected 332 stations of the Météo-France network, the French weather service, depicted in Figure 7b, that belong to the French Mediterranean area shown in Figure 7a. Daily precipitation measurements are collected over the period 1 January 1958 to 31 December 2014 (57 years). In Figure 7b, the size of the plotting symbol is proportional to the length of the observation period available (from 10 to 57 years). The color indicates the percentage of missing values over the observation period (from 0% in light orange to 10% in dark red). Two contour level curves, 400 m and 800 m, of the digital elevation map in dark and light shades of gray recall the orography of the area.

A regular grid (approximately 500 m) is set up to cover the region where the stations lie. Interpolation of the distribution of heavy precipitation is carried out onto the grid. The vector of covariates $\boldsymbol{x}$ is taken as the x and y coordinates (extended Lambert II projections of latitude and longitude).

For each station, the threshold that defines the excesses for the application of the peaks-over-threshold approach (section 2.1) is set to the 98% quantile of the precipitation intensities (the observations greater than 0.1 mm which is the sensitivity of daily rain gauges). This is in line with conventional choices of threshold, see for instance *Roth et al*. [2012]. The resulting overall number of excesses per station ranges from 20

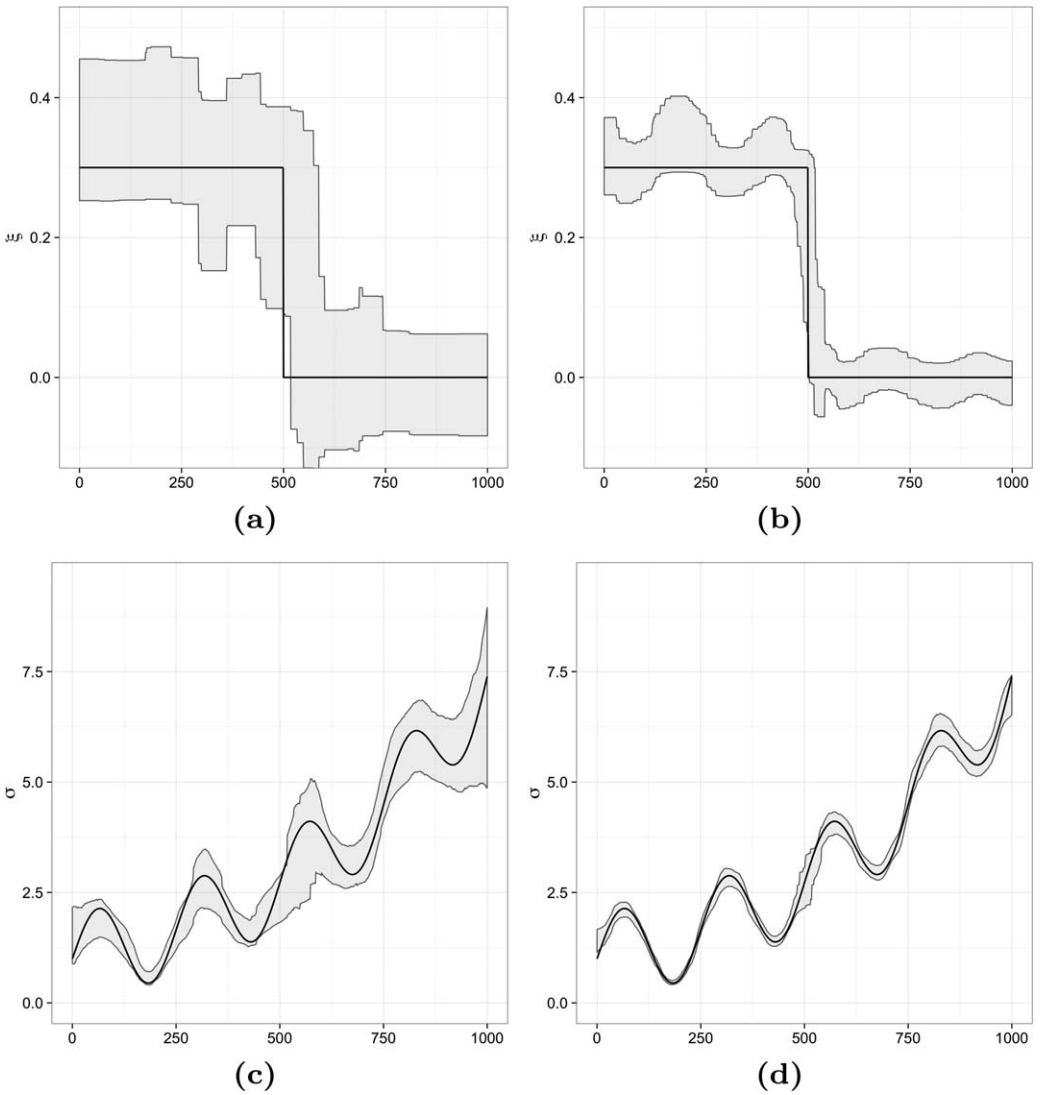

Figure 5. 95% confidence bands (gray) obtained by interpolating on the test set ($x \in \{0, 1, \ldots, 1000\}$) with 1000 replicates (a) and (b) the shape and (c) and (d) the scale parameters of the GP distribution of the two-region generative model (black curves). (a) and (c) Results for the smallest data set (100 sites and a sample size of 25) and (b) and (d) for the largest data set (400 sites and a sample size of 100). The partition size is selected with the negative log-likelihood loss function, see section 2.2.2.

to 191. The threshold and the average number of excesses per year are estimated at each station and then interpolated onto the regular grid with kernel regression (see section 2.2.3 and equation (19)). Figure 8 shows the interpolation results.

4.1. Interpolation Approaches

We compare three approaches to interpolate the distribution of heavy precipitation, i.e., the distribution of the excesses above a high threshold. Each approach differs in the way the shape parameter is assumed to vary in the region.

The first approach considered is the regional peaks-over-threshold model with a single region, i.e., the number of subregions is fixed to one. Therefore, the shape parameter is assumed to be constant (this is equivalent to the *Naveau et al*. [2014] approach, see Appendix A). The shape parameter estimate obtained by pooling the 332 stations of Figure 7b together in equation (11) is $\hat{\xi}_1 = 0.11$.

In the second approach, we let the number of subregions be determined by the data as described in section 2.2.2. For the precipitation data, cross validation is implemented with held-out sets containing four sites

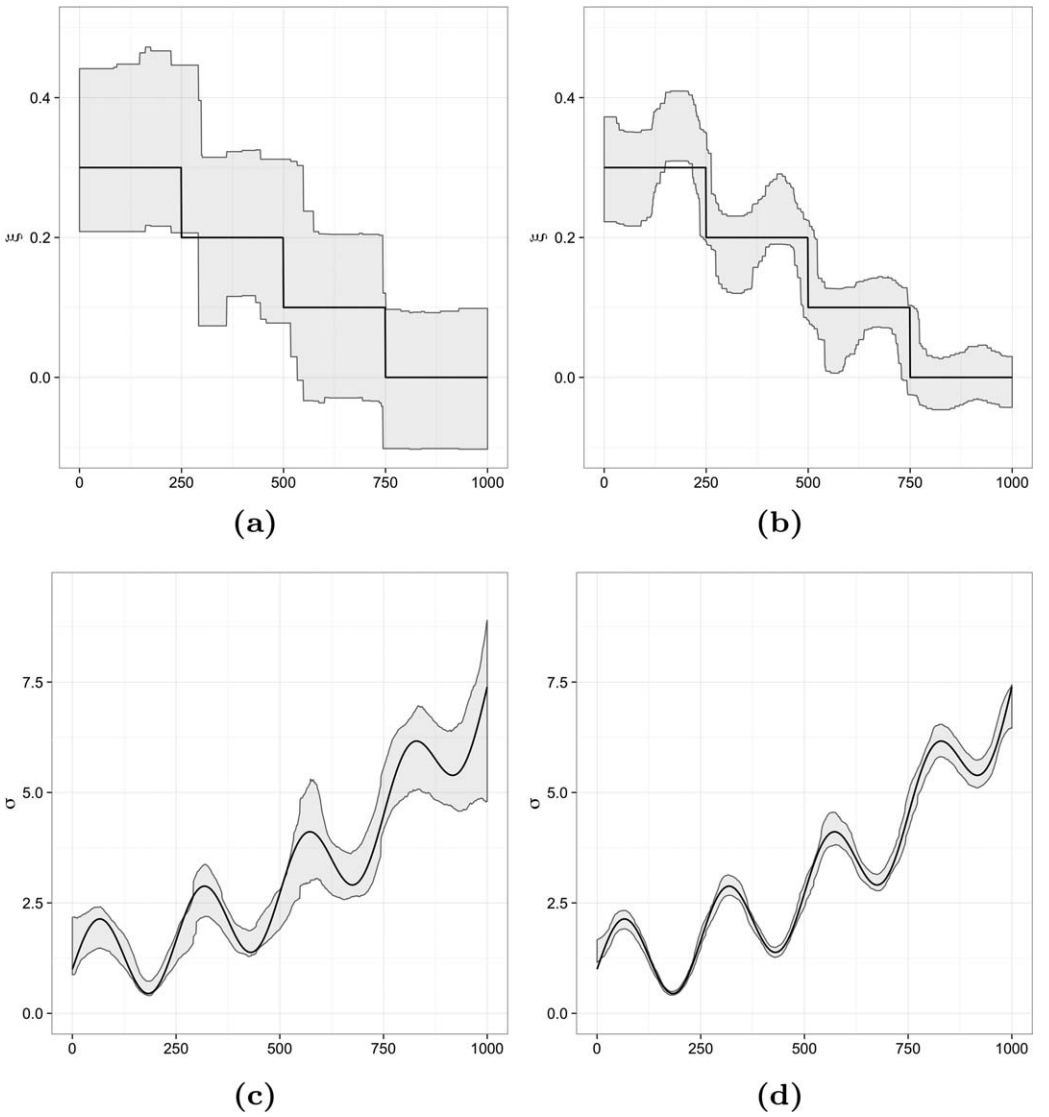

Figure 6. 95% confidence bands (gray) obtained by interpolating on the test set ($x \in \{0, 1, \ldots, 1000\}$) with 1000 replicates (a) and (b) the shape and (c) and (d) the scale parameters of the GP distribution of the four-region generative model (black curves). (a) and (c) Results for the smallest data set (100 sites and a sample size of 25) and (b) and (d) for the largest data set (400 sites and a sample size of 100). The partition size is selected with the negative log-likelihood loss function, see section 2.2.2.

taken at random (there are 83 such sets). A partition of size four yielded the highest performance according to the three loss functions (equations (14–16)). In Figure 9a, we see that the shape parameter estimate varies in a piecewise constant manner across the region. Each subregion has a different hazard level related to the values of the estimated shape parameter, namely $\hat{\xi}_1 = 0.01$, $\hat{\xi}_2 = 0.07$, $\hat{\xi}_3 = 0.12$, and $\hat{\xi}_4 = 0.29$. The difference between the interpolated scale parameter of the regional peaks-over-threshold model with a single region versus the four subregion partition is presented in Figure 10a.

In the third approach, the shape and scale parameters are estimated at each station thanks to equation (6) and then interpolated with kernel regression (see section 2.2.3 and equation (19)). In this approach, the shape parameter can vary almost freely. In Figure 9b, its estimated values range mainly from -0.05 to 0.3 with less than 5% of its values below -0.05 (in dark gray in Figure 9b with a minimum of -0.24). Figure 10b shows the difference between the interpolated scale parameter of this approach and the one of the regional peaks-over-threshold model with four subregions. Less than 1% of the differences are greater than 7 mm in magnitude (in dark gray in Figure 10b).

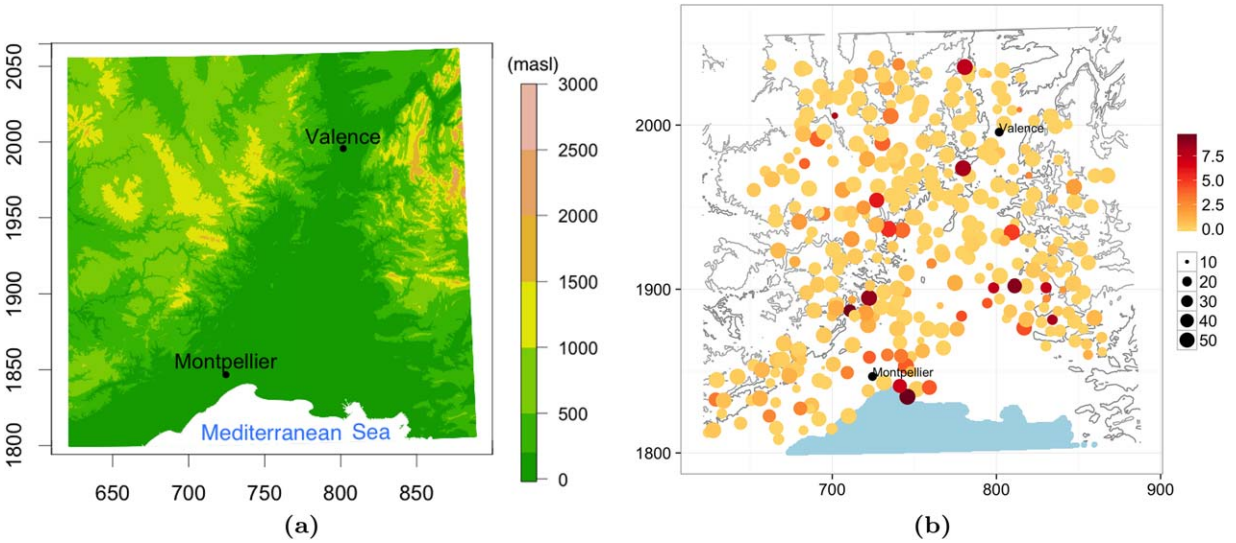

Figure 7. (a) Digital elevation map of the area of interest in the French Mediterranean. (b) 332 rain gauge stations of the Météo-France network (French weather service) covering the period 1 January 1958 to 31 December 2014 (57 years). The size of the symbol is proportional to the length of the observation period (10–57 years) and the color shade (light orange to dark red) indicates the percentage of missing values (0–10%).

4.2. Out-of-Sample Performance Comparison

Cross validation, with held-out sets of size four taken at random, is employed to evaluate the out-of-sample performance of each of the three interpolation approaches described previously. Note that cross validation was also applied to determine the number of hazard subregions of the second interpolation approach. This can be thought of as a form of double layer of cross validation since thanks to the randomization of the sites, the held-out sets are different in each layer [*Arlot and Celisse*, 2010].

With the GP parameter interpolated to the held-out sets, we computed the three loss functions from equations (14–16) on the excesses y_{ik} in order to account for the estimation of *both* the shape and the scale parameters. In addition, the loss functions are computed relatively to the regional peaks-over-threshold model with four hazard subregions. For instance, the negative log-likelihood loss function becomes:

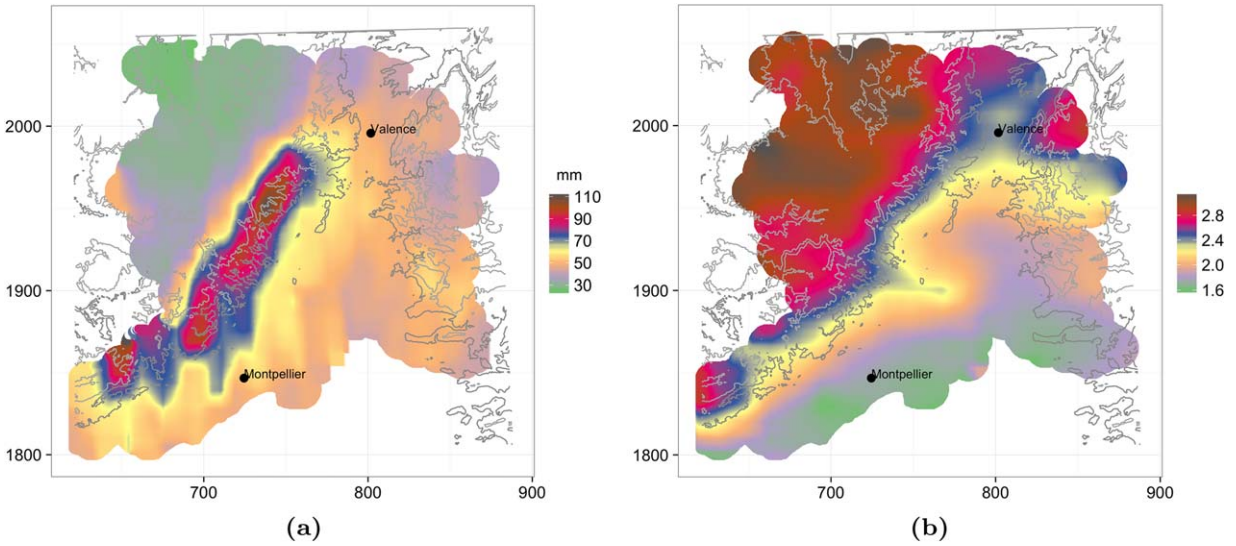

Figure 8. French Mediterranean precipitation data: (a) Threshold defined as the 98% quantile of precipitation intensities (b) Average number of excesses above the threshold per year. Interpolation onto the grid is performed with kernel regression.

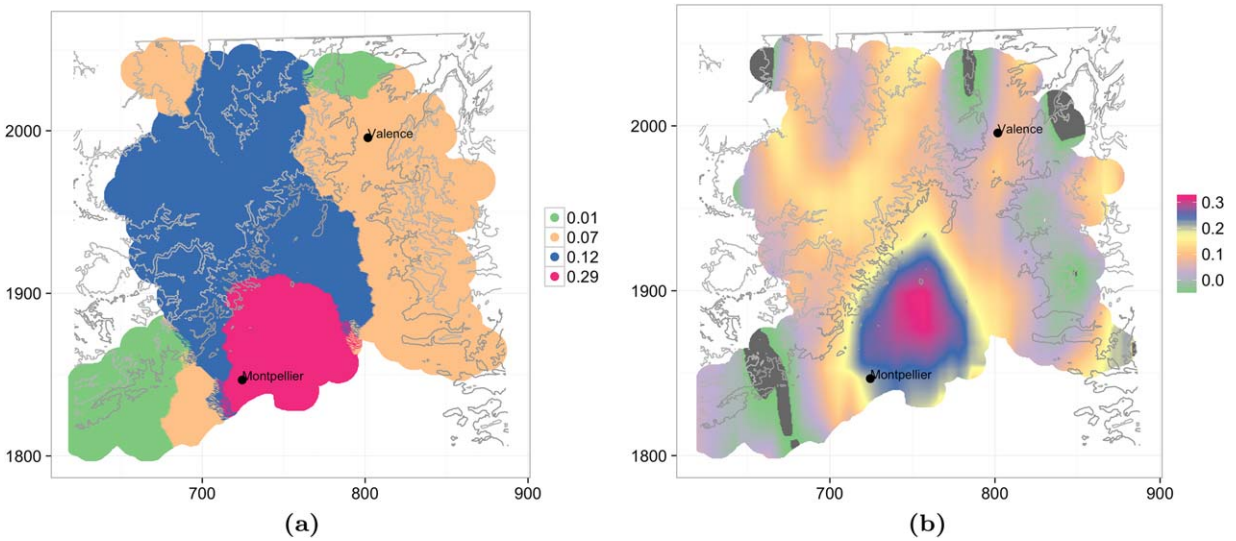

Figure 9. Interpolated shape parameter of the GP distribution onto the grid with (a) the regional peaks-over-threshold model with four subregions and (b) kernel regression applied to at-site shape parameter estimates. Less than 5% of the values are below −0.05 and are shown in dark gray. The shape parameter estimate of the regional peaks-over-threshold model with a single region is $\hat{\xi}_1 = 0.11$.

$$-\sum_{i=1}^{M}\sum_{k=1}^{n_i} \ln g(y_{ik}; \hat{\sigma}_i^O, \hat{\xi}_{\hat{C}_i}^O) + \sum_{i=1}^{M}\sum_{k=1}^{n_i} \ln g(y_{ik}; \hat{\sigma}_i^R, \hat{\xi}_{\hat{C}_i}^R), \tag{21}$$

where g is the density function of the GP distribution, $\hat{\sigma}_i^R$ and $\hat{\xi}_{\hat{C}_i}^R$ are the parameter estimates of the regional model with four hazard subregions while $\hat{\sigma}_i^O$ and $\hat{\xi}_{\hat{C}_i}^O$ are the estimates of one of the other interpolation approaches. Positive values of the relative log-likelihood from equation (21) indicate that the regional model with four hazard subregions outperforms the other interpolation approach. The two other relative loss functions based on the sum-of-squares quantile error in equation (15) and on the Anderson-Darling statistic in equation (16) are built similarly and can be interpreted in the same way.

Figure 11 shows the relative negative log-likelihood loss from equation (21) at each station for the regional model with a single region and the kernel regression interpolation of at-site estimates. At one station

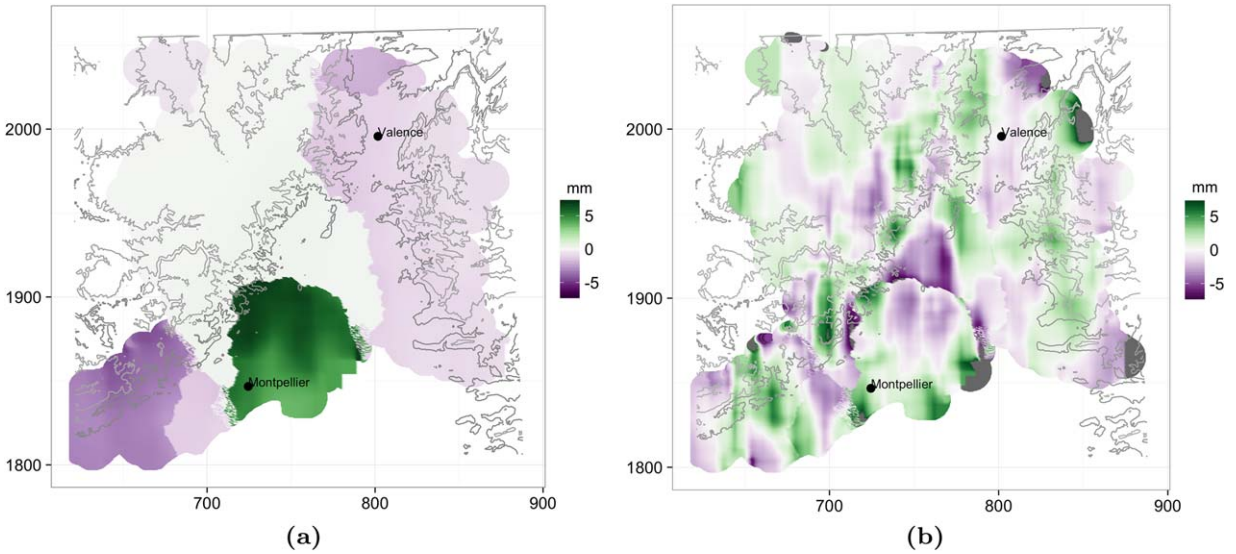

Figure 10. Differences in interpolated scale parameter of the GP distribution onto the grid: (a) The regional peaks-over-threshold model with a single region or (b) The kernel regression interpolation of the at-site estimates minus the interpolated scale parameter of the regional model with four subregions. In Figure 10b, less than 1% of the differences are greater than 7 mm in magnitude and are shown in dark gray.

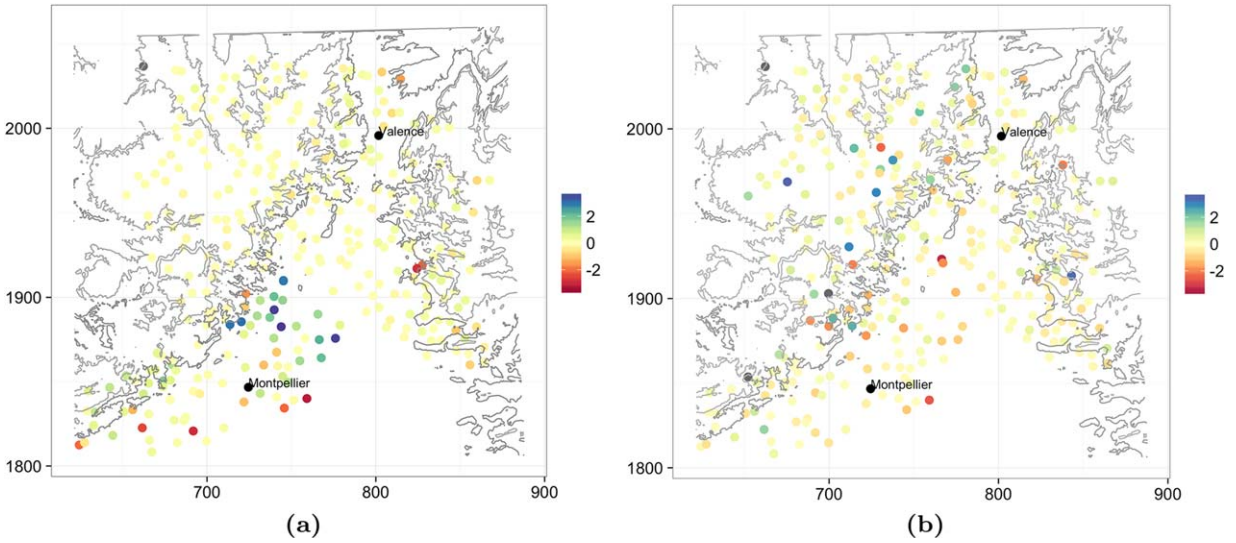

Figure 11. Negative log-likelihood relative to the regional model with four subregions. (a) For the regional model with a single region and (b) for the kernel regression interpolation of the at-site estimates. Positive values (blue shades) indicate that the regional model with four subregions outperforms the other interpolation approach in terms of log-likelihood.

(shown in gray in the north-west of both figures), kernel regression interpolation failed (for μ in the regional peaks-over-threshold model and for σ and ξ in the interpolation of the at-site estimates). In Figure 11b, two stations have values beyond the color scale (–4.5 and 6.9) (also shown in gray). The figures for the other relative loss functions have similar spatial patterns and are provided in the supporting information. The spatial average with standard errors in parentheses of the three relative loss functions is provided in Table 1.

Last, we performed the inference on 330 stations, i.e., with two stations (the closest ones to the cities of Montpellier and Valence, see Figure 7b) kept aside for validation. From the interpolated GP parameters from each of the three interpolation approaches, we computed return level curves (see equation (2)). To obtain the 95% confidence bands shown in Figure 12, we implemented nonparametric spatial block bootstrap as follows. Blocks of 3 days are randomly sampled from the original observations for all the stations simultaneously to preserve both the temporal and spatial dependence present in the observations. The size of the block was determined from the maximum number of consecutive excesses in the precipitation data. Other propositions to take into account the spatial dependence in the uncertainty estimation can be found in *Madsen and Rosbjerg* [1997b]; *Madsen et al*. [2002]; and *Van de Vyver* [2012].

5. Discussion

5.1. Synthetic Data Study

The synthetic data study in section 3 serves two purposes. First it allows to evaluate the procedure to select the number of hazard subregions from section 2.2.2 on data whose generative model is known.

For both the two subregion and the four subregion generative models (see equation (20) and equation (18), respectively), the number of subregions of the generative model is selected approximately in 50% of the bootstrap replicates when the negative log-likelihood loss function is used to assess performance see Figure 4. The picture is more mixed for the other two loss functions, see the supporting information. This percentage does not seem to follow a pattern, in particular it does not clearly increase with the number of sites or the sample

Table 1. Average (Standard Error) Loss Functions Relative to the Regional Model With Four Subregions[a]

	Neg Log-Like	Sum-of-Squares	Anderson-Darling
Single region	**0.1160 (0.04148)**	**784.5 (336.8)**	**0.06201 (0.02805)**
Kernel regression	0.03704 (0.05053)	**641.7 (286.6)**	0.009116 (0.04415)

[a]Positive values means that the regional model with four subregions outperforms the other interpolation approach, results in boldfaced are approximately significant at 95%.

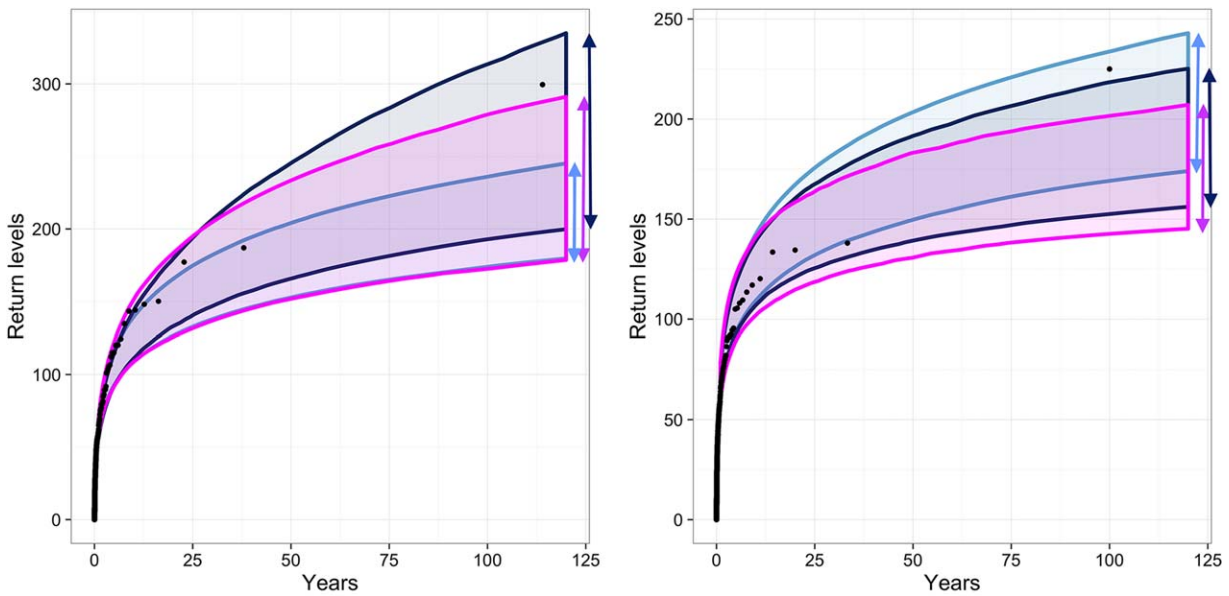

Figure 12. Return level curves from the regional peaks-over-threshold model with a single region (light blue), with four hazard subregions (dark blue) and the kernel regression interpolation of at-site estimates (magenta). The points represent the empirical return levels.

size as we might expect. There is a tendency to select more rather than less subregions than present in the generative model, especially as the number of sites and the sample size get larger. Indeed, the partition with a single region (in red in the figures) is rarely selected.

The focus of the procedure is only on the selection of the number of subregions, not on their identification. As the partitioning or *E-Step* is fairly stable, i.e., a partition of a given size will always define very comparable subregions, the selection of the number of subregions also determines their shape. On the other hand, the partition is built so that within each subregion, the hazard level of the sites is as similar as possible. Thus, if within some subregions the hazard level is too variable, the only possibility to improve the fit of the model is to increase the number of subregions, not to change their shape. In contrast, in regional frequency analysis, subregions are conventionally identified as geographically coherent areas with similar climatic and physical features. Such subregions are often found to be heterogeneous in terms of extremal behavior. Statistical tests are used to assess homogeneity and the subregions are reshaped until they pass the tests [*Madsen and Rosbjerg*, 1997b; *Castellarin et al.*, 2001].

The second purpose of the synthetic data study is to evaluate the inference and interpolation strategy of the regional peaks-over-threshold model, see section 2.2.1. The 95% confidence bands for the shape and scale parameter become narrower as the number of sites and the sample size increases, see Figures 5 and 6 and the supporting information. The width of the 95% confidence bands of the scale parameter is comparable for both generative models. In contrast, with the four subregion generative model, the confidence bands of the shape parameter tend to overlap more between subregions than with the two subregion generative model.

5.2. Daily Precipitation Data Application

The threshold that defines the excesses (the 98% quantile of the precipitation intensities) takes its larger values, about 100 mm, along the Cevennes mountain range, see Figure 8a. As the Cevennes acts as a barrier, the threshold values drop quickly behind it. The spatial pattern of the average number of excesses per year is noticeably different, see Figure 8b. In section 2.1.1, it is defined as $N_{exc}=365.25\,\zeta_u=365.25\,\mathbb{P}(P>u|\boldsymbol{x})=365.25\,\mathbb{P}(P>u|P>0,\boldsymbol{x})\,\mathbb{P}(P>0|\boldsymbol{x})$, where P is the random variable representing daily precipitation (not the excesses, that are represented by Y). By construction, $\mathbb{P}(P>u|P>0,\boldsymbol{x})=0.02$. Therefore, the north-south pattern of N_{exc} is driven by $\mathbb{P}(P>0|\boldsymbol{x})$, the probability of occurrence of positive precipitation intensities.

The shape parameter estimate of the regional peaks-over-threshold model with a single region $\hat{\xi}_1 = 0.11$ is very close to the average of the estimates of the four subregions model which is of 0.12, see Figure 9a. Similarly, the shape parameter estimates of the model with four subregions can be seen as approximately the average of the estimates obtained with kernel regression within each subregion, see Figure 9. Most of the values (84%) in Figure 9b are in the $[0, 0.3]$ interval. Thus, the range of values of the shape parameter estimates is globally the same for the four subregions model and the interpolation with kernel regression. Also, in both models, the higher hazard subregion is located in the south (in pink) with shape parameter estimates close to 0.3. The region starts at the coast, goes up to the foothills of the Cevennes mountain range and is consistent with expert knowledge [*Delrieu et al*., 2005; *Braud et al*., 2014].

The borders of the subregions are relatively smooth in Figure 9a except in a few places such as near the southern east and west borders of the high hazard subregion. The amount of smoothness is controlled by k, the number of neighbors in the k-nearest neighbor rule, which was set to 5. This choice might have an impact on the cross-validation results (held-out sites might be classified into different hazard subregions with other values of k). We leave the selection of the number of neighbors for further studies.

The partition obtained with the regional peaks-over-threshold model with data-driven number of subregions has two subregions with relatively close shape parameter estimates in Figure 9a, $\hat{\xi}_1 = 0.01$ and $\hat{\xi}_2 = 0.07$ in green and orange, respectively. It seems likely that three instead of four hazard subregions would be sufficient. However, we considered only partitions of size 1, 2, 4, and 8. To assess the validity of a partition of size 3, we would have to tune the procedure of selection of the number of subregions to refine the search over more precise partition sizes. Given the results in the synthetic data study, it seems, on one hand, that determining a very precise partition size is not an easy task but on the other hand, a slightly higher number of subregions should not harm much the GP parameter estimates.

The differences between the scale parameter estimates from the three interpolation approaches in Figure 10 are generally small compared to the threshold values in Figure 8a. Indeed, 93% of the differences between the single region and the four subregion model are less than 5 mm in magnitude (see Figure 10a). For the kernel regression interpolation of at-site estimates, this percentage goes up to 97%, see Figure 10b. This is consistent with the findings in the synthetic data study that the scale parameter estimates is not very sensitive to the shape parameter estimate.

The spatial pattern of the cross-validation results for the negative log-likelihood loss function in Figure 11a shows that the single region model is outperformed by the four subregion model in the higher hazard area (shown in pink in Figure 9a). Similar patterns can be seen for the other two loss functions in the supporting information. In contrast, there is no clear spatial pattern for the performance of the kernel regression interpolation of the at-site estimates relative to the four subregion model, see Figure 11b. These conclusions are supported by Figure 12. In Figure 12a, for the city of Montpellier which sits in the higher hazard subregion, the 95% confidence bands of the return level curve of the single region model underestimate the empirical return levels. In Figure 12b, no similar underestimation occurs at the city of Valence which belongs to a low hazard region with shape parameter about 0.07, see Figure 9a. The 95% confidence bands of the return level curve of the kernel regression interpolation are somewhat lower than the empirical return levels and in particular, the larger empirical return level is outside the confidence bands in both cases. The spatial averages of the relative loss functions (with standard errors in parentheses) in Table 1 are all positive and most of them are significant at 95%, indicating that the four subregion model performed better relatively to the other two interpolation approaches.

6. Conclusions

In some regions such as the French Mediterranean, the distribution of heavy precipitation is known to follow specific spatial patterns. Since the shape parameter of the GEV or the GP distribution governs the behavior of the upper tail of the distribution where the extremes lie, it can be related directly to the hazard level. Therefore, when interpolating the distribution of extremes in such a region, we would be tempted to take into account the variability of the hazard level and to let the shape parameter vary in space. However, the shape parameter is difficult to estimate due to the high uncertainty of its estimators and for this reason, it is often assumed to be constant. In this work, we chose a middle ground and assume a moderately varying hazard level. More precisely, it is allowed to vary in a piecewise constant fashion.

The main contributions of this paper are the following. We formulated the regional peaks-over-threshold model as a conditional mixture model of GP distributions. Each component has a smoothly varying scale parameter together with a constant shape parameter and can be thought of as affecting a subregion with constant hazard level. We developed a two-step inference and interpolation strategy that is similar in spirits to the Expectation-Maximization algorithm. In the first step, the partition into subregions with constant hazard level is estimated i.e., each site is assigned to a mixture component. The second step consists in estimating the GP parameter for each subregion while the probability of belonging to each subregion follows from the partitioning of the first step. The two-step strategy relies on two probability weighted moments of the GP. The computations are fast and could be used to initialize maximum likelihood estimators. We proposed a classical cross-validation procedure with three different loss functions to assess performance in order to select the number of subregions.

Thanks to 18 synthetic data sets, we assess how the performance of the regional peaks-over-threshold model is affected by the number of subregions, the variability of the hazard level, the number of sites, and the sample size at each site. To this end, we designed two generative models with different types of partitions: one with two subregions with very different hazard levels and another one with four subregions and more similar hazard levels. We were able to evaluate the following two aspects: the ability of the cross-validation procedure to retrieve the number of subregions of the generative model and the performance of the GP estimators when the number of subregions is selected with cross validation and therefore prone to error.

The synthetic data study indicates that the proposed procedure to select the number of subregions is only partially successful at identifying the number of subregions of the generative model and tend to overestimate it, in particular for larger data sets. Note that conventional homogeneity test in regional frequency analysis also suffer from a lack of power [*Viglione et al*., 2007]. On the other hand, although the selected number of subregions does not tend to the generative model's, the accuracy of the GP parameter estimates is improved with larger data sets. As expected, hazard levels are more clearly determined (i.e., the confidence intervals of the shape parameter estimates overlap less) for synthetic data from the generative model with the two very distinct hazard subregions than with the four subregions with closer hazard levels.

In the French Mediterranean daily precipitation data application, the cross-validation procedure selected four subregions for the regional peaks-over-threshold model according to all three loss functions. We conducted a comparison with two other interpolation approaches. The first approach assumes that the shape parameter is constant in the region while the second approach let it varies smoothly as a function of covariates. The comparison of the out-of-sample performance of the three interpolation approaches leads us to conclude that the assumption of a constant shape parameter is not appropriate for this region. On the other hand, the gain in performance of the four subregion model relative to the smooth interpolation of the at-site estimates is not so clear. As shown in *Carreau et al*. [2013] for the block maxima approach, it is likely that the gain in performance of the regional model would be greater in applications in which the spatial variability of the variable of interest is high compared to the spatial density of the sites such as in sparsely gauged networks, in more arid climates or when considering subdaily precipitation. Potential advantages of the regional model are the partitioning into hazard subregions which could be useful for decision makers and shape parameter estimates that are less likely to take negative values thanks to the pooled sample.

One interesting perspective for this work is to exploit the formulation of the regional peaks-over-threshold model as a mixture of GPs. First, by making the assignment rules to make the partition *soft*, i.e., probabilistic, smooth transitions between subregions could be obtained. More precisely, the probability of belonging to a subregion would vary smoothly, being closer to 1 in the center and gradually decreasing when getting near the borders. Second, as mentioned previously, the mixture parameters could be estimated with an EM algorithm or by maximizing directly the log-likelihood starting from the parameter values provided by the two-step inference strategy developed in this work. Another perspective is to improve the procedure to select the number of subregions. In particular, we could work out a penalty term in order to limit the number of subregions when the differences in shape parameter estimates are low relative to their uncertainty. Last, the regional peaks-over-threshold model could be applied to a sparsely gauged network to better assess the gain in performance of the regional peaks-over-threshold model compared to a direct interpolation of the at-site GP parameter estimates. Most likely, x and y coordinates would not be informative enough and, in addition to altitude, covariates related to the orography could be of interest [*Benichou and*

Le Breton, 1987]. Besides, nonparametric methods, such as the k-nearest neighbor rule and kernel regression, might not perform so well in sparser networks and it might be more appropriate to seek parsimonious parametric models.

Appendix A: Alternative Inference Strategy for a Given Subregion

We present the inference strategy developed in *Naveau et al*. [2014] to estimate the GP parameters in a region with constant shape parameter and smoothly varying scale parameter. It is related to the so-called *M-Step* of the inference strategy in section 2.2.1.

It relies on a different scaled variable $Z \sim G(1,\xi)$. Let α denotes the ratio of the second and third PWMs of Z (replace $r = 1, 2$ and $\sigma(\boldsymbol{x})=1$ in equation (3)), i.e.,

$$\alpha = \frac{\mathbb{E}[Z\bar{G}(Z;1,\xi)]}{\mathbb{E}[Z\bar{G}(Z;1,\xi)^2]}. \tag{A1}$$

As before, let $\mathbb{E}[Y|\boldsymbol{x}]=\mu(\boldsymbol{x})=\sigma(\boldsymbol{x})/(1-\xi)$ be the first PWM of Y given $\boldsymbol{x}$. *Naveau et al*. [2014] expressed ξ and $\sigma(\boldsymbol{x})$ as functions of the aforementioned three PWMs:

$$\xi = \frac{9-4\alpha}{3-2\alpha} \quad \text{and} \quad \sigma(\boldsymbol{x})=\mu(\boldsymbol{x})(1-\xi). \tag{A2}$$

To infer ξ and $\sigma(\boldsymbol{x})$, estimates of α and $\mu(\boldsymbol{x})$ are required. Since $Z=Y(\boldsymbol{x})/\sigma(\boldsymbol{x})$ is not observable, *Naveau et al*. [2014] circumvent this issue by relying on the probability weighted moment of $Y(\boldsymbol{x})/\mu(\boldsymbol{x})$ to estimate α. The latter remains unchanged since it is a ratio and $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ differs by the factor $1-\xi$.

Acknowledgments
This work was partially supported by the Agence Nationale de la Recherche (French Research Agency) through its Blanc program with the projects Floodscale and DADA and through its JCJC program with the project Starmip. Part of this was has also been supported by LEFE-INSU-Multirisk, LEFE-INSU-Cerise, AMERISKA, and A2C2 projects. We are grateful to Météo-France for the precipitation data available at `https://publitheque.meteo.fr` and to Juliette Blanchet (IGE) for the preprocessing.

References

Abadir, K. M., and S. Lawford (2004), Optimal asymmetric kernels, *Econ. Lett.*, *83*(1), 61–68.
Anderson, T. W., and D. A. Darling (1954), A test of goodness of fit, *J. Am. Stat. Assoc.*, *49*(268), 765–769.
Arlot, S., and A. Celisse (2010), A survey of cross-validation procedures for model selection, *Stat. Surv.*, *4*, 40–79.
Balkema, A. A., and L. de Haan (1974), Residual life time at great age, *Ann. Probab.*, *2*(5), 792–804.
Bénichou, P., and O. Le Breton (1987), Prise en compte de la topographie pour la cartographie des champs pluviométriques statistiques, La Météorologie, 7e série, no. 19.
Bishop, C. M. (2011), *Pattern Recognition and Machine Learning, Information Science and Statistics*, Springer, New York.
Blanchet, J., and M. Lehning (2010), Mapping snow depth return levels: Smooth spatial modeling versus station interpolation, *Hydrol. Earth Syst. Sci.*, *14*(12), 2527–2544.
Borga, M., E. Anagnostou, G. Blöschl, and J.-D. Creutin (2011), Flash flood forecasting, warning and risk management: The HYDRATE project, *Environ. Sci. Policy*, *14*(7), 834–844, doi:10.1016/j.envsci.2011.05.017, adapting to Climate Change: Reducing Water-related Risks in Europe.
Braud, I et al. (2014), Multi-scale hydrometeorological observation and modelling for flash flood understanding, *Hydrol. Earth Syst. Sci.*, *18*(9), 3733–3761, doi:10.5194/hess-18-3733-2014.
Burn, D. (1990), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resour. Res.*, *26*(10), 2257–2265.
Carreau, J., and S. Girard (2011), Spatial extreme quantile estimation using a weighted log-likelihood approach, *J. Soc. Française Stat.*, *152*(3), 66–82.
Carreau, J., L. Neppel, P. Arnaud, and P. Cantet (2013), Extreme rainfall analysis at ungauged sites in the South of France: Comparison of three approaches, *J. Soc. Française Stat.*, *154*(2), 119–138.
Castellarin, A., D. Burn, and A. Brath (2001), Assessing the effectiveness of hydrological similarity measures for flood frequency analysis, *J. Hydrol.*, *241*(3), 270–285.
Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, *Springer Series in Statistics*, Springer, London, U. K.
Cooley, D., D. Nychka, and P. Naveau (2007), Bayesian spatial modeling of extreme precipitation return levels, *J. Am. Stat. Assoc.*, *102*(479), 824–840.
Delrieu, G., et al. (2005), The catastrophic flash-flood event of 8–9 September 2002 in the Gard region, France: A first case study for the Cévennes-Vivarais Mediterranean Hydrometeorological Observatory, *J. Hydrometeorol.*, *6*(1), 34–52.
Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Stat. Soc. Ser. B*, *39*(1), 1–38.
Diebolt, J., A. Guillou, and I. Rached (2007), Approximation of the distribution of excesses through a generalized probability-weighted moments method, *J. Stat. Plann. Inference*, *137*(3), 841–857.
Epanechnikov, V. A. (1969), Non-parametric estimation of a multivariate probability density, *Theor. Probab. Appl.*, *14*(1), 153–158.
Evin, G., J. Blanchet, E. Paquet, F. Garavaglia, and D. Penot (2016), A regional model for extreme rainfall based on weather patterns subsampling, *J. Hydrol.*, *541*, 1185–1198.
Fisher, R., and L. H. C. Tippett (1928), Limiting forms of the frequency distribution of the largest and smallest member of a sample, in Cambridge Philosophical Society, vol. 24, pp. 180–190, Cambridge Univ. Press, Cambridge, U. K.
Furrer, R., and P. Naveau (2007), Probability weighted moments properties for small samples, *Stat. Probab. Lett.*, *77*(2), 190–195.
Gnedenko, B. (1943), Sur la distribution limite du terme maximum d'une serie aleatoire, *Ann. Math.*, *44*, 423–453.
Hayfield, T., and J. Racine (2008), Nonparametric econometrics: The np package, *J. Stat. Software*, *27*(5), 1–32.

Hosking, J. R. M., and J. R. Wallis (2005), *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge Univ. Press, Cambridge, U. K.
Li, Q., and J. Racine (2004), Cross-validated local linear nonparametric regression, *Stat. Sin.*, *14*(2), 485–512.
Madsen, H., and D. Rosbjerg (1997a), The partial duration series method in regional index-flood modeling, *Water Resour. Res.*, *33*(4), 737–746.
Madsen, H., and D. Rosbjerg (1997b), Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling, *Water Resour. Res.*, *33*(4), 771–781.
Madsen, H., P. S. Mikkelsen, D. Rosbjerg, and P. Harremoës (2002), Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics, *Water Resour. Res.*, *38*(11), 1239, *doi*:10.1029/2001WR001125.
Nadaraya, E. A. (1964), On estimating regression, *Theor. Probab. Appl.*, *9*(1), 141–142.
Naveau, P., A. Toreti, I. Smith, and E. Xoplaki (2014), A fast nonparametric spatiotemporal regression scheme for generalized Pareto distributed heavy precipitation, *Water Resour. Res.*, *50*, 4011–4017, doi:10.1002/2014WR015431.
Pickands, J. (1975), Statistical inference using extreme order statistics, *Ann. Stat.*, *3*, 119–131.
R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna, Austria.
Renard, B. (2011), A Bayesian hierarchical approach to regional frequency analysis, *Water Resour. Res.*, *47*, W11513, doi:10.1029/2010WR010089.
Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge Univ. Press.
Roth, M., T. Buishand, G. Jongbloed, A. Klein Tank, and J. van Zanten (2012), A regional peaks-over-threshold model in a nonstationary climate, *Water Resour. Res.*, *48*, W11533, doi:10.1029/2012WR012214.
Sang, H., and A. E. Gelfand (2009), Hierarchical modeling for extreme values observed over space and time, *Environ. Ecol. Stat.*, *16*(3), 407–426.
Van de Vyver, H. (2012), Spatial regression models for extreme precipitation in Belgium, *Water Resour. Res.*, *48*, W09549, doi:10.1029/2011WR011707.
Viglione, A., F. Laio, and P. Claps (2007), A comparison of homogeneity tests for regional frequency analysis, *Water Resour. Res.*, *43*, W03428, doi:10.1029/2006WR005095.
Watson, G. S. (1964), Smooth regression analysis, *Sankhyā Ser. A*, *26*, 359–372.