Smooth Copula-based Generalized Extreme Value model and Spatial Interpolation for Sparse Extreme Rainfall in Central Eastern Canada

Fatima Palacios-Rodriguez, Elena Di Bernardino, Mélina Mailhot

Noname manuscript No.
(will be inserted by the editor)

Smooth Copula-based Generalized Extreme Value model and Spatial Interpolation for Sparse Extreme Rainfall in Central Eastern Canada

Fatima Palacios-Rodriguez · Elena Di Bernardino · Mélina Mailhot

Received: date / Accepted: date

Abstract This paper proposes a smooth copula-based Generalized Extreme Value (GEV) model to map and predict extreme rainfall in central eastern Canada. Furthermore, we provide a comparison with different classical interpolation-based approaches. The considered data represents a station network particularly spatially sparse. Furthermore, one observes several missing values and non-concomitant record periods at different stations. We compare the classical GEV parameter interpolation approaches with our smooth GEV modeling approach, in which the parameters are modeled as smooth functions in space through the use of spatial covariates and by using copula-clustering techniques recently introduced in the literature.

Keywords Copula-based Clustering · Dependence Models · Extreme Value Theory · Hydrology · Spatial Interpolation

1 Introduction

Motivation. Heavy rainfall can have disastrous consequences on health and well-being of communities, buildings, infrastructures, transportation systems and public safety. In practice, the interest is, amongst others, from a national safety, risk management and insurance perspectives. For researchers, as Extreme Value Theory (EVT) is an area with great recent innovative results, precipitation levels are very interesting. Throughout the years, flood events became important, both from practical and theoretical perspectives. For example, in Canada, it is recent that insurance companies offer flood protections. Before 2013, homeowners would rely on the Disaster Financial Assistance program offered by the federal, provincial and territorial governments. Now, insurance products are available, and several resources are dedicated to improving flood mapping and mitigation efforts. Models are now adapted with today's knowledge on extreme events, in order to assess risks depending on precipitation appropriately. In other words, it is desirable to set aside safety capital according to a well-evaluated risk.

Extensive literature exists on the spatial mapping or spatial interpolation of extreme rainfall and the approaches are essentially divided in two groups. The first one is mainly composed of techniques to spatially interpolate the station estimates from marginal Generalised Extreme Value distributions in order to provided return level maps. This classical approach is frequently used, *e.g.* in Beguería and

Fatima Palacios-Rodriguez
Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Calle Tarfia sin número, 41012 Seville, Spain, E-mail: fpalacios2@us.es

Elena Di Bernardino
Laboratoire J.A. Dieudonné, UMR CNRS 7351, Université Côte d'Azur, Nice, France. Tel.: (33) (0) 4 89 15 04 95 E-mail: elenadb@unice.fr

Mélina Mailhot
Department of Mathematics and Statistics, Concordia University, 1455 De Maisonneuve Blvd. W., Montréal (QC) Canada H3G 1M8. E-mail: melina.mailhot@concordia.ca

Vicente-Serrano (2006), Kohnová et al. (2009) and Li et al. (2010). A comparison of different naive interpolation methods (inverse distance weighting, nearest neighbor and kriging) for mapping extreme precipitation in central Slovakia can be found in Szolgay et al. (2009). Similar study is performed in Blanchet and Lehning (2010), or in Das et al. (2020), for mapping snow depth return levels. In Hwang et al. (2012), a two step procedure is proposed where a regression and hydrological models are used to interpolate precipitation. Pointwise return levels, in the Cévennes-Vivarais region (southern part of France), based on nearest neighbor estimators, are obtained in Gardes and Girard (2010). A spatial model is proposed in Yoon et al. (2015), assuming no spatial dependency between the stations.

The second group in the extreme return level maps literature is based on the direct estimation of the spatial extremal distribution, without requiring any interpolation, (see *e.g.*, Wi et al. (2016) and Lomba and Alves (2020)). Spatial extreme distributions have received a lot of attention in recent years and they represent a well-founded approach which are theoretically preferred to any interpolation method. A bayesian procedure is presented in Perreault et al. (1999), for eastern Canada and US data, and Asong et al. (2015) deals with the Canadian Prairie region. Several techniques have been developed for the direct estimation of spatial extreme distributions, which involve, among others, extreme-value copulas (see *e.g.*, Joe (1994) and Saad et al. (2015)), max-stable processes (see *e.g.*, Schlather (2002), Padoan et al. (2010)) and Bayesian hierarchical models (see *e.g.*, Cooley et al. (2007)). Notice that the property of max-stablity is classically included in several papers on spatial interpolation of extremes (see *e.g.*, Blanchet and Davison (2011), for extreme snow depth models).
This article focuses on constructing the spatial mapping of maximum precipitation, using the EVT framework for 24h duration rainfall annual maxima in Central Eastern Canada. Despite the fact that several authors have brought efforts in order to provide spatial extreme models for precipitation, the considered dataset used in this article presents at least two interesting aspects which need to be addressed carefully.

Challenging characteristics of the considered dataset. Firstly, we focus here on modeling extreme rainfall for 116 recording stations located in the province of Quebec, Nova Scotia, New Foundland, New Brunswick and Prince Edward Island. This is a vast region, with a small proportion of recording stations and literature is quite scarce for this specific area. Notice that the size of the meteorological stations network in Canada is a major well-known issue already raised by the Canadian Standards Association. Khedhaouiria et al. (2020) consider another Canadian dataset which is more concentrated in a southern Canadian region.
Even given the scarce aspect of the dataset, the obtained results will show a robust performance in the case of considered smooth Generalized Extreme Value (GEV) distribution fitting methods. Notice that the same dataset has been recently analyzed in Perreault et al. (2019) where an interpolation Bayesian hierarchical model is proposed with the spatial effect modeled via Gaussian Markov random fields. Then, our results can be compared with those obtained in Perreault et al. (2019) as suggested in Section 5.

Secondly, the considered dataset presents several missing values and non-concomitant record periods of different stations. The interested reader is referred to Figure 1 for a graphical illustration of this crucial point. It implies that when one aims to model the joint behaviour of the extreme rainfall in the considered area, the estimated marginal distribution uses the complete series for that station, but the copula function representing the dependence (see, *e.g.*, Nelsen (1999)) is only based on the time period where all series were recorded simultaneously. Here, we are facing the problem of estimating parametric multivariate models when unequal amounts of data are available on each variable (see *e.g.*, Patton (2006)). To overcome this issue, we consider the *hybrid copula*, *i.e.*, the extension of the empirical copula obtained by combining an estimator of a multivariate cumulative distribution function with estimators of the marginal cumulative distribution functions for marginal estimators that are not necessarily equal to the margins of the joint estimator (see *e.g.*, Segers (2015)).

Contributions of the present work. The contributions of this article are twofold. *(i)* We propose a smooth GEV model, mixing sophisticated response surfaces for the GEV parameters' models, with a flexible joint dependence framework *via* a hierarchical copula-based model, which takes into account dependence

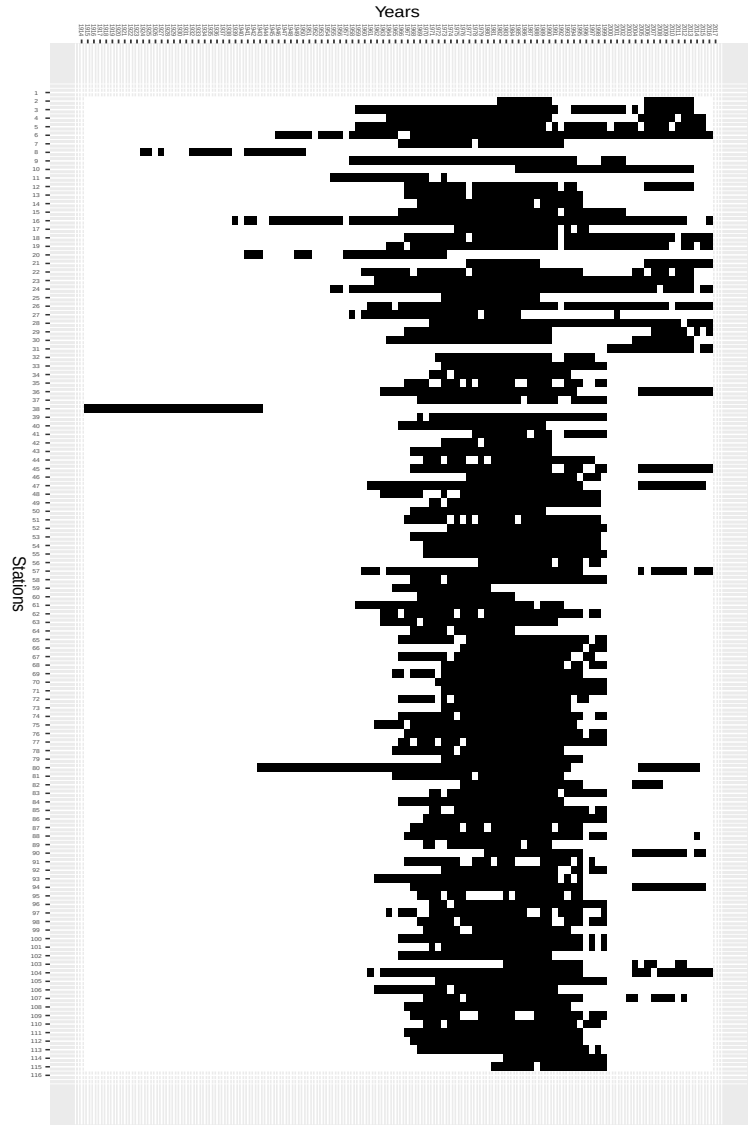

Fig. 1 White cells represent missing data and black ones observed extreme rainfall registered in 116 stations in Center Eastern Canada from 1914 to 2017.

structure between the recording stations. The hierarchical copula structure is detected via a clustering algorithm implemented with an adapted version of the copula-based dissimilarity measure recently proposed in Disegna et al. (2017). The considered dissimilarity measure is consistently estimated by using the previously discussed hybrid copula. *(ii)* We propose an analysis, comparing classical spatial interpolation of individual GEV distributions: polynomial and spline-based regression models, inverse distance weighted and universal kriging models. This comparison is crucial to show the difference between practical commonly implemented routines and the sophisticated approach proposed in this paper, in terms of obtained return level maps (see Figures 9-10).

Outline of the paper. The article is organized as follows. Section 2 presents the considered precipitation dataset. In Section 3, return level maps are obtained through smooth spatial GEV models by using our clustering hierarchical copula-based model with associated copula-based dissimilarity measure. In Section 4 we build return level maps via classical spatial interpolation methods of individual GEV distributions. Section 5 is devoted to evaluating the performance of the proposed methods. Some details on L-moments for GEV parameters are postponed in Appendix A.

2 Considered Center Eastern Canada dataset

We consider rainfall data measured in millimeters (mm), adjusted for snow/ice, registered in 116 stations in Center Eastern Canada for a duration of 24h. The stations are managed by Environment and Climate Change Canada (ECCC) and verify the quality standards imposed by the World Meteorological Organization. Annual maxima precipitation for a 24-hour duration are recorded from 1914 to 2017. Each station possesses precipitation measures for a minimum of 16 years in the specified time range. A specific characteristic of the considered rainfall dataset is the sparsity of the recorded data, as depicted in Figure 1, illustrating the proportion of missing data and concomitant observations across years. The elevation of the studied area covers a wide range given between 5m and 672m above sea level. These annual maxima are publicly available at `climate.weather.gc.ca/prods_servs/engineering_e.html`.

We introduce the following mathematical notation, crucial in the following for the description of the considered models. Let $x_i = \left(x_{(i,1)}, \ldots, x_{(i,T)}\right)$ be the annual rainfall maxima time series of the ith station, for $i \in I := \{1, \ldots, n\}$, observed for T years. Then, our rainfall data can be represented as a sparse $(n \times T)-$matrix, $\mathbf{X}$ (refer to left panel of Figure 2), where $n = 116$ (the number of stations) and $T = 104$ (the length of the considered whole time window). The spatial location of the considered stations is shown in Figure 2 (right panel)[1].

$$\mathbf{X} = \begin{pmatrix} x_{(1,1)} & \cdots & x_{(1,t)} & \cdots & x_{(1,T)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{(i,1)} & \cdots & x_{(i,t)} & \cdots & x_{(i,T)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{(n,1)} & \cdots & x_{(n,t)} & \cdots & x_{(n,T)} \end{pmatrix}$$

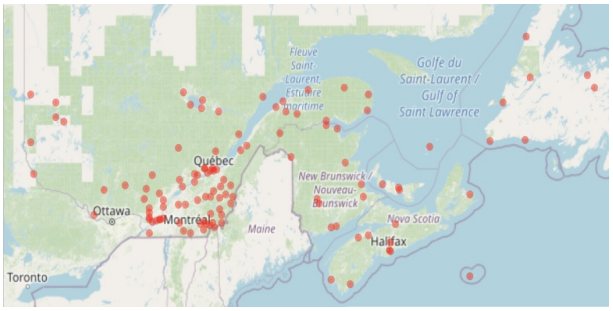

Fig. 2 Left: Considered data $(n \times T)-$matrix. Right: Locations of considered 116 stations in Central Eastern Canada.

As suggested by Figure 1, we introduce the indicator variable

$$I_t^i = \begin{cases} 1, & \text{if } x_{(i,t)} \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Furthermore, we denote by I_f (*resp.* I_v) the non-empty set of indexes of stations used for the fitting (*resp.* validation) procedure. Let $n_f = \text{card}\,(I_f)$ and $n_v = \text{card}\,(I_v)$. Obviously, $I_f \cap I_v = \emptyset$ and $n_f + n_v = n$. In the present paper, we consider $n_f = 95$ and $n_v = 21$. Notice that the arbitrary choice of the n_v validation stations can be influent in the final performance of the proposed models. For this reason, in order to test the robustness of the investigated models, we decide to choose randomly 200 combinations of $n_f = 95$ and $n_v = 21$ stations to perform our study. One of the 200 considered random combinations between fitting and validation stations is displayed in Figure 3.

In the following we will first consider spatial smooth GEV models to derive return level maps for every location $s \in S$, where S represents the overall surface being interpolated. The involved parameters are modeled *via* smooth functions in space by including some significant covariates of the models (see Section 3). Then, we compare the obtained results with classical spatial interpolation methods of individual GEV distributions based on the L-moments estimators (see Section 4 and Appendix A).

[1] Remark that all maps in this paper have been obtained with package `leaftet` in `R`.

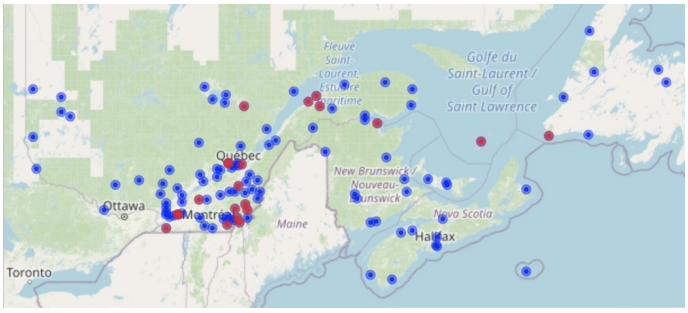

Fig. 3 Spatial locations of one of the 200 considered combinations between fitting stations I_f (blue points) and validation stations I_v (red points) with $n_f = 95$ and $n_v = 21$.

3 Return level maps through spatially smooth copula-based GEV model

3.1 Univariate EVT via block maxima approach

Since $\mathbf{X}$ is defined by annual maxima of precipitation, we focus on the EVT block-maxima approach (see, *e.g.*, Coles (2001) and Ferreira and de Haan (2015)). Let $x_{(i,t)}$ be the annual maxima at the ith station for year t. Then, we write $x_{(i,t)} = \max_j\{z^j_{(i,t)}\}$, where $z^j_{(i,t)}$ represents the precipitation at the ith station the jth day of the considered year t. EVT requires independence or short-range dependence (Leadbetter et al. (1983)). We observe through an additional analysis, near-independence in our precipitation time-series for every considered year and station. Therefore, we can model $x_{(i,t)}$ by means of a GEV distribution with parameters $\Lambda = (\mu, \xi, \sigma)$, *i.e.*,

$$G(x;\Lambda) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The shape parameter $\xi \in \mathbb{R}$ describing the tail behaviour of the distribution is called the extreme value index, $\mu \in \mathbb{R}$ is the location parameter and $\sigma > 0$ the scale parameter. Now, we introduce the return level for the GEV distribution. The return level $q(p;\Lambda)$ associated with the return period $1/p$ $(0 < p \leq 1)$ is the $(1-p)$th quantile of the GEV distribution in (2), *i.e.*, it is expected to be exceeded on average once every $1/p$ years:

$$q(p;\Lambda) = \begin{cases} \mu - \frac{\sigma}{\xi}[1-\{-\log(1-p)\}^{-\xi}], & \xi \neq 0, \\ \mu - \sigma\log\{-\log(1-p)\}, & \xi = 0. \end{cases} \tag{3}$$

Considered covariates. In the following, we consider three classical geographical coordinates as covariates: longitude, latitude and elevation, which are obtained using a digital elevation model. Furthermore, in our analysis, we include the mean precipitation averaged over the 34-year period 1981-2014 of the Canadian Regional Climate Model (CRCM5) driven by the Era-Interim reanalysis. In Figure 4 a graphical representation for this covariate for the considered area is available. More information on this climate reconstruction can be found, for instance, in Bresson et al. (2017). This variable is available on a regular lattice covering the northeastern part of North America through 90000 grid cells, each of which corresponds to an area of 12×12 km^2 in size.

3.2 Smooth models for GEV parameters

In this section, we consider the estimation of a spatially smooth GEV distribution with the joint use of all stations. We aim to model the GEV parameters $\Lambda(s)$ for $s \in S$ from the data as smooth functions in space. Let ζ be one of three GEV parameters and $\tilde{\zeta}$ be the associated interpolated value.
Let consider the following general regression model associated to the function F

$$\zeta(s) = F(y_s^{(1)}, , \ldots, y_s^{(r)}) + \epsilon_s, \tag{4}$$

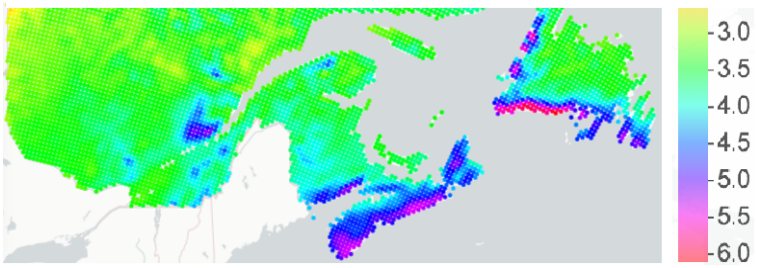

Fig. 4 Mean precipitation at every 12 × 12 km^2 grid cell of the Canadian Regional Climate Model driven by the Era-Interim reanalysis, averaged over the 34-year period 1981-2014.

where ϵ_s is an error term satisfying the ordinary least squares hypothesis. In what follows, the GEV parameter ζ (either μ, ξ or σ) at location s is modeled by (4) without the stochastic (Gaussian) contribution represented by ϵ_s.

Polynomial regression model. Firstly, we consider F in Equation (4) as linear with respect to each covariate. In the present study, we consider covariates $y_s^{(j)}$, for $j \in \{1, \ldots, r\}$, as polynomials of longitude, latitude, altitude and mean precipitation with a maximum polynomial degree of 3 and we take all possible combinations between these covariates with a maximum polynomial interaction degree of 3. This provides a commonly used polynomial regression model for our predictive analysis, *i.e.*, the interpolated value at location s is written as

$$\tilde{\zeta}(s) = \tilde{\beta}_0 + \tilde{\beta}_1\, y_s^{(1)} + \ldots + \tilde{\beta}_r\, y_s^{(r)}, \tag{5}$$

with the previously described covariates $y_s^{(j)}$ and where $\tilde{\beta}_0, \ldots, \tilde{\beta}_r$ are the classical least square estimates of regression parameters (see, *e.g.,* Rencher and Christensen (2012)).

Spline-based regression model. In order to generalize Equation (5), we can model the relation between the covariates with a smooth non-linear function F in (4). To avoid having to deal with the estimation of a large number of parameters, we consider here a partial linearity by the following generalized additive model:

$$\zeta(s) = \beta_0 + \beta_1 y_s^{(1)} + \ldots + \beta_q y_s^{(q)} + F(y_s^{(q+1)}, \ldots, y_s^{(r)}) + \epsilon_s, \tag{6}$$

where ϵ_s is an error term and F is a penalized spline (see Marx and Eilers (1998)). Therefore, the interpolated value at location s is given by

$$\tilde{\zeta}(s) = \tilde{\beta}_0 + \tilde{\beta}_1 y_s^{(1)} + \ldots + \tilde{\beta}_q y_s^{(q)} + \tilde{F}(y_s^{(q+1)}, \ldots, y_s^{(r)}), \tag{7}$$

where $\tilde{F}$ is the estimated penalized spline in Equation (6) obtained by minimizing the sum of squared errors subject to constraints on its parameters, to avoid over-fitting (see, *e.g.*, Section 3 in Ruppert et al. (2003)). A simplified similar framework is considered for instance in Padoan et al. (2010) where F is a linear regression model using only latitude and elevation as covariates. Conversely, here we take into consideration more complex covariate models provided by polynomial regression as in (5) or spline-based regression as in (7) with longitude, latitude, altitude and mean precipitation as covariates. In order to limit the number of possible smooth GEV parameter models, we select here a total of 96 possible models, gathered in Table 1.

In Section 3.3 below, we mix the sophisticated response surfaces for modeling the GEV parameters gathered in Table 1 to a flexible joint dependence framework via a hierarchical copula-based model. Although the assumption of spatial independence between the stations is very unlikely in real life, it can be found in several papers and can provide satisfying results if we fix all our interest in marginal distributions. Nevertheless, the aim of Section 3.3 will be to relax in a tractable way this hypothesis to build a more realistic dependent setting.

Models for μ and ξ		Models for σ	
Chosen model	Covariates	Chosen model	Covariates
Best polynomial regression model	3 geographical coordinates	Best polynomial regression model	3 geographical coordinates
Best polynomial regression model	3 geographical coordinates and mean precipitation	Best polynomial regression model	3 geographical coordinates and mean precipitation
		Best polynomial regression model	3 geographical coordinates and μ parameter
i.e. 2 **polynomial regression models for μ and for ξ**		*i.e.* 3 **polynomial regression models for σ**	
Best spline-based regression model	3 geographical coordinates	Best spline-based regression model	3 geographical coordinates
Best spline-based regression model	3 geographical coordinates and mean precipitation	Best spline-based regression model	3 geographical coordinates and mean precipitation
		Best spline-based regression model	3 geographical coordinates and μ parameter
i.e. 2 **spline-based regression models for μ and for ξ**		*i.e.* 3 **spline-based regression models for σ**	

Table 1 Selected GEV parameter models from Equations (5) and (7) with related considered covariates.

3.3 Log-likelihood for the hierarchical copula-based model

In order to estimate the parameter models of the GEV presented in Table 1, we apply a log-likelihood approach which requires the joint distribution of annual maximum precipitation of the considered fitting stations I_f. To this end, we focus here in multivariate hierarchical copula models, that is, models that are able to capture different dependencies between and within different groups of random variables *via* dependence copula functions. One such class of models is based on nested Archimedean copulas (see *e.g.*, Hofert and Pham (2013)). A (partially) nested Archimedean copula C with two nesting levels and K child copulas (or groups), is given by

$$C(\mathbf{u}) = C_0(C_1(\mathbf{u}_1), \ldots, C_K(\mathbf{u}_K)), \quad \mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_K)^t, \tag{8}$$

where K denotes the dimension of C_0 (*i.e.*, the number of clusters) and each copula C_k is Archimedean with a completely monotone generator ψ_k, for $k \in \{0, \ldots, K\}$ (see, *e.g.*, Nelsen (1999)). In the following, for the sake of simplicity, we consider C_0 as the independent $K-$dimensional copula, *i.e.*, $C_0(v_1, \ldots, v_K) = \prod_{i=1}^{K} v_i$, with $v_i \in [0, 1]$.

Definition 1 (Hierarchical copula log-likelihood) *Denote by $G_i\big(\cdot; \Lambda(s_i)\big)$, the GEV distribution in* (2) *with GEV smooth surface parameters $\Lambda(s_i)$ as in Table 1 associated to the ith station and $g_i\big(\cdot; \Lambda(s_i)\big)$ its density, for $i \in I_f$. Let I_t^i as in* (1)*. Also, let $k \in \{1, \ldots, K\}$ and $I_k = \{i_1^{(k)}, \ldots, i_{d_k}^{(k)}\}$, where $d_k = card(I_k)$, be the set of station indices belonging to the kth cluster, such that $\cup_{k=1}^{K} I_k = I_f$ and $I_k \cap I_{k'} = \emptyset$, $\forall\, k \neq k'$. Denote by c_{θ_k}, the Archimedean copula density for the k^{th} cluster related to model in* (8)*. Let*

$$\mathcal{L}^{\perp}(\Lambda) := \sum_{i \in I_f} \sum_{\substack{t=1 \\ \text{s.t.} I_t^i = 1}}^{T} \ln\big\{g(x_{(i,t)}; \Lambda(s_i))\big\}. \tag{9}$$

Then, we introduce the log-likelihood associated to the hierarchical copula model in (8)

$$\mathcal{L}^{C}(\Lambda) = \sum_{k=1}^{K} \sum_{\substack{t=1 \\ \text{s.t. } I_t^{i_1^{(k)}} = \ldots = I_t^{i_{d_k}^{(k)}} = 1}}^{T} \ln\Big\{c_{\theta_k}\Big(G_{i_1^{(k)}}\big(x_{(i_1^{(k)}, t)}; \Lambda(s_{i_1^{(k)}})\big), \ldots, G_{i_{d_k}^{(k)}}\big(x_{(i_{d_k}^{(k)}, t)}; \Lambda(s_{i_{d_k}^{(k)}})\big)\Big)\Big\} + \mathcal{L}^{\perp}(\mu, \xi, \sigma). \tag{10}$$

Obviously, in the independence setting $\mathcal{L}^{C}(\Lambda)$ in (10) reduces to $\mathcal{L}^{\perp}(\Lambda)$ in (9).

3.4 Adapted copula-based clustering method

Partitioning Around Medoids (PAM) is a well recognized technique to create clusters with a good partitioning using medoids for a given number of clusters K (see, *e.g.*, Kaufman and Rousseeuw (1987) and Kaufman and Rousseeuw (1990), Chapter 2). The PAM algorithm is based on the search for K representative objects or medoids among the observations of the dataset. After finding a set of K medoids, clusters are constructed by assigning each observation to the *nearest* medoid. Next, each selected medoid object x_k and each non-medoid data point x_i are swapped and the objective function is computed. The objective function used is the sum of an appropriate dissimilarity measure $d_{ik}(x_i, x_k)$ computed between the time series of the ith station and the time series of the kth medoid (Reynolds et al. (2006), Schubert and Rousseeuw (2019)). The objective is to improve the quality of the clustering by exchanging selected objects (medoids) and non-selected objects. If the objective function can be reduced by interchanging a selected object with an unselected object, then the swap is carried out. This is continued until the objective function can no longer be decreased.

In the following we will run the PAM algorithm by using an adapted version of the *copula-based dissimilarity measure* recently introduced by Disegna et al. (2017) to detect clusters between spatially near and dependent stations.

Using the latitude and longitude covariates, we can construct additional information on stations, constituted of an $(n_f \times n_f)$ data matrix S, whose generic entry s_{ij} can be interpreted as the *spatial distance* between the ith and jth stations and

$$\widetilde{s}_{ij} = s_{ij}/(\max_{i,j=1,\dots,n_f} s_{ij}), \tag{11}$$

the normalised spatial distance between the ith and jth stations.
Notice that the sparse behaviour of our $(n \times T)$-data matrix $\mathbf{X}$ requires some crucial adaptation of classical clustering copula methods. To this end, we now introduce the notion of the bivariate hybrid empirical copula (see *e.g.*, Segers (2015)).

Definition 2 (Hybrid empirical copula) *Let i and j be fixed, with $i, j \in \{1, \dots, n_f\}$. Consider the $2 \times T$ matrix composed of $(x_{(i,t)}, x_{(j,t)})^\top_{\{t=1,\dots,T\}}$, where $\top$ represents the transpose operator. In each column, one or both entries may be missing. Formally, our observations consist of a sample of independent, identically distributed quadruples*

$$(I_t^i, I_t^j, I_t^i x_{(i,t)}, I_t^j x_{(j,t)}), \text{ for } t \in \{1, \dots, T\}.$$

Then, the hybrid empirical copula is defined

$$\widehat{C}^{ij}(u, v) = \widehat{H}\left(G^{\leftarrow}(u, \widehat{\Lambda}^i_{LM}), G^{\leftarrow}(v, \widehat{\Lambda}^j_{LM})\right), \text{ for } (u, v) \in [0, 1]^2, \tag{12}$$

where

$$\widehat{H}_T(x, y) = \frac{\sum_{t=1}^T \mathbb{1}\{x_{(i,t)} \le x,\, x_{(j,t)} \le y,\; I_t^i = I_t^j = 1\}}{\sum_{t=1}^T \mathbb{1}\{\; I_t^i = I_t^j = 1\}}. \tag{13}$$

The hybrid empirical copula in (12)-(13) is similar to the classical empirical copula process, but now the asymptotic variances and covariances are to be multiplied by the reciprocals of the observation probabilities $\mathbb{P}[I_t^i = 1]$, $\mathbb{P}[I_i^j = 1]$ and $\mathbb{P}[I_t^i = I_t^j = 1]$. Details are given in Segers (2015).
Then, the adapted empirical version of the copula-based dissimilarity measure in Disegna et al. (2017) can be defined as follows.

Definition 3 (Empirical copula-based dissimilarity measure) *Let define*

$$\widehat{d}_{ij} = f(\| \beta(M - \widehat{C}_{ij}) + \widetilde{s}_{ij}\,(1 - \beta)(M - W) \|), \tag{14}$$

where

- $\widetilde{s}_{ij}$ *is the normalised spatial distance in* (11)*;*

- *M is the Fréchet upper-bound copula,* i.e., $M(u,v) = \min(u,v)$*;*
- *W is the Fréchet lower-bound copula,* i.e., $W(u,v) = \max(u+v-1,0)$*;*
- *$\beta \in [0,1]$ is the tuning parameter which reflects the prior belief of the decision maker about the desired influence of the spatial component on the clustering procedure;*
- *$\widehat{C}^{ij}$ is the hybrid copula defined as* (12)-(13)*;*
- *$\|\cdot\|$ is a suitable norm in the copula space and f is an increasing and continuous real-valued function with $f(0) = 0$.*

Notice that the considered copula-based dissimilarity measure in (14) can be formalised as a suitable function of the hybrid empirical copula $\widehat{C}^{ij}$ (expressing the dependence between the ith and jth stations) and the spatial information s_{ij}. The weight of this convex combination is expressed by the magnitude of the β parameter. In Algorithm 1, we detail the steps to estimate the copula-based dissimilarity measure $\widehat{d}_{ij}$ in (14) for our sparse dataset.

In Figure 5, we display the absolute value of the logarithm scale for the dissimilarity measure $\widehat{d}_{ij}$ in (14) obtained *via* Algorithm 1 with $\beta = 0.2$ for fitting stations of one of the 200 combinations of the fitting stations (the same displayed in Figure 3). In addition, in Figure 6, we fix three fitting stations (black dots) and we display the estimated dissimilarity measure in (14) of these stations with respect to all others fitting stations in the considered combination. Unsurprisingly, one can observe that the estimated dissimilarity measure takes the smallest values in the geographical neighborhood of the considered fixed station. Moreover, the dissimilarity measure does not consider only spatial distance but also the copula dependence structure between involved stations.

Algorithm 1 Proposed implementation of a copula-based dissimilarity measure for sparse data

(Step 1) Estimate $\widehat{\Lambda}^i_{LM} = (\widehat{\mu}^i_{LM}, \widehat{\sigma}^i_{LM}, \widehat{\xi}^i_{LM})$, *i.e.*, the L-moments estimators of the GEV parameters relative of the ith station (see Appendix A for further details).

(Step 2) From Equations (2)-(3), estimate the inverse of the marginal parametric estimator of the GEV distribution of the daily annual maxima of the ith station, *i.e.*, $G^{\leftarrow}(\cdot, \widehat{\Lambda}^i_{LM})$.

(Step 3) Fix $\beta \in [0,1]$;

(Step 4) Evaluate $\widetilde{s}_{ij}$ as in (11).

(Step 5) Choose the Crámer-von Mises norm in (14) and $f(\cdot) = \exp(\cdot) - 1$.

(Step 6) Evaluate $n^c_f = \sum_{t=1}^{T} \mathbb{1}\{I^i_t = I^j_t = 1\}$, *i.e.*, the number of common observations in $(x_{(i,t)}, x_{(j,t)})^{\top}_{\{t=1,\ldots,T\}}$.

(Step 7) Fix a threshold value $\overline{n}$. *In the present work $\overline{n} = 10$.*

if $n^c_f \geq \overline{n}$, using (12)-(13) and (Step 2), evaluate $\widehat{d}_{ij}$ as in (14)

if $n^c_f < \overline{n}$, the ith and jth stations are assumed to be independent.

if $n^c_f = 0$, instead of the dissimilarity measure $\widehat{d}_{ij}$, we only consider $f(\widetilde{s}_{ij})$, *i.e.*, f applied on the normalised spatial distance between the ith and jth stations.

The associated `R` code can be found in `dissimilaritymeasure.R` file in the supplementary material CodeR folder.

Finally, in Algorithm 2, we describe how Algorithm 1 can be used to maximize the hierarchical copula log-likelihood $\mathcal{L}^C(\Lambda)$ in Equations (9)-(10).

To conclude this section we propose in Figure 7 an illustration of several obtained clusters of the (partially) nested Archimedean copula model in Equation 8 by choosing $\beta = 0.2$. We fitted different classical Archimedean copulas (Gumbel, Joe, Clayton, Frank) for each cluster and we provided the selection model in terms of the AIC criterion (Akaike (1974)). Obtained Archimedean copulas for each

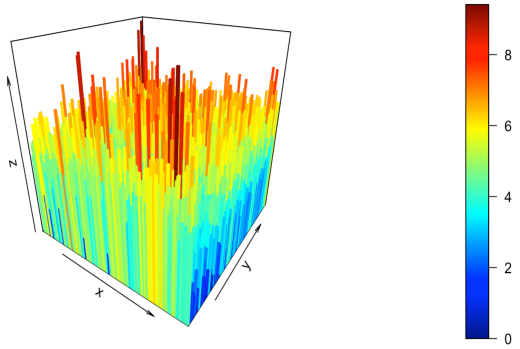

Fig. 5 Absolute value of dissimilarity measure in logarithm scale abs(ln($\widehat{d}_{ij}$)) for the combination of fitting stations displayed in Figure 3. Here we consider $\beta = 0.2$.

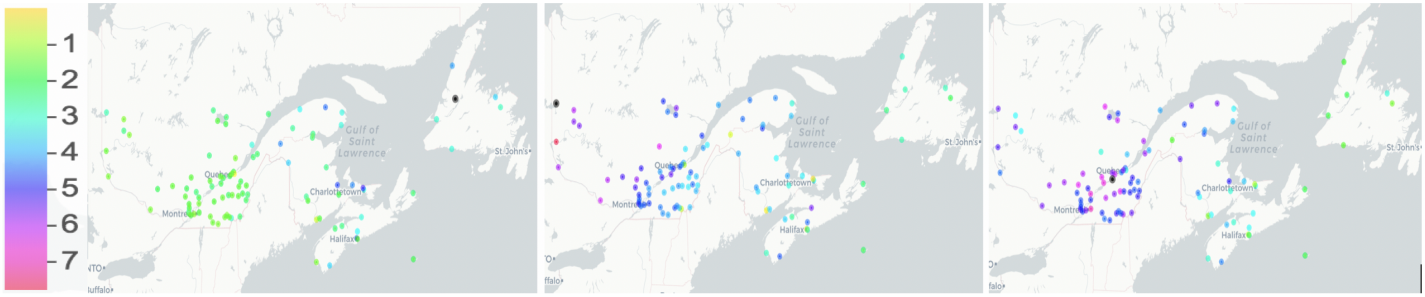

Fig. 6 Absolute value of dissimilarity measure in logarithm scale abs(ln($\widehat{d}_{ij}$)) for three fixed stations (black dots) with respect to the others fitting stations. Here we consider $\beta = 0.2$. We consider as fitting stations the combination displayed in Figure 3.

Algorithm 2 Proposed implementation for smooth copula-based GEV method

(Step 1) Choose several values for K (*i.e.*, the number of clusters) and β.

(Step 2) For these input parameters (K, β), estimate the copula-based dissimilarity measure $\widehat{d}_{ij}$ via Algorithm 1.

(Step 3) Provide the clustering of stations by running the PAM algorithm with $\widehat{d}_{ij}$ from (Step 2).

(Step 4) Select optimal (K^*, β^*) in (Step 1) with respect to the classical Average Silhouette Width (ASW) criterion.

(Step 5) Estimate the best Archimedean copula density $c_{\widehat{\theta}_k}$ in terms of the AIC criterion for each cluster with $k \in \{1, \ldots, K^*\}$.

(Step 6) Using $c_{\widehat{\theta}_k}$ from (Step 5) and marginal GEV parameter models from Table 1 maximize the log-likelihood $\mathcal{L}^C(\Lambda)$ in Equations (9)-(10).

The associated `R` code can be found in `smoothDEPENDENCEmodel.R` file (resp. `smoothINDEPENDENCEmodel.R` file for the independence setting of Equation (9)) in the supplementary material CodeR folder.

cluster are Gumbel with $\hat{\theta} = 1.656$ (left panel), Clayton with $\hat{\theta} = 0.595$ (center panel) and Joe with $\hat{\theta} = 1.119$ (right panel).

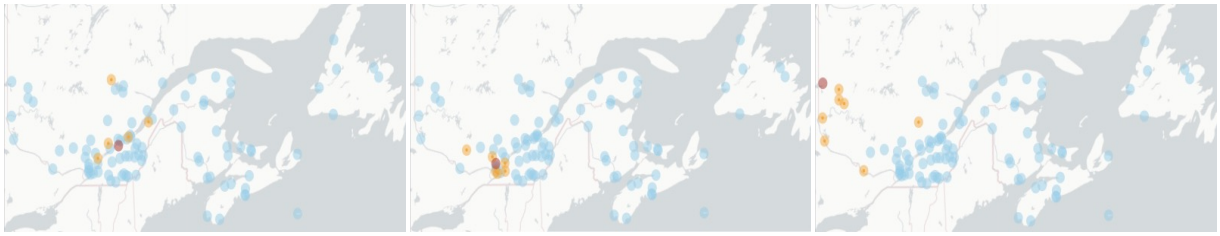

Fig. 7 Illustration for several obtained clusters. Centroid stations are presented by dark-red dots, element stations in each cluster by orange ones. Obtained Archimedean copulas for each cluster are Gumbel with $\hat{\theta} = 1.656$ (left panel), Clayton with $\hat{\theta} = 0.595$ (center panel) and Joe with $\hat{\theta} = 1.119$ (right panel).

4 Return level maps through classical spatial interpolation of individual GEV distribution

In the following, we will be interested in presenting naive interpolation routines used in practice to estimate return levels for every location $s \in S$ by interpolating individual GEV distributions.
For that purpose, we present in this section several exact and inexact techniques to derive the spatial interpolators $\tilde{\xi}(s)$, $\tilde{\mu}(s)$ and $\tilde{\sigma}(s)$ all based on the L-moments estimators $\hat{\xi}_{LM}$, $\hat{\mu}_{LM}$ and $\hat{\sigma}_{LM}$ in Appendix A and by considering covariates previously introduced in Table 1.
Let $\hat{\zeta} := \hat{\zeta}_{LM}$ be the L-moments estimator (either $\hat{\xi}_{LM}$, $\hat{\mu}_{LM}$ or $\hat{\sigma}_{LM}$) used in the interpolation (see Appendix A). In the rest of this section, for sake of simplicity we will drop the LM notation.

4.1 Exact interpolation techniques

Inverse distance weighted (IDW). The inverse distance weighted method provides an exact interpolation, *i.e.*, the interpolated value $\tilde{\zeta}(s_i)$ at station s_i is equal to the estimated value $\hat{\zeta}_i$ used in the interpolation. The IDW method is widely known as the basic one in the interpolation literature (Burrough (1986)). This method is based in the assumption that all the points on the earths surface are interdependent on the basis of distance. IDW technique provides satisfactory results when the number of points in the considered area is large and the points are uniformly distributed. However, it presents certain weaknesses (for details, the reader is referred to Achilleos (2011)). This interpolation technique implies that the influence of surrounding stations is reduced by large distances. In addition, distances are attenuated by weighting factors. Let us denote d_i, for $i = 1, \ldots, n_f$, the distance between the interpolating location s_i and the interpolated location s. Then, the interpolated value at location s is defined by

$$\tilde{\zeta}(s) = \frac{\sum_{i=1}^{n_f} \frac{\hat{\zeta}_i}{d_i}}{\sum_{i=1}^{n_f} \frac{1}{d_i}}.$$

We consider IDW with gradient correction (Nalder and Wein (1998)) in order to take into account the dependence between parameters and covariates. Let $y_s^{(1)}, , \ldots, y_s^{(r)}$ denote r covariates recorded for each station s, the interpolated value at location s is defined by

$$\tilde{\zeta}(s) = \frac{\sum_{i=1}^{n_f} \frac{\hat{\zeta}_i + \beta_1 (y_s^{(1)} - y_{s_i}^{(1)}) + \ldots + \beta_r (y_s^{(r)} - y_{s_i}^{(r)})}{||l_s - l_{s_i}||}}{\sum_{i=1}^{n_f} \frac{1}{||l_s - l_{s_i}||}}, \tag{15}$$

where l_s is the two-dimensional coordinate (longitude, latitude) of the location s, the parameters $\beta_1, \ldots, \beta_r$ correspond to the values that minimize the cross-validation score

$$\sum_{i=1}^{n_f} (\hat{\zeta}_i - \tilde{\zeta}_{-i}(s_i))^2, \tag{16}$$

with $\tilde{\zeta}_{-i}(s_i)$ the interpolated ζ at s_i when this station is not considered in Equation (15). Detailed steps are provided in Algorithm 3.

Algorithm 3 Proposed implementation for IDW method

(Step 1) Consider altitude and mean precipitation covariates, denoted $(y_s^{(1)}, y_s^{(2)}) = (a_s, \bar{m}_s)$ at a given station with two-dimensional coordinate l_s.

(Step 2) Equation (15) can be written as

$$\tilde{\zeta}(s) = \frac{\sum_{i=1}^{n_f} \frac{\hat{\zeta}_i + \beta_a(a_s - a_{s_i}) + \beta_{\bar{m}}(\bar{m}_s - \bar{m}_{s_i})}{||l_s - l_{s_i}||}}{\sum_{i=1}^{n_f} \frac{1}{||l_s - l_{s_i}||}}.$$

(Step 3) Estimate parameters β_a and $\beta_{\bar{m}}$ by minimizing the cross-validation score in (16).

Universal Kriging. The main principle of kriging is to compute the best linear unbiased estimator of $\zeta(s)$ by the calculation of a weighted average of the known values of ζ in the neighborhood of s. The most general case, universal kriging, was set out in Matheron (1969). Unlike the simple kriging, the expectation of random function model $\zeta(s)$ is allowed to vary spatially. In universal kriging, it is assumed that

$$\mathbb{E}[\zeta(s)] = \beta(s) \equiv \sum_{j=0}^{r} \beta_j f_j(s), \tag{17}$$

where f_j are known functions and the β_j, $j = 0, 1, \ldots, r$, are unknown coefficients. Usually, $f_0(s) = 1$, $\forall s$, which guarantees that the constant-mean case is included in the model. The model for universal kriging is given by

$$\zeta(s) = \beta(s) + G(s), \tag{18}$$

where $G(s)$ is a zero-mean Gaussian process which defines the spatial dependence. In order to predict ζ in Equation (18), we need to estimate the β_j parameters with $j = 0, 1, \ldots, r$ and, the variogram associated to $G(s)$ (see Chapters 5 and 6 in Diggle and Ribeiro (2007)). The mean square error predictor of $\zeta(s)$ is defined by

$$\tilde{\zeta}(s) = \tilde{\beta}(s) + \sum_{i=1}^{n_f} \lambda_i(s) \left(\hat{\zeta}_i - \tilde{\beta}(s) \right),$$

where $\tilde{\beta}(s) = \sum_{j=0}^{r} \hat{\beta}_j f_j(s)$ with $\hat{\beta}_j$ denoting the estimator of β_j in Equation (17), $j = 0, \ldots, r$, and $\lambda_i(s)$, $i = 1, \ldots, n_f$, the prediction weights (see Section 3.4. in Chilès and Delfiner (2009)). If the variogram of $G(s)$ is supposed to be continuous at the origin, the nugget effect (*i.e.*, microscale variations) is not considered. In the above case, kriging is an exact interpolation technique (see Section 3.2.1 in Cressie (1993)). One of the most applied version of universal kriging is when functions $f_j(s)$, $j = 1, \ldots, r$ are considered as explanatory variables. That is, if we assume that $\beta(s)$ in Equation (17) is explained by r covariates, $y_s^{(1)}, , \ldots, y_s^{(r)}$, then Equation (18) can be written as

$$\zeta(s) = \beta_0 + \sum_{j=1}^{r} \beta_j y_s^{(j)} + G(s). \tag{19}$$

Detailed steps are gathered in Algorithm 4. The `R` code associated to Algorithm 3 and Algorithm 4 can be found in the `classicalinterpolationtechniques.R` file in the supplementary material CodeR folder.

4.2 Inexact interpolation techniques

Let consider polynomial and spline-based regression models presented in Section 3.2. Since the error term ϵ_s, techniques from Equation (4) do not provide exact interpolations. Firstly, Algorithm 5 presents steps for the implementation of the proposed polynomial regression method. Secondly Algorithm 6

Algorithm 4 Proposed implementation for Universal Kriging method

(Step 1) Consider $\beta(s)$ in (17) as a first order polynomial on the two-dimensional coordinates l_s, with mean and altitude covariates.

(Step 2) Assume that the variogram of G in (19) is continuous at the origin.

(Step 3) Estimate the variogram of G via maximum likelihood method for several covariance functions. *We consider exponential, spherical, circular, cubic, Matérn and Gneiting covariance functions.*

(Step 4) Choose the covariance function associated to the best fitting in terms of the AIC criterion.

Algorithm 5 Proposed implementation for polynomial regression method

(Step 1) Consider covariates as polynomials of longitude, latitude, altitude and mean precipitation with a maximum degree of 3.

(Step 2) Take all possible combinations between covariates built in Step 1 with a maximum interaction degree of 3.

(Step 3) Choose the combination in Equation (5) associated to the best model by AIC criterion.

Algorithm 6 Proposed implementation for spline-based regression method

(Step 1) Using (7), we fix the interpolated value at location s as
$$\tilde{\zeta}(s) = \tilde{\beta}_0 + \tilde{\beta}_1 a_s + \tilde{\beta}_2 \bar{m}_s + \tilde{F}(l_s).$$

(Step 2) Fix 10000 combinations of 15 knots among $n_f = 95$ fitting stations.

(Step 3) Select the best model associated to the combination of 15 knots that provides the lowest value of generalized cross-validation (GCV) score (see, e.g., Section 4.2.3 in Wood (2017)).

gathers steps for the proposed spline-based regression method. In the present study, the spline is a function of the coordinates and altitude and mean precipitation covariates are considered linearly.
The `R` code associated to Algorithm 5 and Algorithm 6 to can be found in `classicalinterpolationtechniques.R` file in the supplementary material CodeR folder.

5 Quality of predictions

In order to evaluate the quality of the proposed models, we introduce several measures to measure the accuracy of the several fits. Particularly, we are interested in the quality of the interpolated distributions. To this end, we study the difference between accuracy measures for 95 fitting stations and 21 validation stations. A consistent way to measure the quality is to compare the goodness-of-fit of the quantiles of the interpolated GEV parameters versus the observed ones. Let $z_{s_i}^{(1)}, \ldots, z_{s_i}^{(m)}, \ldots, z_{s_i}^{(M)}$ denote the $M = 30$ empirical quantiles at location s_i, where the mth value $z_{s_i}^{(m)}$ is associated to a probability $p_m = \frac{m-1/2}{M}$. Also, $\tilde{q}_{p_m,s_i}$ denotes the $(1-p_m)$ quantile of the interpolated GEV distribution at location s_i, obtained by Equation (3) replacing Λ by the interpolated values $\widetilde{\Lambda}$ with $p = p_m$. The considered goodness-of-fit scores are gathered in Table 2.

From the GEV interpolated parameters through the different proposed techniques in Sections 3 and 4, we calculate the corresponding scores gathered in Table 2 for 200 combinations of index sets I_f and I_v, *i.e.*, for 200 combinations of fitting and validation stations. The obtained boxplots are gathered in Figure 8. Furthermore, the associated medians and standard deviations are displayed in Tables 3 and 4. In these tables, the smooth independent GEV model refers to Equation (9).

Relative Mean Squared Error	$RMSE = \sqrt{\frac{1}{\widetilde{n}\,M}\sum_{i=1}^{\widetilde{n}}\sum_{m=1}^{M}\left(z_{s_i}^{(m)} - \tilde{q}_{p_m,s_i}\right)^2}$
Mean Absolute Error	$MAE = \frac{1}{\widetilde{n}\,M}\sum_{i=1}^{\widetilde{n}}\sum_{m=1}^{M}\lvert z_{s_i}^{(m)} - \tilde{q}_{p_m,s_i}\rvert$
Maximum Prediction Error	$MPE = \max_{i\in\{1,\ldots,\widetilde{n}\}}\max_{m\in\{1,\ldots,M\}}\lvert z_{s_i}^{(m)} - \tilde{q}_{p_m,s_i}\rvert$
Bias	$B = \frac{1}{\widetilde{n}\,M}\sum_{i=1}^{\widetilde{n}}\sum_{m=1}^{M}\left(z_{s_i}^{(m)} - \tilde{q}_{p_m,s_i}\right)$

Table 2 Considered goodness-of-fit scores. Here we consider $M = 30$, $\widetilde{n} = n_f$ for the fitting stations analysis and $\widetilde{n} = n_v$ for the validation one.

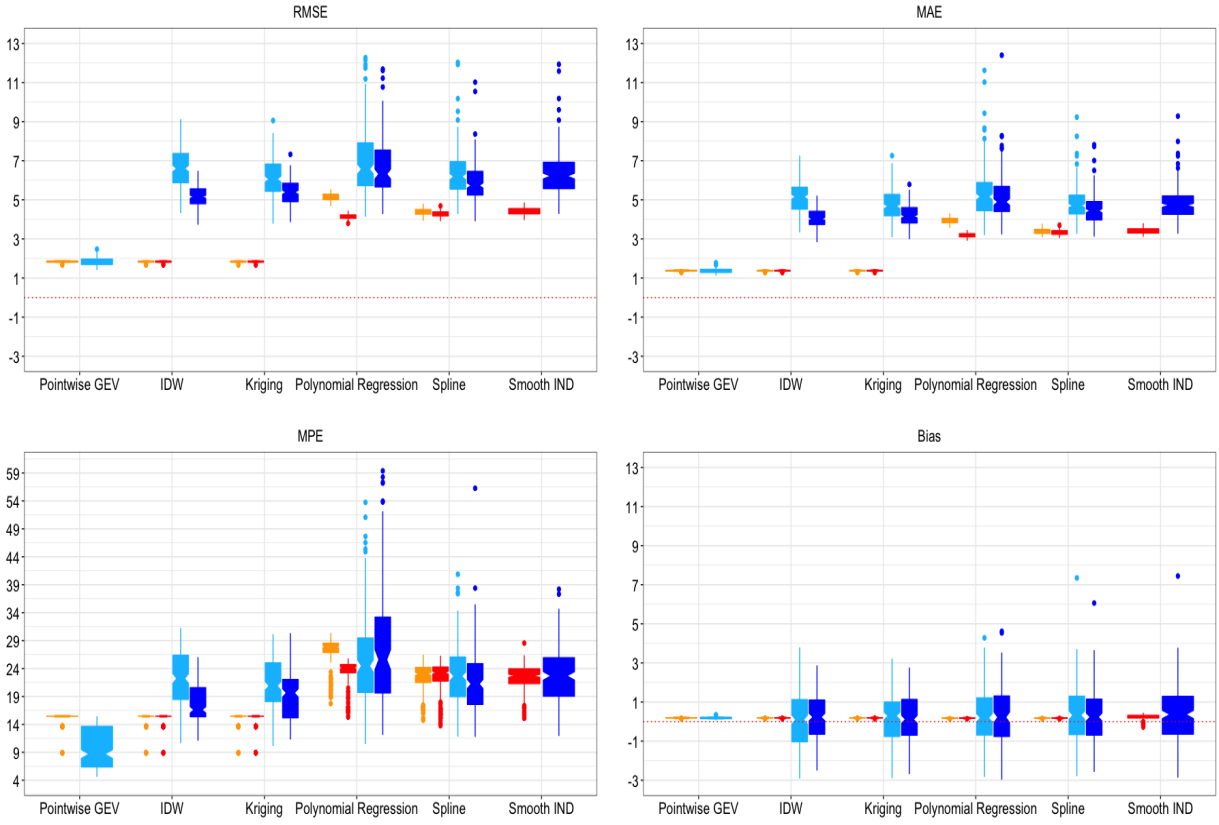

Fig. 8 Boxplots of obtained scores from Table 2 for 200 combinations between fitting and validation stations for each interpolation technique. Models with three geographical covariates are displayed in orange boxplots for fitting stations and in sky-blue boxplots for validation ones. Models with the mean precipitation in addition to the three geographical covariates are displayed in red boxplots for fitting stations and in dark -blue boxplots for validation ones.

In this analysis, we consider models with the three geographical covariates (see Table 3 and associated orange and sky-blue boxplots in Figure 8) and models where the mean precipitation covariate is additionally taken into account (see Table 4 and associated red and dark-blue boxplots in Figure 8). We also introduce the scores by considering the quantiles with pointwise L-moments estimators for the GEV parameters in Appendix A, called *Pointwise GEV* in Table 3 and Figure 8. This means that, for all methods validation stations, the scores correspond to the predictions except for the *Pointwise GEV* scores which are fitting scores. Then, the *Pointwise GEV* scores can be interpreted as lower bounds of the error that would result from a prediction. We refer the interested reader to Figure 12 of Appendix A, for a pointwise return level map associated to these *Pointwise GEV* estimators. Obviously, since

	Fitting stations				Validation stations			
	RMSE	MAE	MPE	B	RMSE	MAE	MPE	B
1. Pointwise GEV	1.85 (0.04)	1.38 (0.02)	15.45 (1.18)	0.19 (0.01)	1.81 (0.18)	1.36 (0.11)	8.68 (3.74)	0.18 (0.05)
2. IDW	1.85 (0.04)	1.38 (0.02)	15.45 (1.18)	0.19 (0.01)	6.64 (1.01)	5.17 (0.79)	22.12 (4.44)	0.15 (1.51)
3. Polynomial regression	5.16 (0.18)	3.95 (0.15)	27.91 (3.01)	0.18 (0.01)	6.62 (2.42)	5.16 (1.24)	25.04 (12.79)	0.14 (1.64)
4. Spline	4.40 (0.16)	3.39 (0.13)	23.02 (2.74)	0.18 (0.01)	6.20 (4.51)	4.71 (1.55)	22.83 (22.83)	0.30 (1.93)
5. Kriging	1.85 (0.04)	1.38 (0.02)	15.45 (1.18)	0.19 (0.01)	6.06 (0.94)	4.67 (0.74)	20.90 (4.67)	0.27 (1.35)

Table 3 Median scores from Table 2 for 200 combinations between fitting and validation stations for each interpolation technique. Associated standard deviations are displayed in brackets. We consider here the three geographical coordinates as covariates.

	Fitting stations				Validation stations			
	RMSE	MAE	MPE	B	RMSE	MAE	MPE	B
1. IDW	1.85 (0.04)	1.38 (0.02)	15.45 (1.18)	0.19 (0.01)	5.16 (0.57)	4.04 (0.48)	16.72 (4.06)	0.21 (1.13)
2. Polynomial regression	4.16 (0.13)	3.21 (0.12)	24.13 (2.81)	0.18 (0.01)	6.41 (2.08)	4.89 (1.19)	25.68 (11.13)	0.21 (1.52)
3. Spline	4.28 (0.14)	3.32 (0.12)	23.25 (3.08)	0.18 (0.01)	5.74 (3.67)	4.44 (1.27)	21.44 (19.75)	0.23 (1.67)
4. Kriging	1.85 (0.04)	1.38 (0.02)	15.45 (1.18)	0.19 (0.01)	5.39 (0.62)	4.16 (0.52)	19.63 (4.23)	0.11 (1.18)
5. Smooth independent GEV	4.41 (0.17)	3.41 (0.14)	22.69 (2.67)	0.24 (0.13)	6.22 (4.50)	4.73 (1.55)	23.03 (22.82)	0.36 (1.94)

Table 4 Median scores from Table 2 for 200 combinations between fitting and validation stations for each technique. Associated standard deviations are displayed in brackets. We consider here the three geographical coordinates and the mean precipitation as covariates.

IDW and kriging techniques provide exact interpolations, their results exactly correspond with the ones from *Pointwise GEV* parameters estimation.

Table 3 suggests that when using only longitude, latitude and elevation as covariates, kriging performs better, since almost all considered scores are lower. Spline and IDW perform similarly. The less performing models seems to be the polynomial regression ones, both in terms of median values (see Table 3) and of sensitivity of the combinations between fitting and validation stations (see boxplots in Figure 8). Prediction seems to quickly deteriorate away from the fitting stations. Indeed, results for the combinatory validation stations (sky-blue and dark-blue boxplots in Figure 8) are relatively poor compared to those for the fitting stations (orange and red ones). This considerations is true in particular for RMSE and MAE scores. Conversely, performances seem to be more stable in terms of bias. Moreover due to the small number of considered validation stations ($n_v = 21$) and the discrepancy between n_f and n_v, the variance of the sky-blue and dark-blue boxplots in Figure 8 is considerably large. In addition, we can observe that the choice of fitting and validation stations produces a larger impact over the polynomial regression techniques (see variability of boxplots for polynomial regression in Figure 8 for instance in terms of MPE).

In Figure 8, it can be observed that the behaviour of models with three geographical covariates and mean precipitation as covariates slightly improve models with only the three geographical covariates. Since essentially all the considered scores decrease, the first 4 lines to Table 4 show the global improvement compared to Table 3, when using additionally the mean precipitation as a covariate. All interpolation methods have analogous performance and still universal kriging and IDW perform slightly better. Note that in Table 4 (line 5), error measurements from the independent smooth GEV model, based on Equation (9), are quite high compared to those when a GEV is fitted to each station separately (first line of Table 3). Notice that these errors cannot directly be compared since the individual GEV fitting uses all available information at the validation stations for parameter estimation, while this information is not used in the parameter estimation of the smooth GEV model.

To illustrate the performance of the smooth copula-based GEV model proposed in Section 3, we now consider the particular combination between fitting and validation stations previously displayed in Figure 3. For this combination, our error measures are gathered in Table 5. One can observe that the quality in this last model (Table 5, line 7) remains globally more stable in the validation stations with respect to the considered error measures, when other methods deteriorate quickly the quality of the predictions. These results show a better performance of the considered smooth GEV distribution fitting, in particular in this sparse station network situation.

	Fitting stations				Validation stations			
	RMSE	MAE	MPE	B	RMSE	MAE	MPE	B
1. IDW	1.87	1.39	15.45	0.19	4.88	3.96	15.86	−1.29
2. Polynomial regression	4.10	3.22	24.18	0.18	6.02	4.60	19.04	−0.25
3. Spline	4.48	3.50	22.35	0.17	3.91	3.12	15.29	−0.84
4. Kriging	1.87	1.39	15.45	0.19	5.06	4.00	15.17	−1.42
5. Smooth independent GEV	4.69	3.60	22.43	0.18	4.28	3.27	15.30	−0.61
7. Smooth copula-based GEV	4.69	3.61	22.53	0.33	4.27	3.23	15.25	−0.44

Table 5 Goodness-of-fit scores from Table 2 for the combination between fitting and validation stations displayed in Figure 3 for each technique. We consider here the three geographical coordinates and the mean precipitation as covariates.

Associated return level maps for IDW, Kriging, Spline and smooth copula-based GEV models in Table 5 are displayed in Figures 9-10.

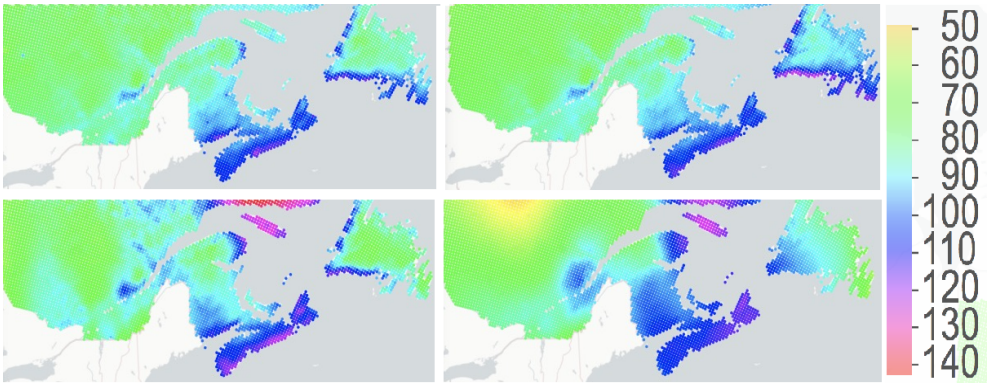

Fig. 9 Obtained 20-years precipitation return level maps in mm in Central Eastern Canada. First row: IDW (left panel), Kriging (right panel). Second row: Spline (left panel) and smooth copula-based GEV (right panel).

One can appreciate in Figures 9-10 a global similar behavior for the IDW (first row, left panel), kriging (first row, right panel) and spline (second row, left panel) methods. Conversely, a slight different return level map can be observed for the smooth copula-based GEV method (second row, right panel), due to the considered local spatial and copula dependencies. Furthermore, remark that the smooth return level maps can be computed from the fitted model without any further interpolation. Finally, one can compare return levels in Figures 9-10 with the map recently obtained in Perreault et al. (2019) for the same Central Eastern Canada rainfall dataset, where the spatial effect is modeled via Gaussian Markov random fields. Indeed, the 24h duration 20-years precipitation return level map in Perreault et al. (2019) shows a very similar behavior of smooth copula-based one (see second row, right panel in Figure 9).

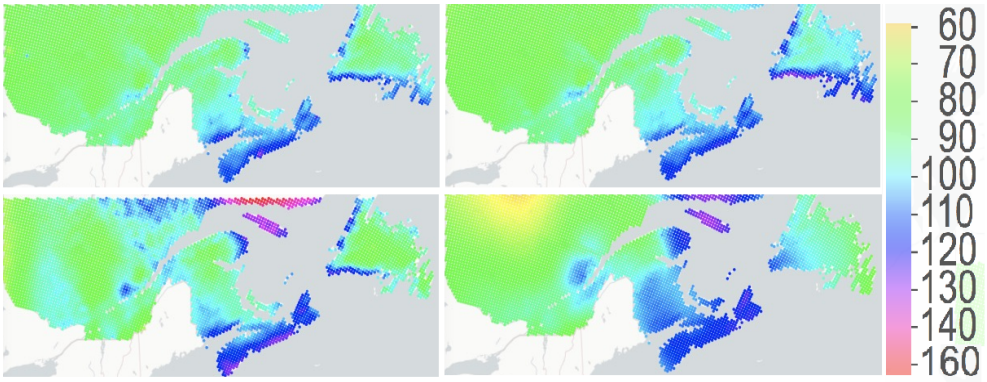

Fig. 10 Obtained 40-years precipitation return level maps in mm in Central Eastern Canada. First row: IDW (left panel), Kriging (right panel). Second row: Spline (left panel) and smooth copula-based GEV (right panel).

Acknowledgements Funding in partial support of this work was provided by the SIMONS Foundation and the Center for Mathematical Research with the Program Simons-CRM. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. We are grateful to Mitacs Globalink Research Award for finance support. The authors thank Jonathan Jalbert (Polytechnique Montral, Canada) for useful discussions.

References

Achilleos, G. A. (2011). The inverse distance weighted interpolation method and error propagation mechanism creating a DEM from an analogue topographical map. *Journal of Spatial Science*, 56(2):283–304.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Asong, Z., Khaliq, M., and Wheater, H. (2015). Regionalization of precipitation characteristics in the canadian prairie provinces using large-scale atmospheric covariates and geophysical attributes. *Stochastic Environmental Research and Risk Assessment*, 29(3):875–892.

Beguería, S. and Vicente-Serrano, S. M. (2006). Mapping the Hazard of Extreme Rainfall by Peaks over Threshold Extreme Value Analysis and Spatial Regression Techniques. *Journal of Applied Meteorology and Climatology*, 45:108–124.

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, 5(3):1699–1725.

Blanchet, J. and Lehning, M. (2010). Mapping snow depth return levels: smooth spatial modeling versus station interpolation. *Hydrology and Earth System Sciences*, 14(12):2527–2544.

Bresson, E., Laprise, R., Paquin, D., Thériault, J., and de Elía, R. (2017). Evaluating the ability of crcm5 to simulate mixed precipitation. *Atmosphere-Ocean*, 55(2):79–93.

Burrough, P. A. (1986). *Principles of Geographical Information Systems for Land Resources Assessment*. Oxford University Press, Oxford.

Chilès, J. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, Inc.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian Spatial Modeling of Extreme Precipitation Return Levels. *Journal of the American Statistical Association*, 102:824–840.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Das, S., Zhu, D., and Yin, Y. (2020). Comparison of mapping approaches for estimating extreme precipitation of any return period at ungauged locations. *Stochastic Environmental Research and Risk Assessment*, 34(8):1175–1196.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer-Verlag New York.

Disegna, M., DUrso, P., and Durante, F. (2017). Copula-based fuzzy clustering of spatial time series. *Spatial Statistics*, 21:209 – 225.

Ferreira, A. and de Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, 43(1):276–298.

Gardes, L. and Girard, S. (2010). Conditional extremes from heavy-tailed distributions: an application to the estimation of extreme rainfall return levels. *Extremes*, 13:177–204.

Hofert, M. and Pham, D. (2013). Densities of nested Archimedean copulas. *Journal of Multivariate Analysis*, 118:37 – 52.

Hosking, J. R. M. and Wallis, J. R. (1997). *Regional Frequency Analysis: An Approach based on L-Moments.* Cambridge University Press, Cambridge, UK.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.

Hwang, Y., Clark, M., Rajagopalan, B., and Leavesley, G. (2012). Spatial interpolation schemes of daily precipitation for hydrologic modeling. *Stochastic environmental research and risk assessment*, 26(2):295–320.

Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics*, 22(1):47–64.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. *In: Dodge Y (ed) Statistical Data Analysis Based on the L_1 Norm and Related Methods*, pages 405–416.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons.

Khedhaouiria, D., Mailhot, A., and Favre, A.-C. (2020). Regional modeling of daily precipitation fields across the great lakes region (canada) using the cfsr reanalysis. *Stochastic Environmental Research and Risk Assessment*, 34(9):1385–1405.

Kohnová, S., Parajka, J., Szolgay, J., and Hlavčová, K. (2009). *Mapping of Gumbel Extreme Value Distribution Parameters for Estimation of Design Precipitation Totals at Ungauged Sites*, pages 129–136. Springer Netherlands.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer Verlag, New York.

Li, M., Shao, Q., and Renzullo, L. (2010). Estimation and spatial interpolation of rainfall intensity distribution from the effective rate of precipitation. *Stochastic Environmental Research and Risk Assessment*, 24(1):117–130.

Lomba, J. S. and Alves, M. I. F. (2020). L-moments for automatic threshold selection in extreme value analysis. *Stochastic Environmental Research and Risk Assessment*, 34(3):465–491.

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28(2):193–209.

Matheron, G. (1969). Le Krigeage Universel. *In: Fontainebleau: Cahiers du Centre de Morphologie Mathématique, vol. 1. École des Mines de Paris.*

Nalder, I. A. and Wein, R. W. (1998). Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest. *Agricultural and Forest Meteorology*, 92:211–225.

Nelsen, R. B. (1999). *An introduction to copulas*, volume 139 of *Lecture Notes in Statistics.* Springer-Verlag, New York.

Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105:263–277.

Patton, A. J. (2006). Estimation of multivariate models for time series of possibly different lengths. *Journal of Applied Econometrics*, 21(2):147–173.

Perreault, L., Haché, M., Slivitzky, M., and Bobée, B. (1999). Detection of changes in precipitation and runoff over eastern canada and us using a bayesian approach. *Stochastic Environmental Research and Risk Assessment*, 13(3):201–216.

Perreault, L., Jalbert, J., and Genest, C. (2019). Interpolation of extreme precipitation of multiple durations in eastern canada. Conference: 11th International Conference on Extreme Value Analysis (EVA 2019), available on researchgate.net/publication/344563326.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis.* Wiley Series in Probability and Statistics.

Reynolds, A., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms*, 5:475–504.

Ruppert, D., Wand, M. P., and Carrol, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics.

Saad, C., El Adlouni, S., St-Hilaire, A., and Gachon, P. (2015). A nested multivariate copula approach to hydrometeorological simulations of spring floods: the case of the Richelieu River (Québec, Canada) record flood. *Stochastic Environmental Research and Risk Assessment*, 29(1):275–294.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes : Statistical Theory and Applications in Science, Engineering and Economics*, 5(1):33–44.

Schubert, E. and Rousseeuw, P. J. (2019). Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In *Similarity Search and Applications*, pages 171–187. Springer International Publishing.

Segers, J. (2015). Hybrid copula estimators. *Journal of Statistical Planning and Inference*, 160:23 – 34.

Szolgay, J., Parajka, J., Kohnová, S., and Hlavčová, K. (2009). Comparison of mapping approaches of design annual maximum daily precipitation. *Atmospheric Research*, 92:289–307.

Wi, S., Valdés, J. B., Steinschneider, S., and Kim, T.-W. (2016). Non-stationary frequency analysis of extreme precipitation in south korea using peaks-over-threshold and annual maxima. *Stochastic environmental research and risk assessment*, 30(2):583–606.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall.

Yoon, S., Kumphon, B., and Park, J.-S. (2015). Spatial modelling of extreme rainfall in northeast thailand. *Procedia Environmental Sciences*, 26:45–48. Spatial Statistics conference 2015.

A L-moments for GEV parameters

An illustration that L-moments are efficient in estimating parameters of a wide range of distributions from small sample sizes is presented in Hosking and Wallis (1997). Since we have a small number of observations for several stations (see Figure 1), we consider the L-moments estimators in order to estimate the GEV parameters. The L-moments estimators for the GEV distribution parameters $\hat{\Lambda}_{LM} = (\hat{\mu}_{LM}, \hat{\xi}_{LM}, \hat{\sigma}_{LM})$ in (2) are defined as

$$\hat{\xi}_{LM} = 7.8590c + 2.9554c^2, \quad \hat{\sigma}_{LM} = \frac{\hat{\beta}_0 \hat{\xi}_{LM}}{(1 - 2^{-\hat{\xi}_{LM}})\Gamma(1 + \hat{\xi}_{LM})}, \quad \hat{\mu}_{LM} = \hat{\beta}_0 - \frac{\hat{\sigma}_{LM}}{\hat{\xi}_{LM}}[1 - \Gamma(1 + \hat{\xi}_{LM})],$$

with $c = 2/(3 + \hat{\tau}_3) - \log(2)/\log(3)$ and $\hat{\tau}_3 = \frac{6\hat{\beta}_2 - 6\hat{\beta}_1 + \hat{\beta}_0}{2\hat{\beta}_1 - \hat{\beta}_0}$, where $\hat{\beta}_r$ are suitable estimators of the probability weighted moments of order r (see Hosking et al. (1985) for more details), for $r = 0, 1, 2$.

The return levels over each location are calculated by plugging the L-moments estimators above in Equation (3). By considering the estimated return levels, the return level plots for 3 locations with different altitudes are depicted in Figure 11. These panels represent $q(p; \hat{\Lambda}_{LM})$ versus $-\ln(1-p)$ on a logarithm scale, and provide the highest value expected to be exceeded once every r years for any return period r on $x-$axis. From Equation (3), when $\xi < 0$, the return level plot is convex with asymptotic limit as $p \to 0$ at $\mu - \frac{\sigma}{\xi}$; when $\xi > 0$, the plot is concave with not finite bound; when $\xi = 0$, the plot is linear. In Figure 11 one can appreciate the quality of the fitting of GEV L-moments estimators to our data.

Figure 12 shows the resulting pointwise 20-years and 40-years return levels for the considered 116 stations. Such a map is nevertheless difficult to interpret and can only give information for the few locations where data are available. In practice, spatial return levels as in Figures 9-10, rather than pointwise estimates as in Figure 12, would be of much higher value.

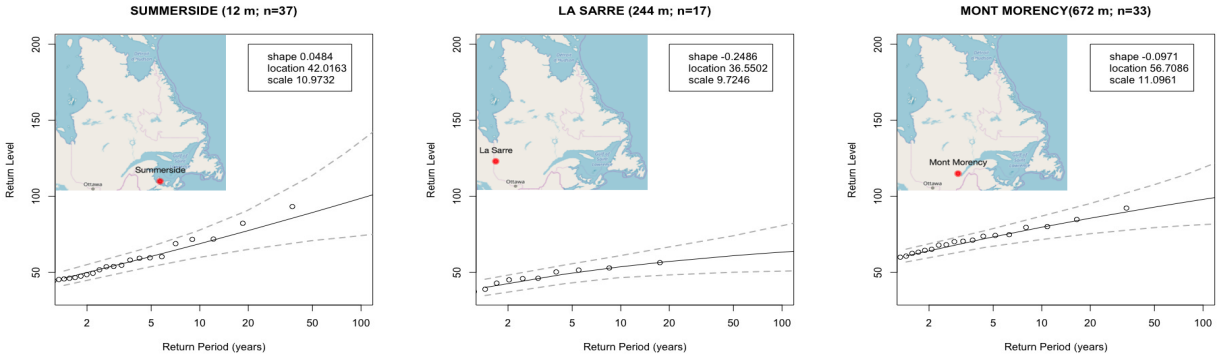

Fig. 11 Locations of 3 selected stations and adequacy of fitted GEV model through the associated return level plots.

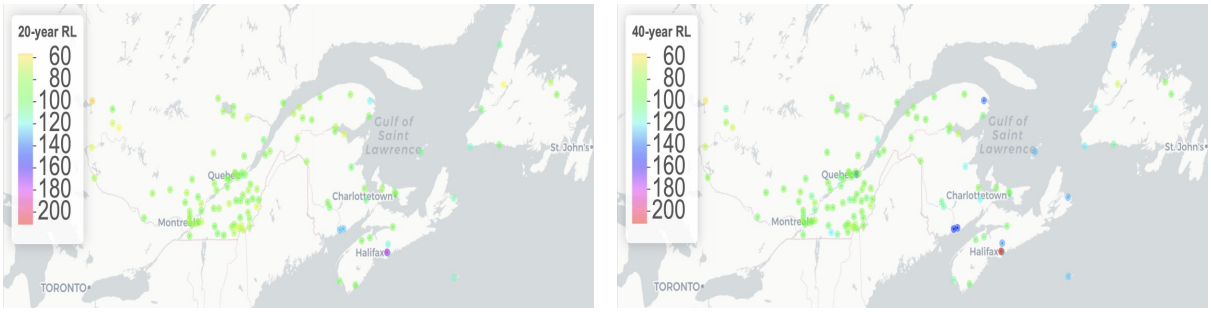

Fig. 12 Pointwise 20-years (left) and 40-years (right) precipitation return level map in mm for the considered 116 stations in Central Eastern Canada.