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

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

BY

Fatima Palacios-Rodriguez[1], Elena Di Bernardino[2], Melina Mailhot[3]

Abstract

This paper proposes a smooth copula-based Generalized Extreme Value (GEV) model to map and predict extreme rainfall in Central Eastern Canada. The considered data contains a large portion of missing values, and one observes several non-concomitant record periods at different stations. The proposed two-steps approach combines GEV parameters' smooth functions in space through the use of spatial covariates and a flexible hierarchical copula-based model to take into account dependence 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 introduced in the literature. Finally, we compare the classical GEV parameter interpolation approaches with the proposed smooth copula-based GEV modeling approach.

Keywords: Copula-based Clustering, Extreme Value Theory, Hydrology, Spatial Interpolation, Missing values, Non-concomitant record periods.

1 Introduction

Heavy rainfall can have disastrous consequences on health of financial systems 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, see, *e.g.*, Beirlant et al. (2004), de Haan and Ferreira (2006)) 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. Statistical 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 random variable evaluated as precisely as possible in terms of distribution, measurements and variability.

[1]Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Calle Tarfia sin número, 41012 Seville, Spain. Email: fpalacios2@us.es.

[2]Laboratoire J.A. Dieudonné, UMR CNRS 7351, Université Côte d'Azur, Nice, France. Tel.: (33) (0) 4 89 15 04 95. Email: elenadb@unice.fr.

[3]Department of Mathematics and Statistics, Concordia University, 1455 De Maisonneuve Blvd. W., Montréal (QC) Canada H3G 1M8. Email: melina.mailhot@concordia.ca.

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 spacial interpolation techniques to estimate precipitation quantities from marginal extreme value distributions in order to provide return level maps. This classical approach is frequently used, *e.g.* in Beguería and Vicente-Serrano (2006), Cooley et al. (2007) and Kohnová et al. (2009). A comparison of different traditional 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).

The second group in the extreme return level maps literature is based on the direct estimation of the spatial extremal distribution. 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. 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.*, Smith (1990), Schlather (2002), Padoan et al. (2010), Davison et al. (2012), Reich and Shaby (2012), Ribatet et al. (2012), Huser et al. (2019)) and Bayesian hierarchical models (see *e.g.*, Johannesson et al. (2021)). 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). An extreme rainfall model assuming regional dependence is proposed in Yoon et al. (2015) and random fields approaches are presented in Sang and Gelfand (2009) and Sang and Gelfand (2010).

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.

In this paper, we first focus 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 (116 stations for more than 1,000,000 km^2) of recording stations compared to Lehmann et al. (2016), for example, where 872 and 1348 stations are available on a 156,000 and 580,000 km^2 regions, respectively, and literature is quite scarce for this specific area. Note 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 focusses on a southern Canadian region.

Even given the scarce aspect of the dataset, the obtained results show a robust performance with the considered smooth Generalized Extreme Value (GEV) distribution fitting methods. The same dataset has been analyzed in Perreault et al. (2022) where a Bayesian hierarchical interpolation model is proposed with the spatial effect modeled via Gaussian Markov random fields (see, *e.g.,*

Rue and Held (2005)). Results of Section 4 can be compared with those of Perreault et al. (2019).

Also, we consider a dataset which presents several missing values and non-concomitant record periods at different stations. The interested reader is referred to Figure 21 (see Appendix E) for a graphical illustration of this crucial point. It implies that when one aims to model the joint behaviour of the extreme rainfall at a specific station, the estimated marginal distribution uses the complete series for a given 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 cope with this situation, we consider the *hybrid copula model* (see *e.g.*, Segers (2015)). This model is an extension of the classical empirical copula, obtained by combining an estimator of a multivariate cumulative distribution with estimators of the marginal cumulative distributions. Note that, in the missing data framework, the considered marginal estimators are not necessarily equal to the margins obtained through the joint estimator.

The main contribution of the present work is the proposed spatial smooth GEV model, mixing response surfaces for the GEV parameters' models, with a flexible joint dependence framework via a hierarchical copula, taking into account the spatio-temporal dependence structure between the recording stations. The proposed model for the return level maps estimation is based on several inference steps, in order to deal with the challenges of the data, concisely described in the following. The interested reader is referred to Section 2 for the data description and Section 3.4 for the associated detailed algorithms.

First, in order to handle the dependence between stations, we estimate our adapted version of the copula-based dissimilarity measure recently proposed in Disegna et al. (2017) (see Equation (14)). For each couple of stations, this measure is based on a convex combination between the spatial distance and the copula behaviour. Pseudo-observations for the copula are built by using the point-wise GEV quantiles with parameters estimated via the $L-$moment method. Then, on these pseudo-observations series of different lengths and non-concomitant years, we consistently estimate the dissimilarity by using the hybrid empirical copula (see Algorithm 1).
Second, we build a hierarchical copula model via a PAM clustering algorithm based on the estimated dissimilarity, considering the spatial dependence between stations (see Algorithm 2).
Finally we include the GEV smooth surface parameters' with polynomial regression and spline models and we maximize the obtained hierarchical loglikelihood (see Equations (9)- (10)). Considered covariates included in the considered polynomial/spline models are presented in Table 1. The resulting smooth functions for the GEV parameters are used to provide the estimated return level maps via Equation (3).

Furthermore 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 approach proposed in this paper, using recently introduced tools, in terms of obtained return level maps (see Figure 8).

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. Obtained results on considered rainfall data are presented in Section 4. Conclusion and perspectives are discussed in Section 5. Appendix A is devoted to illustrate, via simulation studies, the performance of the proposed dissimilarity measure and associated clustering algorithm. In Appendix B we provide a more comprehensive simulation study to evaluate the estimation performance for the final return level maps and compare it to alternative methods (the Gaussian latent model in Zhang et al. (2020) and the hierarchical max-stable model in Reich and Shaby (2012)). In Appendix C we build return level maps via classical spatial interpolation methods of individual GEV distributions. Further details on L-moments for GEV parameters are postponed in Appendix D. Finally additional information are provided in Appendix E.

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 shortage of the recorded data, as depicted in Figure 21, 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, which will be used 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 $(n \times T)-$matrix, $\mathbf{X} = (x_i)_{i=1,\ldots,n}$, where $n = 116$ (number of stations) and $T = 104$ (length of the considered whole time window).
The spatial location of the considered stations is shown in Figure 1. Remark that all maps in this paper have been obtained with package `leaftet` in `R`.

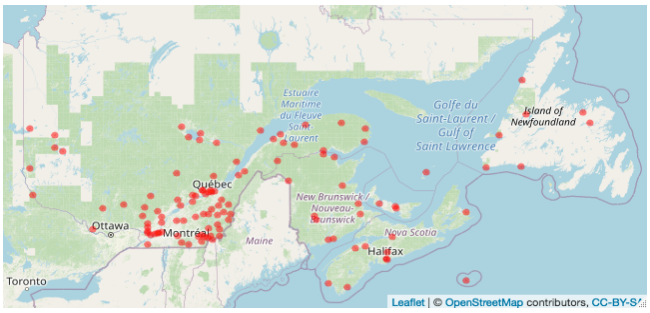

Figure 1: Locations of considered 116 stations in Central Eastern Canada. The latitude coordinates vary in the range $[43.71, 50.24]$ and the longitude ones in $[-79.23, -54.57]$.

As suggested by Figure 21, 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 indices 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 $n_f + n_v = n$. Here, we consider $n_f = 95$ and $n_v = 21$. Note 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 2. The sensibility of the considered statistical routines with respect to the choice of combinations of n_f and n_v stations is investigated in Appendix E.

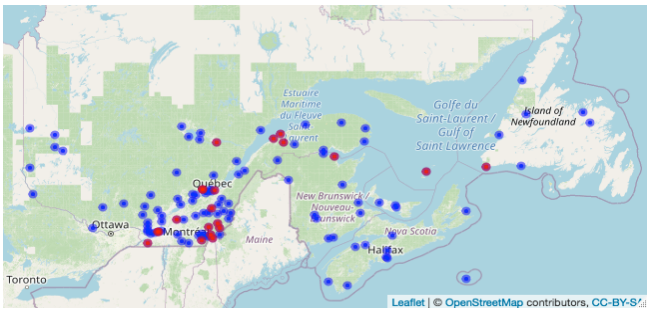

Figure 2: Spatial locations of one of the 200 considered combinations between fitting stations I_f (blue points) and validation stations I_v (red points) with $n_f = 95$ and $n_v = 21$.

In the following we 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 (see Appendix C) based on the L-moments estimators (see Appendix D).

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

In this section, we describe the components of the spatial smooth GEV model. We detail the GEV model used, the polynomial and spline models and the copula model used for the GEV parameters, which takes into account the peculiarity of the dataset presented in Section 2.

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)) and we do 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:

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

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 75% quantile precipitation over the 34-year period 1981-2014 of the Canadian Regional Climate Model (CRCM5) driven by the ERA-Interim reanalysis (see, *e.g.,* Dee et al. (2011)). From CRM5, variables are 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. We consider the coordinates of the centers of these cells and we attribute to each station in Figure 1 the 75% quantile precipitation value of the spatially nearest center-cell. More information on this climate reconstruction can be found, for instance, in Bresson et al. (2017).

3.2 Smooth models for GEV parameters

In this section, we consider the estimation of a spatial 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 (either μ, ξ or $\ln(\sigma)$) and $\tilde{\zeta}$ be the associated interpolated value. Consider the following general regression model associated to the function F

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

where $y_s^{(j)}$, for $j \in \{1, \ldots, r\}$ are covariates and ϵ_s is an error term. It is assumed that ϵ_s are independent and identically distributed by a normal distribution with mean 0 and constant variance. In what follows, the GEV parameter ζ at location s is modeled by (4) without the stochastic (Gaussian) contribution represented by ϵ_s. To corroborate the constant variance hypothesis, we provide a spatial map of $\hat{\epsilon}_s$ to check if in our specific dataset there exists some spatial pattern. The interested reader is referred to Figure 22 in Appendix E.

Polynomial regression model 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 75% quantile 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)). A simplified similar framework is considered in Padoan et al. (2010) where the parameters are modeled by linear regression using latitude, longitude and elevation as covariates.

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 with radial basis functions (see, *e.g.,* 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)). Here we take into consideration covariate models provided by polynomial regression as in (5) or spline-based regression as in (7) with longitude, latitude, altitude and 75% quantile precipitation as covariates. Table 1 gathers the considered covariate models for GEV parameters.

Table 1: Considered GEV covariate parameter models from polynomial Regression in (5) (PR model) and spline-based regression in (7) (spline model). The considered covariate are the three geographical coordinates (long, lat, alt), the 75% quantile precipitation (75% quantile prec) and the μ parameter.

	μ and ξ parameters	$\ln(\sigma)$ parameter
Considered models	Considered covariates	Considered covariates
PR model	• (long, lat, alt) • (long, lat, alt, 75% quantile prec)	• (long, lat, alt) • (long, lat, alt, 75% quantile prec) • (long, lat, alt, μ parameter)
Spline model	• (long, lat, alt) • (long, lat, alt, 75% quantile prec)	• (long, lat, alt) • (long, lat, alt, 75% quantile prec) • (long, lat, alt, μ parameter)

As mentioned above, polynomial regression models in Table 1 are polynomials of considered covariates with a maximum degree of 3. We take here all possible degree combinations (see details in Algorithm 5) and we select the best linear regression model with the help of AIC (see Akaike (1974)). Furthermore for the proposed spline-based regression model in Table 1 we consider 10000 combinations of 15 knots among $n_f = 95$ fitting stations and we select the best spline model via the generalized cross-validation (GCV) score (see details in Algorithm 6).
Table 1 allows us to limit the number of possible smooth GEV parameter models by considering a total of 4(models of μ) $\times$4(models of ξ) $\times$6(models of $\ln(\sigma)$)= 96 models combinations.

In Section 3.3 below, we integrate the response surfaces for modeling the GEV parameters gathered in Table 1 (see also Equations (5) and (7)) 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 (see for instance Blanchet and Lehning (2010)). Nevertheless, the aim of Section 3.3 will be to relax in a tractable way this hypothesis to build a more realistic spatial dependent setting.

3.3 Log-likelihood for the hierarchical copula-based model

In order to estimate the parameter models $\widetilde{\beta}_j$ 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 on multivariate hierarchical copula models; 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 (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 Equations* (5) *and* (7) *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 k^{th} 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 in the nested 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)

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

Obviously, in the spatial independence setting $\mathcal{L}^C(\Lambda)$ in (10) reduces to $\mathcal{L}^{\perp}(\Lambda)$ in (9). This spatial independent smooth GEV model $\mathcal{L}^{\perp}(\Lambda)$ in (9) was previously proposed and discussed by Blanchet and Lehning (2010).

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 k^{th} 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

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

the normalised spatial distance between the ith and jth stations.
Notice that the large proportion of missing values 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 (Hybrid empirical copula)**.** *Let i and j be fixed, with $i, j \in \{1, \ldots, n_f\}$. Consider the $2 \times T$ matrix composed of $(x_{(i,t)}, x_{(j,t)})^{\top}_{\{t=1,\ldots,T\}}$, 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*

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

with I_t^i *as in* (1)*. Let* $\widehat{\Lambda}_{LM}^i$ *the L-moments estimators of the GEV parameters relative of the* i*th station (see Appendix D for further details). Then, the hybrid empirical copula is defined*

$$\widehat{C}^{ij}(u,v) = \widehat{H}\left(q(u,\widehat{\Lambda}_{LM}^i), q(v,\widehat{\Lambda}_{LM}^j)\right), \text{ for } (u,v) \in [0,1]^2, \tag{12}$$

where $q(\cdot)$ *is as in* (3) *and*

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

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 (Empirical hybrid copula-based dissimilarity measure)**.** *We define the dissimilarity measure*

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

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* $f(0) = 0$*.*

Note 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. It seems that this choice $f(\cdot) = \exp(\cdot) - 1$ is the most convenient to highlight small differences among dissimilarity values, but clearly other functions f could be used provided that f is increasing with $f(0) = 0$.

In Algorithm 1, we detail the first part of the proposed inference procedure, *i.e.*, the estimation of the dissimilarity measure $\widehat{d}_{ij}$ in (14) for our dataset.

- First, we build the pseudo-observations of our spatio-temporal rainfall maxima, using the point-wise GEV quantiles with parameters estimated via $L-$moments (see steps 1-2 in Algorithm 1 and Appendix D).
- Second, we estimate the dissimilarity measure on these pseudo-observations via the hybrid copula presented in (12)-(13) (see steps 3-7 in Algorithm 1).

Algorithm 1 Proposed implementation of a copula-based dissimilarity measure

(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.

(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 i–th station, *i.e.*, $q(\cdot, \widehat{\Lambda}^i_{LM})$ and build the pseudo-observations for each station.

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

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

(Step 5) Choose the Crámer-von Mises L^2 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 identical observations in $(x_{(i,t)}, x_{(j,t)})^\top_{\{t=1,\dots,T\}}$.

(Step 7) Fix a threshold value $\overline{n}$.

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

if $n^c_f < \overline{n}$, the i–th and j–th 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.

In Algorithm 2, we detail the second part of the proposed inference procedure, *i.e.*, the maximization of the hierarchical log-likelihood $\mathcal{L}^C(\Lambda)$ in Equations (9)-(10) in terms of the GEV smooth surface parameters as in (5) and (7).

- For a fixed number of clusters K and β in (14) we provide the spatial clustering of stations, using the PAM algorithm (steps 1-3 in Algorithm 2).
- Then, we estimate the Archimedean copula density via the canonical maximum likelihood on the clustered pseudo-observations (see steps 4-5 in Algorithm 2).
- Finally, we maximize the hierarchical loglikelihood (see (9)- (10)) in terms of the GEV smooth surface parameters' *via* polynomial regression and spline models with the covariates presented in Table 1 (see step 6 in Algorithm 2).

A crucial step in the proposed construction is the optimal selection of parameters (K, β) (see step 7 in Algorithm 2). To this aim we introduce in Table 2 several well-known normalised scores, used in the following as selection criteria. The normalization is motivated by the spatial heterogeneity in extreme precipitation in our large region of interest. We measure via these scores the quality of the goodness-of-fit of the quantiles of the interpolated GEV parameters versus the observed ones. Let $z^{(1)}_{s_i}, \dots, z^{(m)}_{s_i}, \dots, z^{(M)}_{s_i}$ denote the $M = 30$ empirical quantiles at location s_i, where the mth value $z^{(m)}_{s_i}$ is associated to a probability $p_m = \frac{m-1/2}{M}$, for $m = 1, \dots, M$. Also, $\tilde{q}_{p_m, s_i}$ denotes the

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 the input parameter 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 for K clusters with $\widehat{d}_{ij}$ from **(Step 2)**.

(Step 4) Estimate the Archimedean copula density $c_{\widehat{\theta}_k}$ via canonical maximum likelihood on the pseudo-observations for $k \in \{1, \ldots, K\}$.

(Step 5) Select the best one in terms of the AIC criterion for each cluster with $k \in \{1, \ldots, K\}$.

(Step 6) Let $c_{\widehat{\theta}_k}$ for $k \in \{1, \ldots, K\}$ as in **(Step 5)**. Let define the marginal GEV smooth surfaces $\mu(s)$, $\xi(s)$ and $\ln(\sigma(s))$ as in Equations (5) and (7) with covariate models gathered in Table 1. Maximize the log-likelihood $\mathcal{L}^C(\Lambda)$ in Equation (10) with respect to the vector of parameters $\widetilde{\beta}$ in Equations (5) and (7).

(Step 7) Select (K^*, β^*) to minimize normalised scores in Table 2.

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.

$(1 - p_m)$ quantile of the interpolated GEV distribution at location s_i, obtained by Equation (3) replacing Λ by the interpolated values $\widetilde{\Lambda}$ with $p = p_m$.

Table 2: Considered normalised scores. Here $M = 30$, $\widetilde{n} = n_f$ for the fitting stations analysis and $\widetilde{n} = n_v$ for the validation one.

Normalised Root Mean Squared Error	$NRMSE = \frac{\sqrt{\frac{1}{\widetilde{n}\,M}\sum_{i=1}^{\widetilde{n}}\sum_{m=1}^{M}\left(z_{s_i}^{(m)} - \tilde{q}_{p_m,s_i}\right)^2}}{\frac{1}{\widetilde{n}\,M}\sum_{i=1}^{\widetilde{n}}\sum_{m=1}^{M} z_{s_i}^{(m)}}$
Normalised Mean Absolute Error	$NMAE = \frac{\sum_{i=1}^{\widetilde{n}}\sum_{m=1}^{M}\lvert z_{s_i}^{(m)} - \tilde{q}_{p_m,s_i}\rvert}{\sum_{i=1}^{\widetilde{n}}\sum_{m=1}^{M}\lvert z_{s_i}^{(m)}\rvert}$
Normalised Maximum Prediction Error	$NMPE = \frac{\max_{i\in\{1,\ldots,\widetilde{n}\}}\max_{m\in\{1,\ldots,M\}}\lvert z_{s_i}^{(m)} - \tilde{q}_{p_m,s_i}\rvert}{\max_{i\in\{1,\ldots,\widetilde{n}\}}\max_{m\in\{1,\ldots,M\}}\lvert z_{s_i}^{(m)}\rvert}$

The output of Algorithm 1 and Algorithm 2 is a vector $\widetilde{\beta}^*$, obtained via the score measures detailed in Table 2, of parameters $\widetilde{\beta}_j$ for the smooth functions in space of GEV parameters in Equations (5) and (7). Notice that, the choice of the hierarchical dependence copula model in Equation (10) directly impacts the obtained vector $\widetilde{\beta}^*$. Finally, by using $\widetilde{\beta}^*$, we can easily construct the associated precipitation return level maps (see Figure 8). In Appendix A we investigate the performance of Algorithm 1 (steps 3-7) and Algorithm 2 (steps 1-3) in several simulated data-set. Appendix B is devoted to a simulation study to evaluate our inference procedure for the final return level maps and compare it to alternative methods.

4 Results on considered Central Eastern Canada data-set

Behaviour of Algorithm 1 and Algorithm 2

In this section, we illustrate the behaviour of Algorithm 1 and Algorithm 2 on the Central Eastern Canada data-set previously displayed in Figure 2. We run our estimation with $\overline{n} = 10$ (see step 7 of Algorithm 1). The output of Algorithm 1 and Algorithm 2 for several values of β and K is gathered in Figure 3 in terms of the normalised maximum prediction error in Table 2.

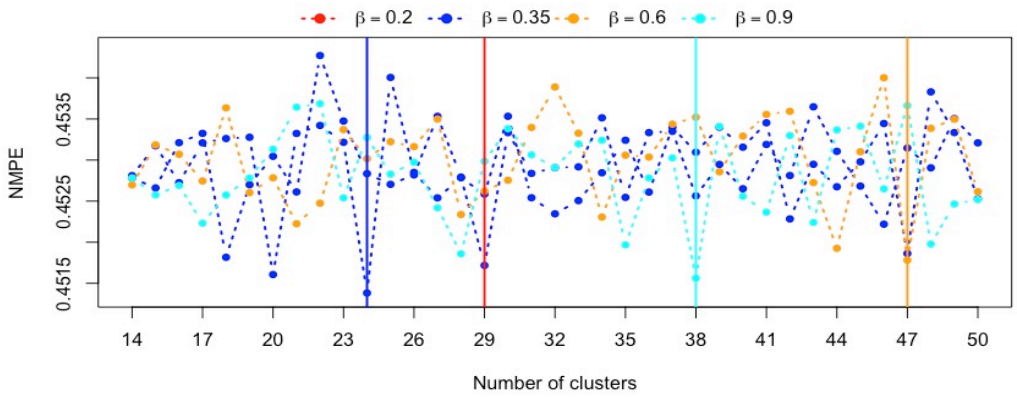

Figure 3: Normalised MPE versus K (numbers of clusters) as in Table 2 for $\beta \in \{0.2, 0.35, 0.6, 0.9\}$. We consider GEV covariate parameter models with three geographical coordinates (long, lat, alt) and 75% quantile covariates (see Table 1).

By step 7 in Algorithm 2, this permits to select $K^* = 24$ and $\beta^* = 0.35$. We underline that we analysed Algorithm 1 and Algorithm 2 for a large range of possible values for β. In Figure 3 we only display some representative small, intermediate and large values of β for the sake of readability.

In Figure 4, we display the obtained absolute value of the logarithm scale of $\widehat{d}_{ij}$ in (14) with $\beta^* = 0.35$. In addition, in Figure 5, we fix 3 fitting stations (black dots) and we display the obtained $\widehat{d}_{ij}$ in (14) with $\beta^* = 0.35$ of these stations with respect to all others fitting stations. 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 between stations but also the copula dependence structure between involved time-series ($\beta^* = 0.35$ in the convex combination in (14)).

Finally, we display some results from steps 4 and 5 in Algorithm 2 for 3 among $K^* = 24$ obtained clusters (see Figure 6). 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. Obtained Archimedean copulas for each cluster are Clayton with $\hat{\theta} = 0.55$ (first panel) and Gumbel with $\hat{\theta} = 1.396$ (second panel). Remark that the considered Archimedean families describe three different situations in terms of tail properties: asymptotic independence in the lower and upper tails (Frank copula), asymptotic dependence in the lower tails (Clayton copula), and asymptotic dependence in the upper tails (Gumbel and Joe copulas).

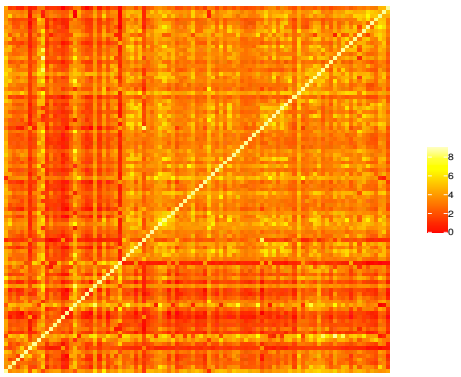

Figure 4: Absolute value of dissimilarity measure in (14) in logarithm scale, *i.e.*, abs(ln($\widehat{d}_{ij}$)), with $\beta^* = 0.35$ for each pair of fitting stations (see blue stations in Figure 2).

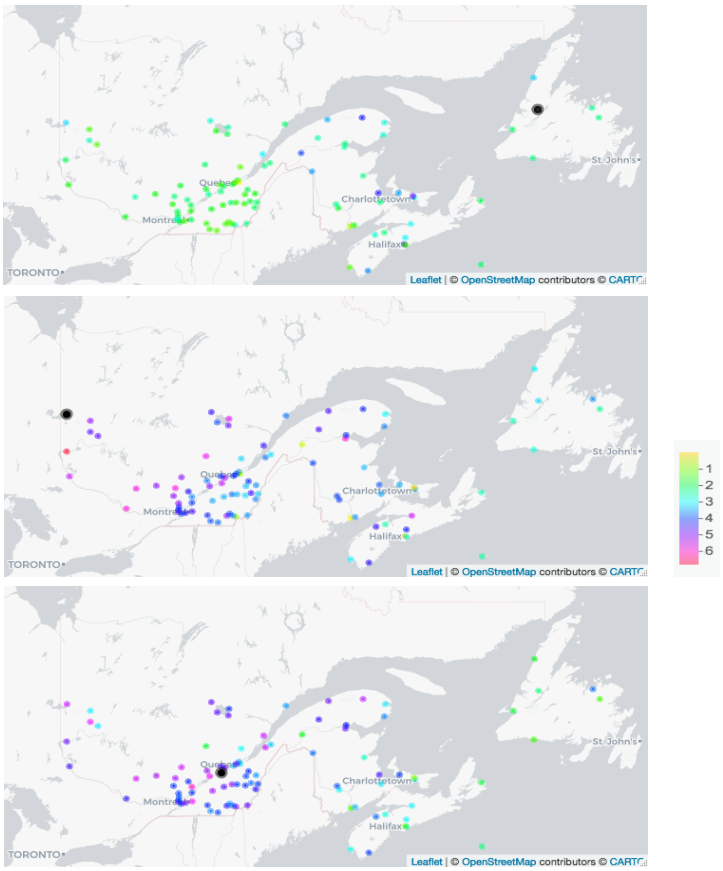

Figure 5: Absolute value of dissimilarity measure in (14) in logarithm scale, *i.e.*, abs(ln($\widehat{d}_{ij}$)), with $\beta^* = 0.35$ for three fixed stations (black dots) with respect to the others fitting stations.

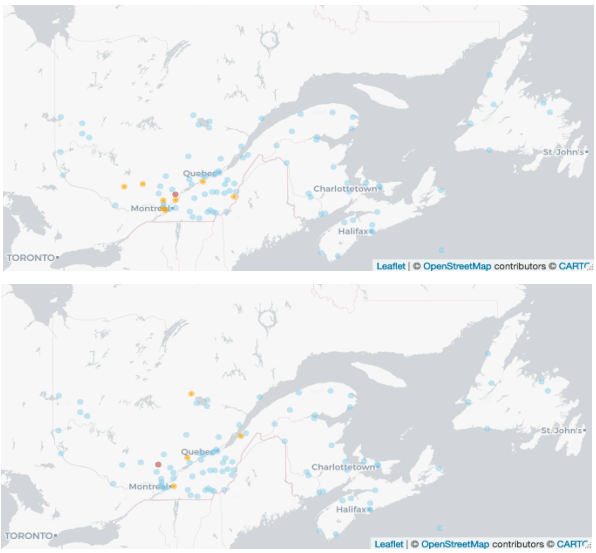

Figure 6: Illustration for two obtained clusters. Centroid stations are presented by dark-red dots, element stations in each cluster by orange ones. Obtained Archimedean copulas for each cluster are Clayton with $\hat{\theta} = 0.55$ (first panel) and Gumbel with $\hat{\theta} = 1.396$ (second panel).

Quality of predictions and comparison with other methods

By using the specific parameters selection presented in the previous section, we provide in Table 3 the normalised error measures for the three geographical coordinates (long, lat, alt) and the 75% quantile precipitation (75% quantile prec) covariate parameter models (see Table 1). Furthermore in Table 4, we compare the performance of the proposed method with some classical techniques of spatial interpolation of individual GEV distributions. The interested reader is referred to Appendix C for a short overview of these methods. In Appendix E we also calculate the corresponding scores for 200 combinations of fitting and validation stations in order to test the sensibility of the considered statistical routines with respect to the choice of combinations of n_f and n_v stations.

Table 3: Normalised goodness-of-fit scores from Table 2 for fitting and validation stations displayed in Figure 2 by using the smooth copula-based GEV model proposed in Section 3, for $K^* = 24$ and $\beta^* = 0.35$. We consider here the three geographical coordinates and the 75% quantile precipitation as GEV covariate parameter models.

Considered covariates	Fitting stations			Validation stations		
	NRMSE	NMAE	NMPE	NRMSE	NMAE	NMPE
(long, lat, alt, 75% quantile prec)	0.13	0.08	0.45	0.11	0.08	0.24

Finally we compare in Table 5 the performance of the spatial surfaces for GEV parameters constructed by the latent variable model for Gaussian Process-based simulation response surface modeling in Zhang et al. (2020) and the hierarchical max-stable spatial model in Reich and Shaby (2012). Notice that a preliminary madogram extremal coefficient estimation associated to our rainfall data clearly shows spatial dependence. The latent Gaussian model and the hierarchical

Table 4: Normalised goodness-of-fit scores from Table 2 for 95 fitting and 21 validation stations displayed in Figure 2 by using classical techniques of spatial interpolation of individual GEV distributions described in Appendix C. We consider here the three geographical coordinates and the 75% quantile precipitation as GEV covariate parameter models.

	Fitting stations			Validation stations		
	NRMSE	NMAE	NMPE	NRMSE	NMAE	NMPE
1. IDW	0.06	0.04	0.17	0.12	0.09	0.24
2. Polynomial regression	0.11	0.07	0.32	0.14	0.10	0.38
3. Spline	0.12	0.08	0.40	0.10	0.07	0.24
4. Kriging	0.06	0.04	0.17	0.12	0.09	0.25

max-stable one are implemented by using `R` functions `LVGP_fit` and `hkevp.fit` in `R` packages `LVGP` and `hkevp`, respectively (see Table 5). The training inference is based on the considered $n_f = 95$ fitting stations and the testing results are evaluated on the $n_v = 21$ validation stations (see Figure 2). The `R` code associated to Table 5 can be found in the `scoreslatentmaxstable.R` file in the supplementary material CodeR folder.

Table 5: Normalised goodness-of-fit scores from Table 2 for 21 validation stations displayed in Figure 2 by using the Gaussian latent model in Zhang et al. (2020) and the hierarchical max-stable model in Reich and Shaby (2012). We consider here the three geographical coordinates and the 75% quantile precipitation as GEV covariate parameter models.

	Validation stations		
	NRMSE	NMAE	NMPE
1. Latent model	0.47	0.37	0.85
2. Hierarchical max-stable spatial model	0.52	0.43	0.83

In Tables 3-5, we illustrated the goodness-of-fit in a global spatial sense. With respect the classical techniques, we can observe in Table 4 that our proposed method provides similar goodness-of-fit as polynomial regression and spline. Furthermore, our method presents a better global performance in comparison with the Gaussian latent model in Zhang et al. (2020) and the hierarchical max-stable model in Reich and Shaby (2012) (see Table 5).

Having a detailed look at the goodness-of-fit for individual station is also interesting. In Figure 7 we show the QQ-plots for two fitting stations (first row) and two validation ones (second row) with spline interpolation method detailed in Algorithm 6 (see Appendix C) (orange points) and proposed smooth copula model (green stars). We also display the latent Gaussian model (blue stars) and the hierarchical max-stable model (red stars) prediction results (Figure 7, second row). The three geographical coordinates and the 75% quantile precipitation are used here as GEV covariate parameter models. By observing Figure 7, our smooth copula-based GEV model seems perform as well as the spline one for instance in station 61 and slightly better in stations 3 and 107 et 55, in particular for extreme quantiles. The extremal fitting of our method performs slightly better than to the latent Gaussian model (blue stars) and the hierarchical max-stable model (red

stars) in stations 55 and 61.

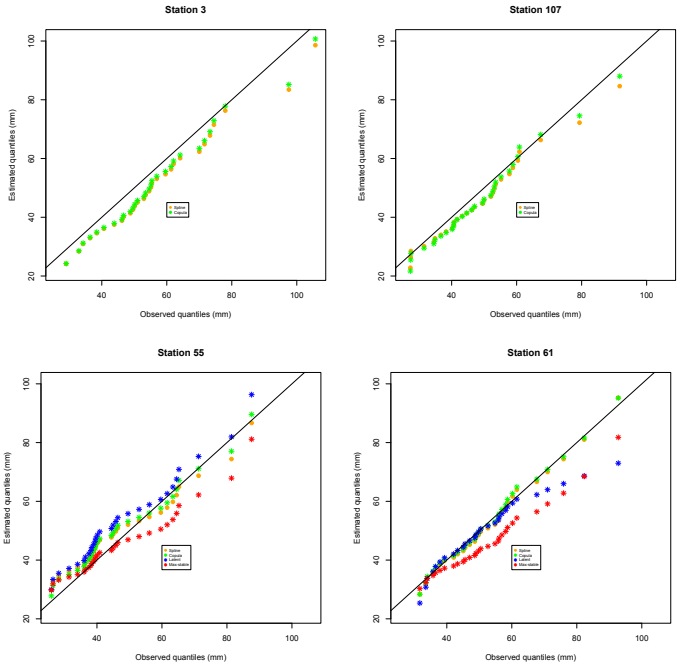

Figure 7: QQ-plots for two fitting stations (first row) and two validation ones (second row) by using spline interpolation method in Algorithm 6 (orange points), proposed smooth copula model (green stars), the latent Gaussian model (blue stars), the hierarchical max-stable model (red stars). The three geographical coordinates and the 75% quantile precipitation are used here as GEV covariate parameter models. The coordinates of station 3 are $(47.983, -66.333)$, of station 107 $(48.516, -72.266)$, of station 55 $(45.600, -70.866)$ and of station 61 (45.500, -73.583).

Due to the limited data-set, a crucial point in our analysis is to quantify estimation uncertainty. Firstly we focus on the uncertainty in Algorithm 1. In our copula-based model, from Algorithm 1 (steps 1-2 and 7), it is possible to quantify the uncertainty given by central limit theorems of L-moments (see, *e.g.*, Hosking and Wallis (1997) and Section 3.6.1 in de Haan and Ferreira (2006)) and from asymptotic results of hybrid copula (see, *e.g.*, Segers (2015) and Boulin et al. (2022)). Delta method and continuous-mapping techniques provide the final confidence intervals. As previously remarked by Blanchet and Lehning (2010) in the case of the simplified independent smooth GEV model $\mathcal{L}^{\perp}(\Lambda)$ in (9), this additional information regarding model uncertainty is an important advantage. For more details the interested reader is referred to Section 6.2 in Blanchet and Lehning (2010).

Secondly in Algorithm 2 (steps 5-6), we have to additionally consider the uncertainty propagated through the classical MLE estimation of the hierarchical Archimedean copula model. The interested reader is referred also to Johannesson et al. (2021) where the propagation of uncertainty in a recent two-steps inference approach is analysed.

Obtained precipitation return level maps

Associated return level maps obtained via for IDW, kriging, spline and smooth copula-based GEV models in Tables 3-5 are spatially displayed in Figure 8. One can appreciate 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 (fourth panels), due to the considered local spatial ($\widetilde{s}_{ij}$) and copula dependencies ($\widehat{C}_{ij}$) in (14). As already remarked locally in Figure 7, for stations 55 and 61, in Figure 8 we can appreciated the global ability of the smooth copula-based GEV to model the spatial dependence for eventually high return levels in comparison with other techniques. Furthermore, remark that this smooth return level maps can be computed from the fitted model without any further interpolation. Finally, one can compare return levels in Figure 8 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 (left column, fourth row in Figure 8).

5 Conclusion

The statistical methodology presented in this paper, based on a hybrid hierarchical smooth GEV copula model, can be useful to infer on any location in the area of study. It takes into account the spatio-temporal structure, using the distance between dependent clusters, by means of a dissimilarity measure designed to handle missing data. Note that other papers tackle the problem of missing data, using hierarchical max-stable process construction such as in Reich and Shaby (2012) or conditioning of a latent process in the context of extremes, such as in Zhang et al. (2021).

Advanced approaches for estimating the GEV parameter surfaces have recently been introduced in the literature, such as the Max-and-Smooth approach of Johannesson et al. (2021) for flood frequency data or the approach of Sass et al. (2021) based on the spatial GEV by introducing fused lasso and fused ridge penalty for parameter regularization. The hybrid hierarchical smooth GEV copula model allows the direct mapping of quantiles by dealing simultaneously with the non-concomitant record periods between recording stations.

An interesting future extension of the proposed method is the quantification of the impact of the portion of missing values and non-concomitant record periods at different stations on the final performance of Algorithm 1 et Algorithm 2. This crucial point can be explored via the study of the variance of the empirical hybrid copula-based dissimilarity measure $\widehat{d}_{ij}$ in (14). Some useful preliminary results on this issue are provided by Segers (2015) and recently by Boulin et al. (2022). Then a comparison on the ability to handle missingness would be an interesting future work.

Acknowledgements: The authors express their gratitude to two anonymous Referees and Associate Editor for their valuable comments on this article. The authors thank Jonathan Jalbert (Polytechnique Montréal, Canada) for useful discussions. Funding in 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

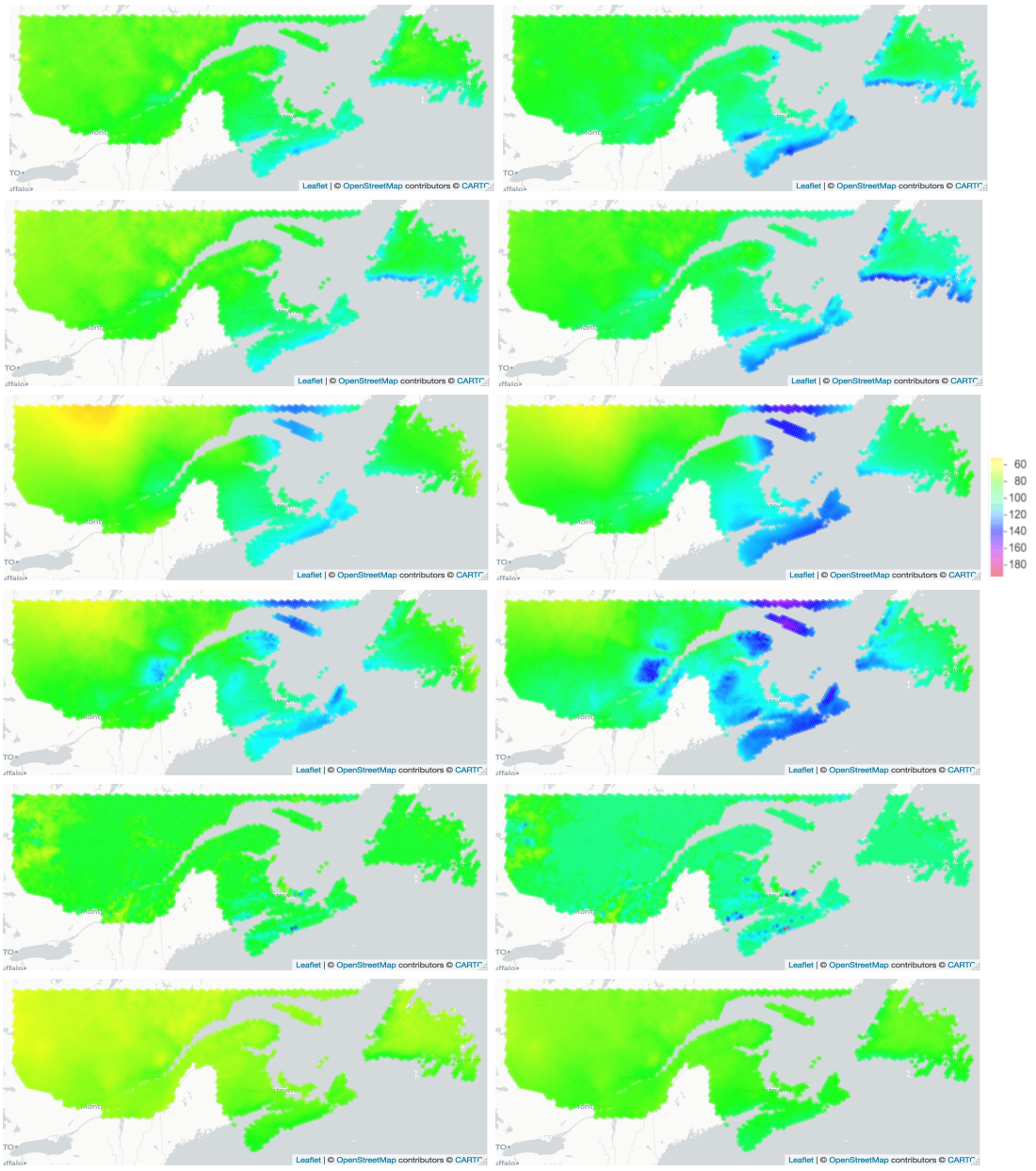

Figure 8: Obtained 20-years (left column) and 40-years (right column) precipitation return level maps in mm in Central Eastern Canada. Models: IDW (first row), kriging (second row), spline (third row), smooth copula-based GEV (fourth row), latent variable model for Gaussian Process-based simulation response surface modeling (fifth row) and hierarchical max-stablespatial model (sixth row).

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.

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.

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.

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

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.

Boulin, A., Di Bernardino, E., Laloë, T., and Toulemonde, G. (2022). Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework. *Journal of Multivariate Analysis*, 192.

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.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical science*, 27(2):161–186.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory. An Introduction.* Springer Series in Operations Research and Financial Engineering. Springer: New York.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hersbach, H., Hólm, E., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A., Monge-Sanz, B., Morcrette, J., Park, B., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J., and Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137:553–597.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. *Bulletins de l'Académie Royale de Belgique*, 65(1):274–292.

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

Disegna, M., D'Urso, 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.

Huser, R., Dombry, C., Ribatet, M., and Genton, M. G. (2019). Full-likelihood inference for max-stable processes. *Stat*, 8(1):e218.

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.

Johannesson, A. V., Siegert, S., Huser, R., Bakka, H., and Hrafnkelsson, B. (2021). Approximate Bayesian inference for analysis of spatio-temporal flood frequency data. *Annals of Applied Statistics*, (to appear).

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.

Lehmann, E. A., Phatak, A., Stephenson, A., and Lau, R. (2016). Spatial modelling framework for the characterisation of rainfall extremes at different durations and under climate change. *Environmetrics*, 27(4):239–251.

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., 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)*.

Perreault, L., Jalbert, J., and Genest, C. (2022). Interpolation of precipitation extremes on a large domain toward idf curve construction at unmonitored locations. *Journal of Agricultural, Biological and Environmental Statistics*, 27.

Reich, B. J. and Shaby, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *The Annals of Applied Statistics*, 6(4):1430–1451.

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.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22:813–845.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields. Theory and Applications*. Chapman and Hall/CRC.

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

Rüschendorf, L. (1976). Asymptotic distributions of multivariate rank order statistics. *Ann. Statist.*, 4(5):912 – 923.

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.

Sang, H. and Gelfand, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16:407–426.

Sang, H. and Gelfand, A. E. (2010). Continuous spatial process models for spatial extreme values. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(1):49–65.

Sass, D., Li, B., and Reich, B. J. (2021). Flexible and fast spatial return level estimation via a spatially fused penalty. *Journal of Computational and Graphical Statistics*, 30(4):1124–1142.

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.

Smith, R. L. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

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.

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.

Zhang, L., Shaby, B. A., and Wadsworth, J. L. (2021). Hierarchical transformed scale mixtures for flexible modeling of spatial extremes on datasets with many locations. *Journal of the American Statistical Association*, 0(0):1–13.

Zhang, Y., Tao, S., Chen, W., and Apley, D. (2020). A latent variable approach to Gaussian process modeling with qualitative and quantitative factors. *Technometrics*, 62(3):291–302.

A Simulation studies for $\widehat{d}_{ij}$ and associated clustering algorithm

In this section we illustrate the performance of Algorithm 1 and Algorithm 2 in several simulated datasets. In this study, we focus on the complete data setting, *i.e.*, without missing data. In this case, Equation (13) is then equal to the classical empirical distribution function (see, *e.g.*, Rüschendorf (1976), Deheuvels (1979)). We reproduce here 3 possible scenarios: a situation characterized by spatial well distinguished clusters (Model M1), by spatial overlapping clusters (Model M2) and by a balance between Model M1 and Model M2 (Model M3).

We generate $K = 5$ clusters for models in Models M1 and M2 and $K = 10$ in Model M3. From Definition 3.3, recall that I_k is the set of station indices belonging to the k^{th} cluster with $d_k = \text{card}(I_k)$, the size of k^{th} cluster. Furthermore here $\cup_{k=1}^{K} I_k = 80$ stations and $I_k \cap I_{k'} = \emptyset$, $\forall\, k \neq k'$. Denote by C_{θ_k} the d_k-dimensional Archimedean copula with dependence parameter θ_k for the k^{th} cluster related to model in (8).

To generate the k^{th} cluster and the associated dependent time-series of pseudo-observations of length $T = 40$, we implement the simulation procedure gathered in the following Algorithm 3.

Algorithm 3 Simulation procedure for spatial clusters and dependent time-series

(Step 1) For each $k \in \{1, \ldots, K\}$, consider d_k the size of the k^{th} cluster.

(Step 2) The two spatial coordinates $\mathbf{m}_k$ of k^{th} centroid are simulated uniformly in a consider spatial domain.

(Step 3) Define Σ_k the variance-covariance matrix relative to the k^{th} cluster.

(Step 4) The two spatial coordinates of station $\mathbf{s}_i^k$ associated to the centroid $\mathbf{m}_k$ are generated via bivariate gaussian vector, *i.e.*, $\mathbf{s}_i^k \sim \mathcal{N}_2(\mathbf{m}_k, \Sigma_k)$, for $i \in \{1, \ldots, d_k\}$.

(Step 5) Generate time-series of pseudo-observations from C_{θ_k} (the d_k-dimensional Archimedean copula associated to the k^{th} cluster) of length $T = 40$.

Model M1: predominant spatial information In this first study, we implement Algorithm 3 by considering clearly spatially separated cluster locations (see Figure 9).

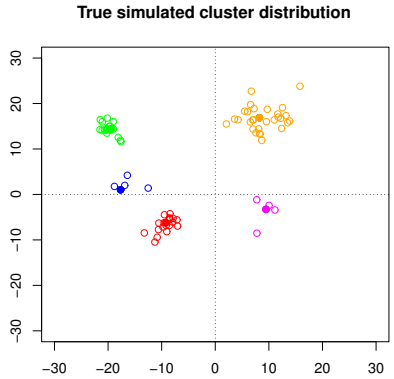

Figure 9: Random generation of Model M1, with a predominant spatial dissimilarity between clusters. Here $K = 5$. The cluster dimensions are $d_1 = d_2 = 20$, $d_3 = 30$, $d_4 = d_5 = 5$.

Furthermore, we generate the time-series of pseudo-observations by using the following copula models:

Cluster label	Archimedean copula	θ_k	d_k	Locations color in Figure 9
1	Frank	2	20	red
2	Gumbel	5	20	green
3	Gumbel	1.5	30	orange
4	Independent	–	5	blue
5	Frank	12	5	magenta

In this setting, we expect a small weight β in the convex combination in the proposed empirical copula-based dissimilarity measure in (14). This is exactly what we observe in Figure 10. Left column of Figure 10 displays the ASW criterion in our PAM algorithm with dissimilarity in (14) for $\beta = 0.01$ (first row) and $\beta = 0.25$ (second row). For more details about definition and interpretation of Average Silhouette Width (ASW) criterion in PAM algorithm, the reader is directed to Chapter 2 in Kaufman and Rousseeuw (1990). For each β, the associated cluster labeling of stations is represented in the right column of Figure 10. In particular, the PAM algorithm with $\beta = 0.01$ perfectly identifies $K = 5$ clusters and the correct cluster labeling of all 80 stations.

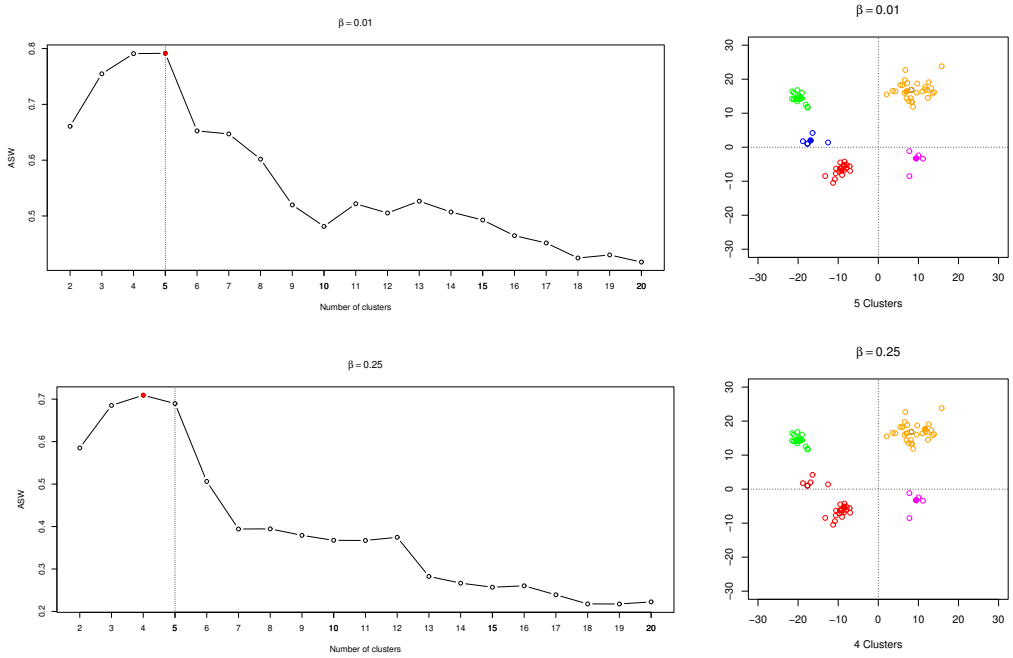

Figure 10: Model M1 with true clusters spatial distribution as in Figure 9. **Left column**: ASW criterion for two values of β in (14). The true value $K = 5$ is displayed in dashed vertical line. **Right column**: associated cluster labeling of the considered 80 stations.

As expected, when β increases the 1^{st} and 4^{th} clusters (red and blue points, respectively) are merged into an unique cluster (see the second row of Figure 10). Indeed in this case, the importance of time-series dependence, increases in the convex combination. Roughly speaking, time-series exhibiting similar dependence structures, in particular in extremes, *i.e.*, Frank copula with parameter 2 and independent copula, are clustered together.

Model M2: predominant dependence information In this second study, we implement Algorithm 3 by considering spatially overlapping clusters (see Figure 11). Furthermore, we generate the time-series of pseudo-observations by using the same d_k-dimensional Archimedean copulas C_{θ_k} as for the previous table of Model M1.

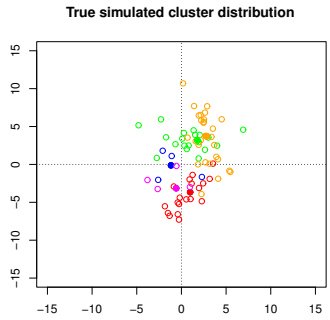

Figure 11: Random generation of Model M2 with spatially overlapping clusters. Here $K = 5$. The cluster dimensions are $d_1 = d_2 = 20$, $d_3 = 30$, $d_4 = d_5 = 5$.

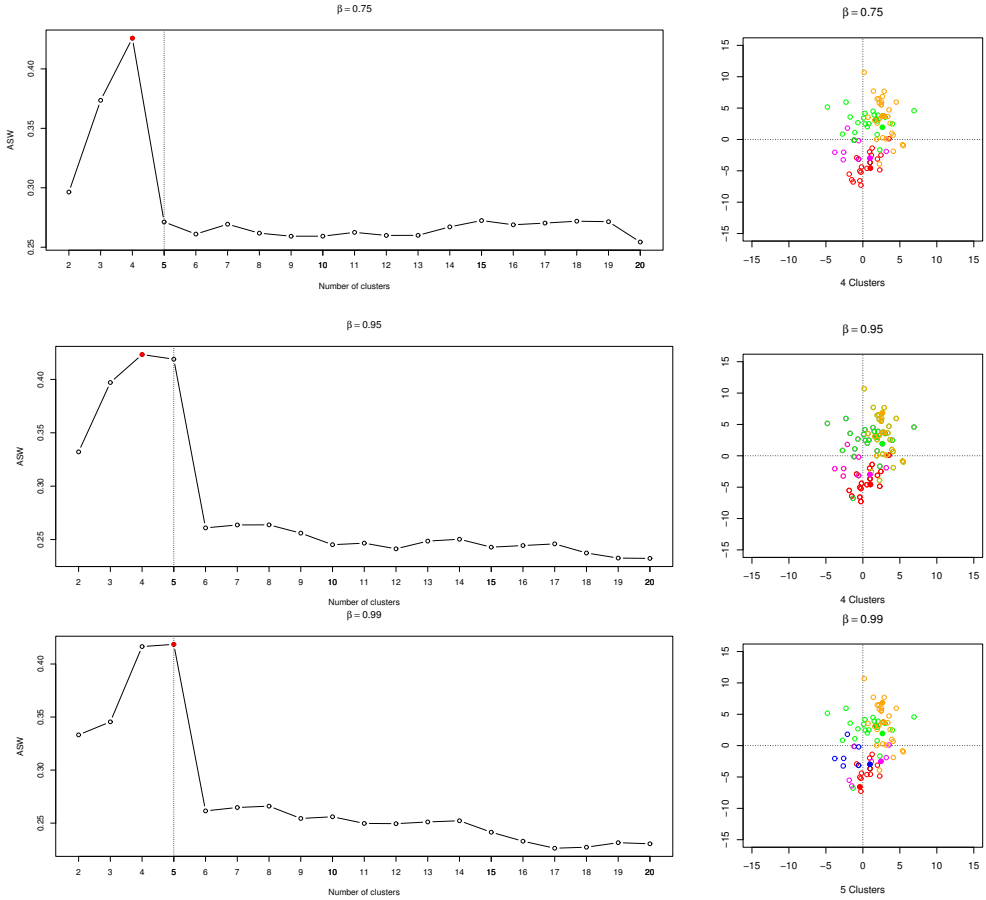

Figure 12: Model M2 with true clusters spatial distribution as in Figure 11. **Left column**: ASW criterion for several values of β in Equation (14). The true value $K = 5$ is displayed in dashed vertical line. **Right column**: associated cluster labeling of the considered 80 stations.

Conversely to the previous Model M1, we expect here $\beta \approx 1$ in Equation (14) in order to build clusters exclusively via the dependence structure criterion. Figure 12 displays the ASW criterion for $\beta \in \{0.75, 0.95, 0.99\}$ (left column) and the corresponding cluster labeling of stations (right column). In particular, the PAM algorithm with $\beta = 0.99$ is able to identify $K = 5$ clusters and the correct cluster labeling of 90% of stations with respect to the true distribution in Figure 11.

Model M3: a balance between spatial and dependence information In this last study, we implement Algorithm 3 by considering a mixture between spatial and dependence information (see Figure 13).

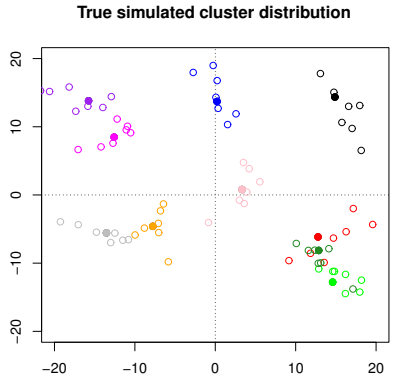

Figure 13: Random generation of Model M3 with a mixture between spatial and dependence information. Here $K = 10$. The cluster dimensions are $d_k = 8$, for $1 \leq k \leq K$.

Furthermore, we generate the time-series of pseudo-observations by using the following copula models:

Cluster label	Archimedean copula	θ_k	d_k	Locations color in Figure 13
1	Frank	2	8	red
2	Gumbel	5	8	green
3	Gumbel	1.5	8	orange
4	Independent	–	8	blue
5	Frank	12	8	magenta
6	Joe	2	8	pink
7	Clayton	10	8	black
8	Clayton	2	8	forestgreen
9	Joe	4	8	purple
10	Joe	5	8	gray

For small values of β only 6 clusters are identified essentially by using the spatial proximity (see first row of Figure 14). By increasing β new clusters appear. In particular with $\beta = 0.25$ the algorithm is able to identify two very different copula-structures: Joe copula (purple points) and Frank one (magenta points). The value $\beta = 0.5$ allows us to correctly identify the number of clusters ($K = 10$) and the correct cluster labeling of 79 stations with respect to the true distribution in Figure 13 (*i.e.*, 98.75% of stations). If β becomes too large the algorithm produces uncorrected clusters labeling (see third row in Figure 14) or artificial small clusters (see fourth and fifth rows in Figure 14).

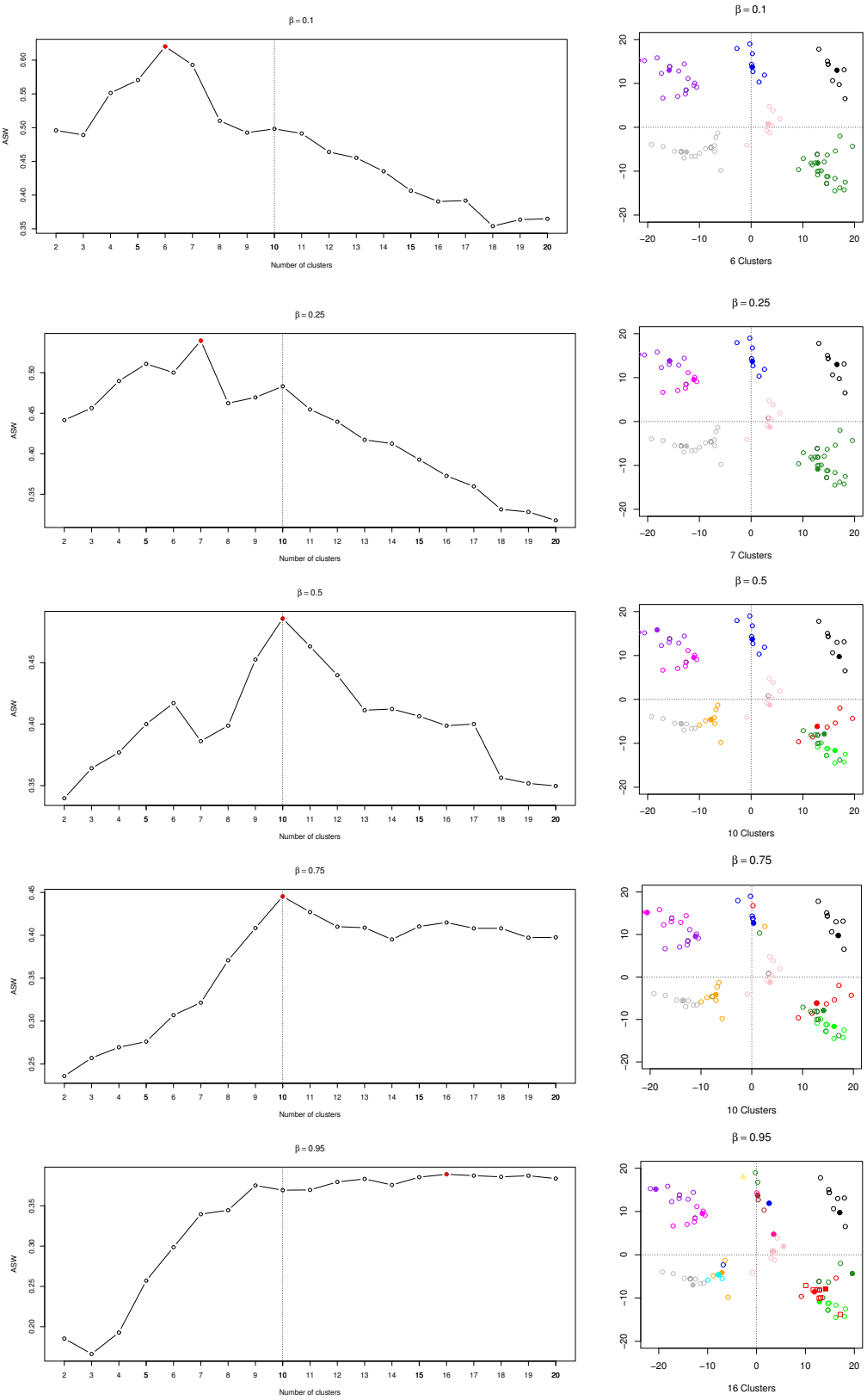

Figure 14: Model M3 with true clusters spatial distribution as in Figure 13. **Left column**: ASW criterion for several values of β in Equation (14). The true value $K = 10$ is displayed in dashed vertical line. **Right column**: associated cluster labeling of the considered 80 stations.

B Simulation studies for the estimation of the final return level maps

In order to evaluate the performance of the proposed spatial smooth GEV copula-model for the final return level maps, a simulation study is presented in this section. As in Appendix A, we take into account the complete data setting, *i.e.*, without missing data. Firstly, we construct the true simulated dataset where the dependent structure between stations and the spatial surfaces of GEV parameters are known. To this end, we start with the construction of the hierarchical copula model in (8) with $K = 7$ clusters of 116 stations located as in Figure 15.

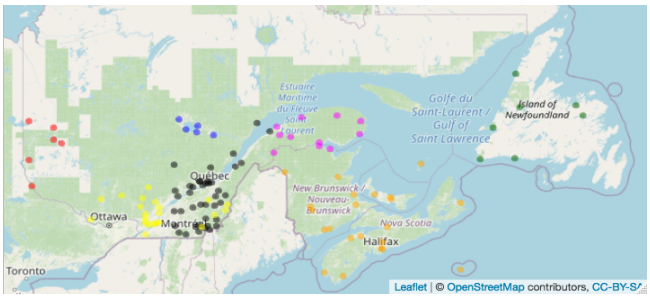

Figure 15: Simulated clusters via the hierarchical copula model in (8) for the 116 stations.

Then, we generate time-series of pseudo-observations of length $T = 60$ from the d_k-dimensional Archimedean copulas C_{θ_k} as described in Table 6.

Table 6: Generated d_k-dimensional Archimedean copulas for spatial clusters in Figure 15 and associated estimated ones (last two columns).

Cluster label	Color in Figure 15	true copula C_{θ_k}	d_k	θ_k	estimated copula $C_{\hat{\theta}_k}$	$\hat{\theta}_k$
1	red	Joe	6	4	Joe	4.15
2	blue	Frank	5	5	Frank	5.58
3	yellow	Gumbel	24	2	Frank	3.83
4	black	Frank	41	1	Frank	1.16
5	magenta	Joe	11	3	Joe	3.27
6	orange	Gumbel	22	3	Gumbel	3.03
7	green	Joe	7	2.5	Joe	2.47

Furthermore, we generate the GEV parameters surfaces by simulated realizations of Gaussian random fields and we get the simulated dataset by applying the GEV quantile in (3) to the previous time-series pseudo-observations of the previous hierarchical copula model. Using this procedure, we simulate flexible spatial surfaces for GEV parameters with nested copula spatial dependence structure. Then, we can easily calculate the true return level maps associated to these GEV

surface parameters (see Figure 18, first panel).

Now, we implement Algorithm 1 and Algorithm 2 (Steps 1-3) for different values of β and K. Particularly, we take $K \in \{1, \ldots, 7\}$ and $\beta \in \{0.7, 0.8, 0.9, 0.99\}$. In Figure 16, we present the classical ASW criterion of PAM method from Algorithm 2 (Step 3). We observe in Figure 16 that the maximum ASW value is reached for $K^* = 7$ and $\beta^* = 0.7$. Notice that this choice of K and β parameters allow us to a correct cluster covering of 100% of stations with respect to the true cluster model in Figure 15.

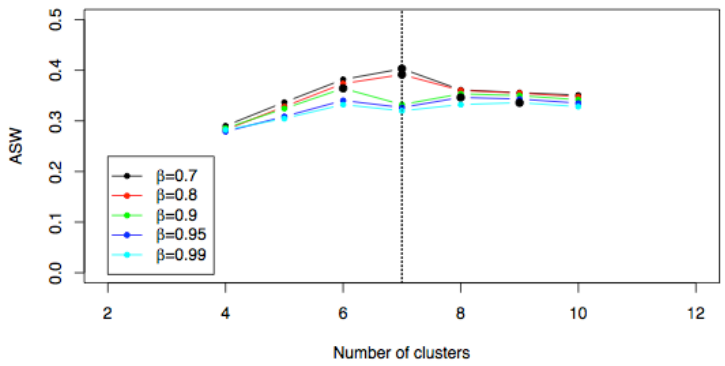

Figure 16: ASW criterion for the true simulated dataset by considering $K \in \{1, \ldots, 7\}$ and $\beta \in \{0.7, 0.8, 0.9, 0.99\}$.

By using this clustering, we estimated by canonical maximum likelihood, the spatial copulas model and associated parameters $\widehat{\theta}_k$. Results are gathered in Table 6 (see Algorithm 2, Step 4).

In Table 7 we gathered the normalised scores in Table 2, evaluated by using the true generated quantiles, for the $n_v = 21$ validation stations (see Figure 2). The fitting procedure is implemented on the $n_f = 95$ stations. The three geographical coordinates and the 75% quantile precipitation are used here as GEV covariate parameter models (see Table 1). These normalised scores are evaluated as in Table 2 by using the true generated quantiles.
Firstly, as lower bounds, we present the scores obtained by replacing $\tilde{q}_{p_m,s_i}$ by the L-moments quantile estimators in Appendix D at level probability p_m for each testing station s_i (see *Pointwise GEV* row in Table 7). In the second line, we display the scores associated to the proposed smooth copula-based model. The last two lines are devoted to two alternative methods (already present in Section 4), *i.e.*, the latent variable model for Gaussian Process-based simulation response surface modeling in Zhang et al. (2020) and the hierarchical max-stable spatial model in Reich and Shaby (2012), respectively. The performance on the 21 validation stations, evaluated via NRMSE, NMAE and NMPE in Table 7, seems to be similar for the considered methods. However, if we focus on the local behaviour, by analysing the individual QQ-plots for the validation stations in Figure 17, the tail fitting is realised slightly better by smooth copula-based GEV model.

Table 7: Normalised goodness-of-fit scores from Table 2 for 21 validation stations displayed in Figure 2 by using the L-moments estimators, the smooth copula-based GEV model, the Gaussian latent model in Zhang et al. (2020) and the hierarchical max-stable model in Reich and Shaby (2012). We consider here the three geographical coordinates and the 75% quantile precipitation as GEV covariate parameter models.

	NRMSE	NMAE	NMPE
1. Pointwise GEV	0.07	0.04	0.12
2. Smooth copula-based model	0.28	0.14	0.64
3. Latent process	0.26	0.11	0.66
4. Hierarchical max-stable spatial model	0.37	0.22	0.71

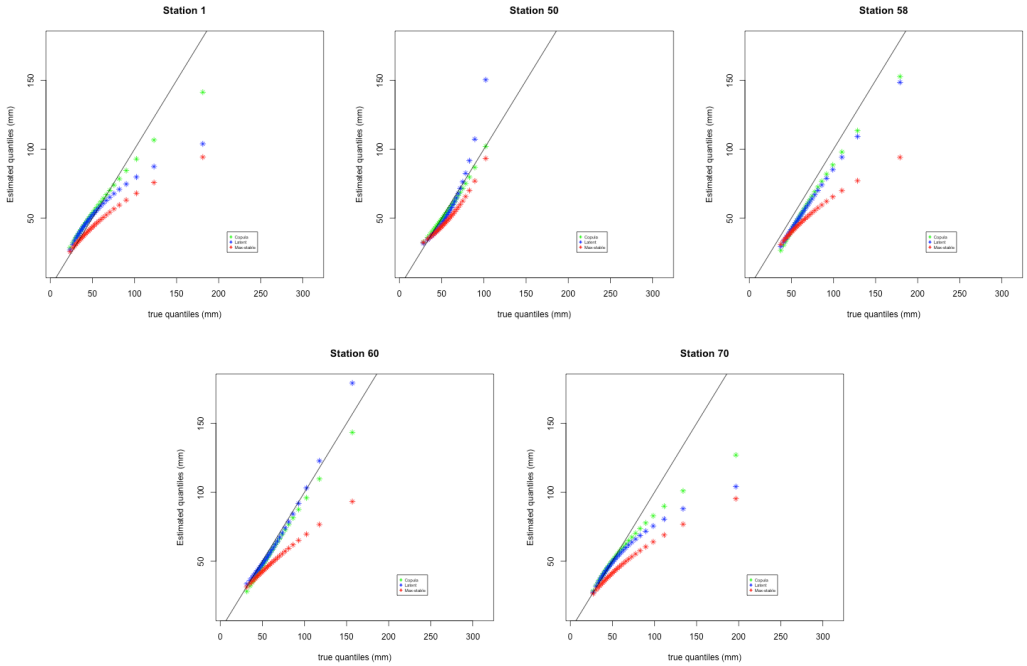

Figure 17: QQ-plots for the five validation stations plotted in Figure 2 by using smooth copula-based model (green stars), the latent Gaussian model (blue stars) and the hierarchical max-stable model (red stars). The coordinates of station 1 are $(47.900, -65.833)$, station 50 $(45.283, -71.200)$, station 58 $(45.633, -71.367)$, station 60 $(45.500, -73.617)$ and station 70 $(46.250, -71.217)$.

Finally the global spatial return level maps displayed in Figure 18 provided by the smooth copula-based GEV model (second row in Figure 18) reproduces in a more adequate way the true return level map in the first row of Figure 18 with respect to the considered competitor methods.

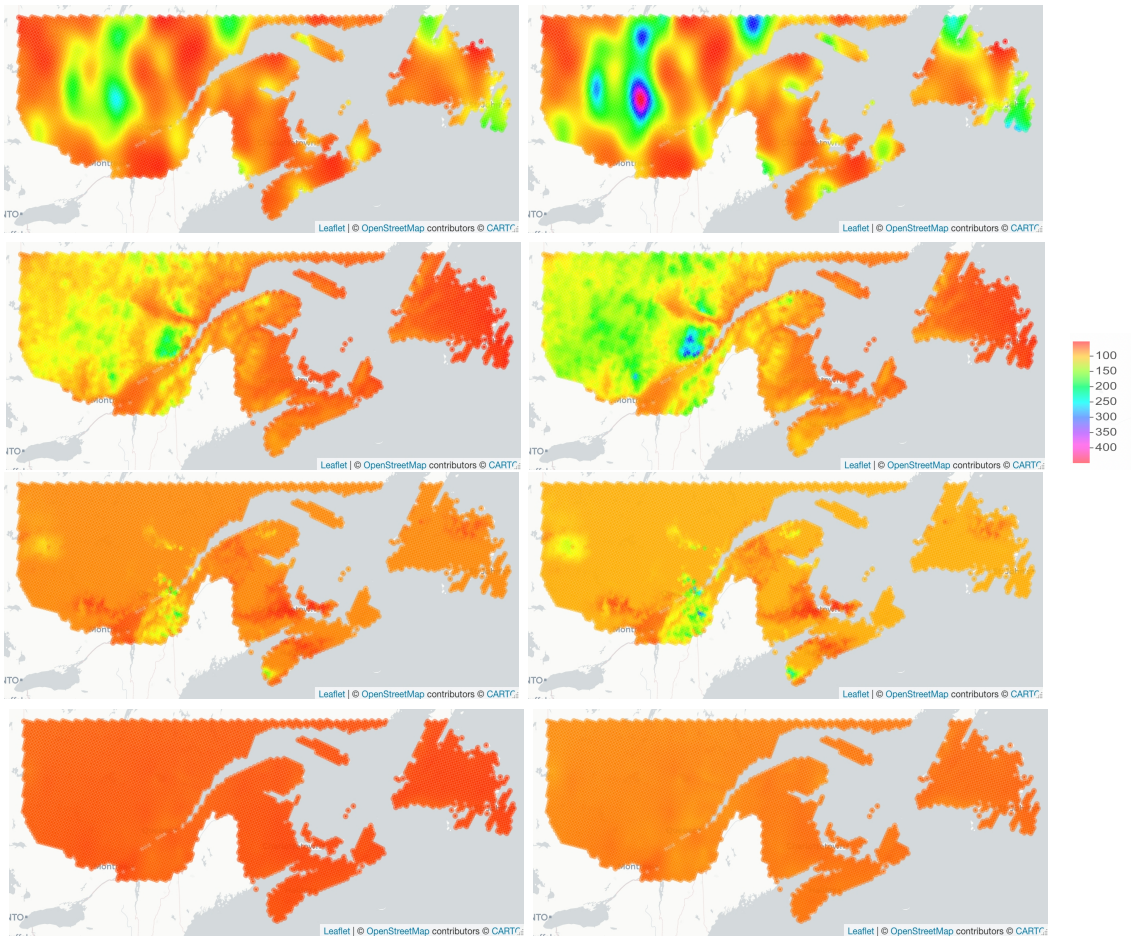

Figure 18: Obtained 20-years (left column) and 40-years (right column) precipitation return level maps in mm. True return level map (first line), via smooth copula-based GEV model (second line), via latent variable model for Gaussian Process-based simulation response surface modeling (third row) and via hierarchical max-stable spatial model (fourth row).

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

In the following, we will be interested in presenting 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 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 D and by considering covariates previously introduced in Table 1. In this setting we say "*exact technique*" when the interpolated value at station s_i is equal to the estimated value used in the interpolation. 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 D). In the rest of this section, for sake of simplicity we will drop the LM notation.

Exact interpolation techniques

Inverse distance weighted (IDW) The IDW method is widely known as the basic one in the interpolation literature (Burrough (1986)). This method relies on the assumption that all the points on the earth's 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 75% quantile precipitation covariates, denoted $(y_s^{(1)}, y_s^{(2)}) = (a_s, q_s)$ at a given station with two-dimensional coordinate l_s.

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

$$\tilde{\zeta}(s) = \frac{\sum_{i=1}^{n_f} \frac{\hat{\zeta}_i + \beta_a(a_s - a_{s_i}) + \beta_q(q_s - q_{s_i})}{||l_s - l_{s_i}||}}{\sum_{i=1}^{n_f} \frac{1}{||l_s - l_{s_i}||}}.$$

(Step 3) Estimate parameters β_a and β_q 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

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

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 75% quantile precipitation 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.

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.

Inexact interpolation techniques

Let us 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. First, Algorithm 5 presents steps for the implementation of the proposed polynomial regression method. Second Algorithm 6 gathers steps for the proposed spline-based regression method. In the present study, the spline is a function of the coordinates and altitude and 75% quantile precipitation covariates are considered linearly.

Algorithm 5 Proposed implementation for polynomial regression method

(Step 1) Consider covariates as polynomials of longitude, latitude, altitude and 75% quantile 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

$$\tilde{\zeta}(s) = \tilde{\beta}_0 + \tilde{\beta}_1 a_s + \tilde{\beta}_2 q_s + \tilde{F}(l_s).$$

(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)).

The `R` code associated to Algorithm 5 and Algorithm 6 to can be found in `classicalinterpolationtechniques.R` file in the supplementary material CodeR folder.

D 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 21), 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

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

with $c = 2/(3+\hat{\tau}_3) - \ln(2)/\ln(3)$ 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 19. 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 19 one can appreciate the quality of the fitting of GEV L-moments estimators to our data.

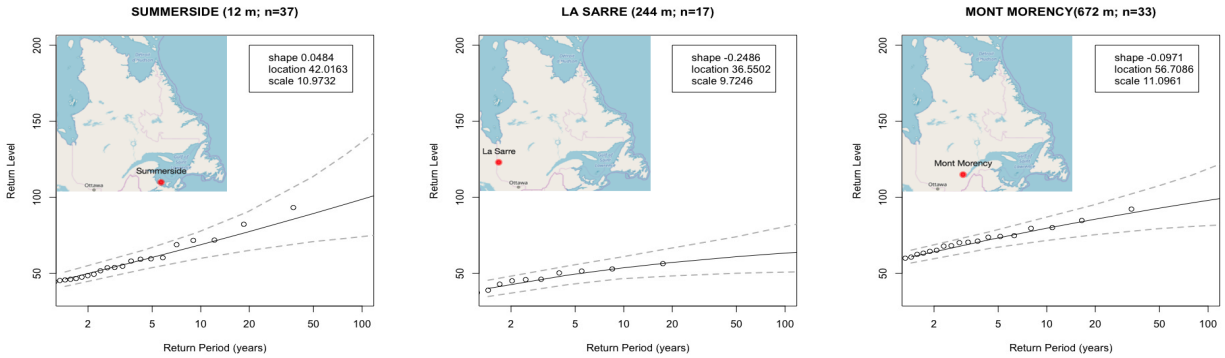

Figure 19: Locations of 3 selected stations and adequacy of fitted GEV model through the associated return level plots.

Figure 20 shows the resulting pointwise 20-years, 30-years, 40-years and 100-years return levels for the 116 stations considered. Such a map is nevertheless difficult to interpret and can only give information for the few locations where data is available. In practice, spatial return levels as in Figure 8, rather than pointwise estimates as in Figure 20, would be of much higher value.

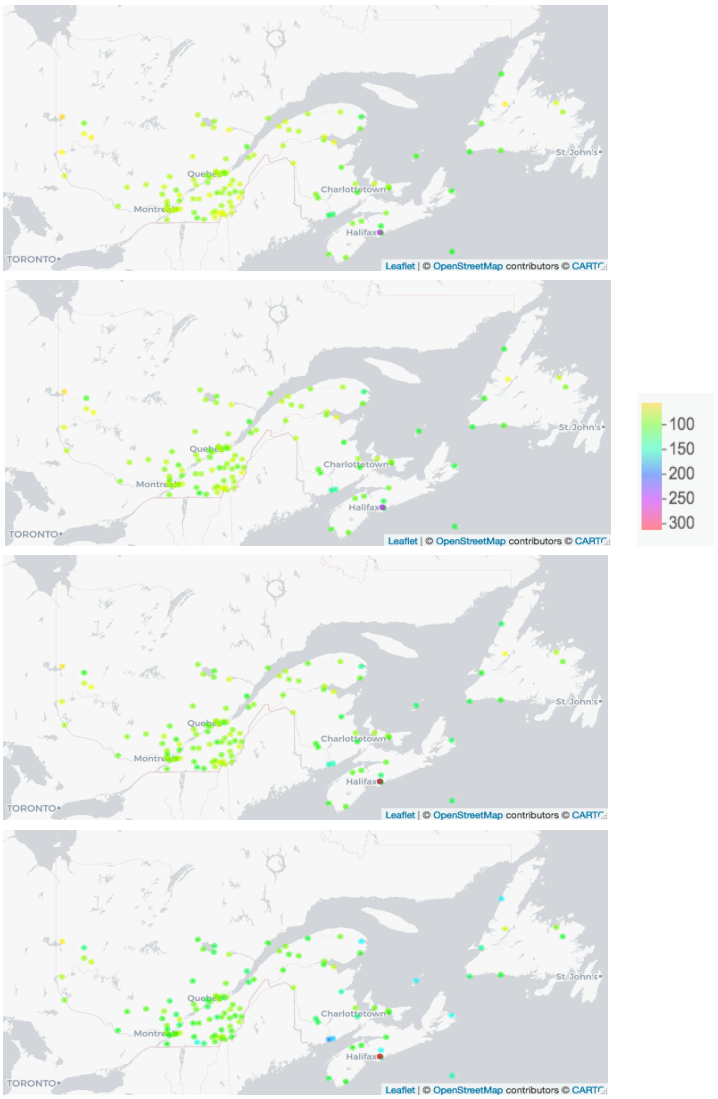

Figure 20: From top to bottom: pointwise 20-years, 30-years, 40-years, and 100-years precipitation return level map in mm for the considered 116 stations in Central Eastern Canada.

E Supplementary materials

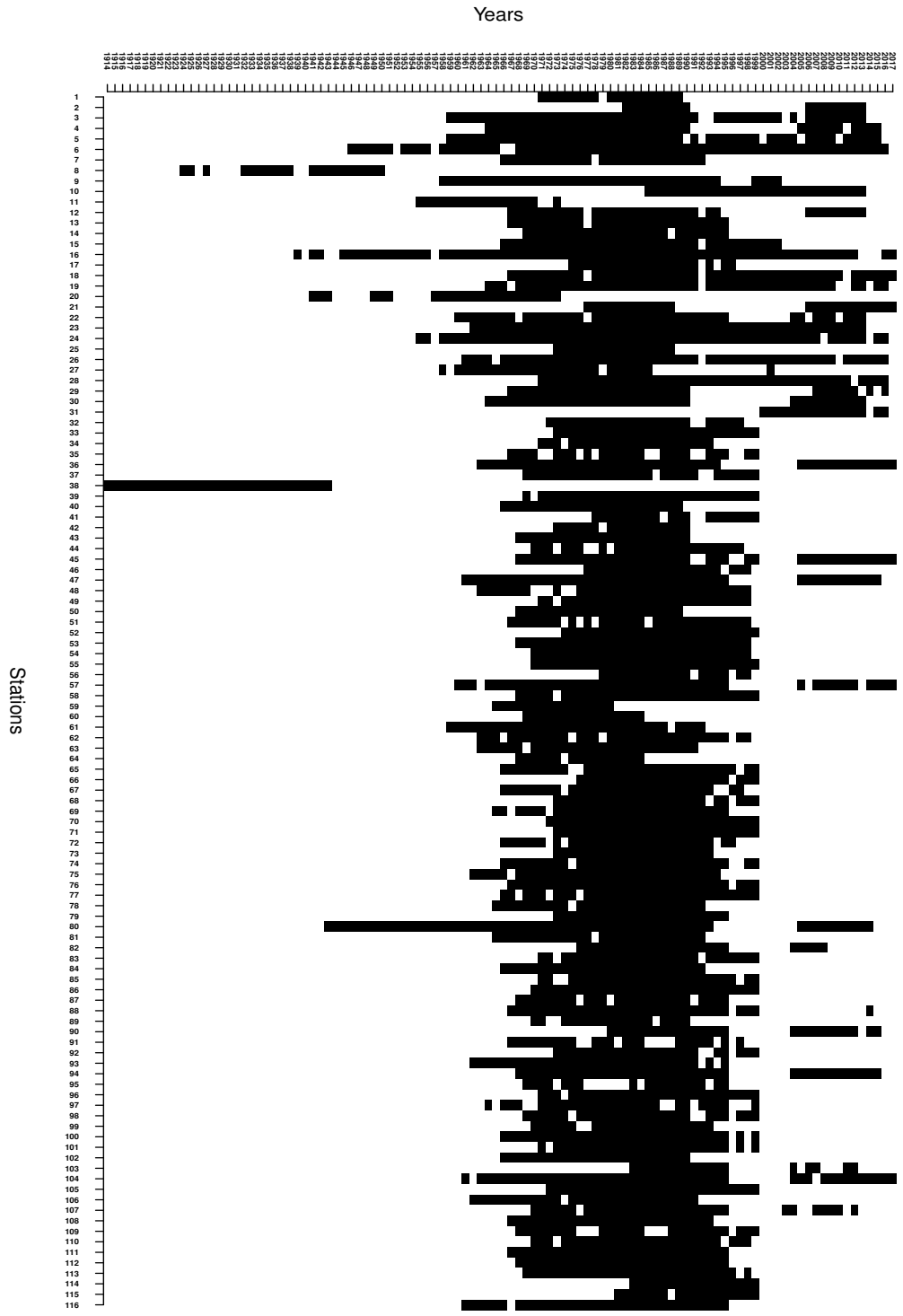

Figure 21: White cells represent missing data and black ones observed extreme rainfall registered in 116 stations in Center Eastern Canada from 1914 to 2017.

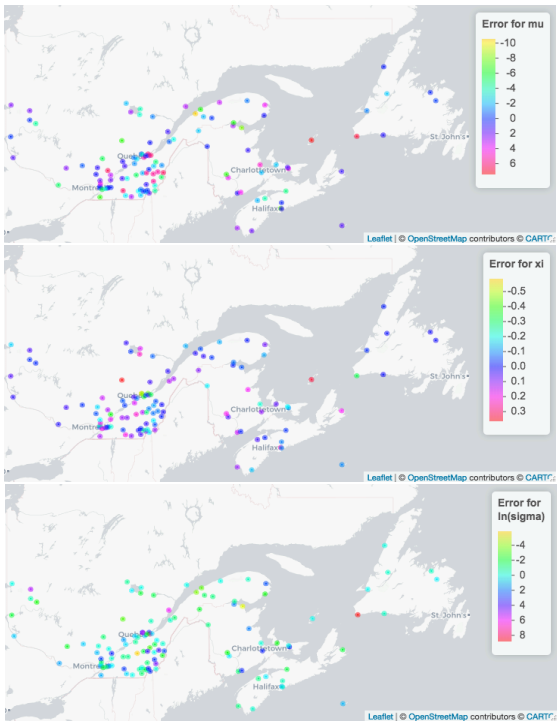

Figure 22: From top to bottom: spatial map of $\hat{\epsilon}_s$ for polynomial regression model in (4) associated to μ, to ξ and to $\ln(\sigma)$ for fitting stations in Figure 2 (blue stations).

The choice of combinations of fitting and validation stations

In the following we analyse the sensibility of the classical techniques of spatial interpolation of individual GEV distributions explained in Section C with respect to the choice of combinations of n_f and n_v stations. To this end, 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. Recall that the covariance function used for the kriging method is the one who provides the smallest AIC criterion among exponential, spherical, circular, cubic, Matérn and Gneiting covariance functions (see Algorithm 4). The obtained boxplots are gathered in Figure 23. Furthermore, the associated median and standard deviation values are displayed in Tables 8 and 9.

Table 8: 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		
	NRMSE	NMAE	NMPE	NRMSE	NMAE	NMPE
1. Pointwise GEV	0.06 (0.0015)	0.04 (0.0006)	0.17 (0.0255)	0.06 (0.0068)	0.04 (0.0027)	0.18 (0.0384)
2. IDW	0.06 (0.0015)	0.04 (0.0006)	0.17 (0.0255)	0.47 (0.0427)	0.11 (0.0130)	0.29 (0.0861)
3. PR	0.13 (0.0049)	0.09 (0.0027)	0.41 (0.0599)	0.16 (0.0650)	0.11 (0.0239)	0.40 (0.4819)
4. Spline	0.12 (0.0053)	0.08 (0.0024)	0.38 (0.0589)	0.16 (0.0420)	0.11 (0.0169)	0.41 (0.2780)
5. Kriging	0.06 (0.0015)	0.04 (0.0006)	0.17 (0.0255)	0.14 (0.0197)	0.10 (0.0129)	0.27 (0.0898)

Table 9: Median scores from Table 2 for 200 combinations between fitting and validation stations for each interpolation technique. The associated standard deviations are displayed in brackets. We consider here the three geographical coordinates and the 75% quantile precipitation as covariates.

	Fitting stations			Validation stations		
	NRMSE	NMAE	NMPE	NRMSE	NMAE	NMPE
1. IDW	0.06 (0.0015)	0.04 (0.0006)	0.17 (0.0255)	0.13 (0.0168)	0.09 (0.0084)	0.27 (0.0824)
2. PR	0.11 (0.0040)	0.07 (0.0023)	0.29 (0.0346)	0.18 (0.1185)	0.11 (0.0310)	0.56 (0.9791)
3. Spline	0.11 (0.0050)	0.08 (0.0022)	0.35 (0.0543)	0.15 (0.0407)	0.10 (0.0151)	0.40 (0.2892)
4. Kriging	0.06 (0.0015)	0.04 (0.0006)	0.17 (0.0255)	0.13 (0.0190)	0.09 (0.0097)	0.27 (0.0933)

In this analysis, we consider models with the three geographical covariates (see Table 8 and associated orange and sky-blue boxplots in Figure 23) and models where the 75% quantile precipitation covariate is additionally taken into account (see Table 9 and the associated red and dark-blue boxplots in Figure 23).

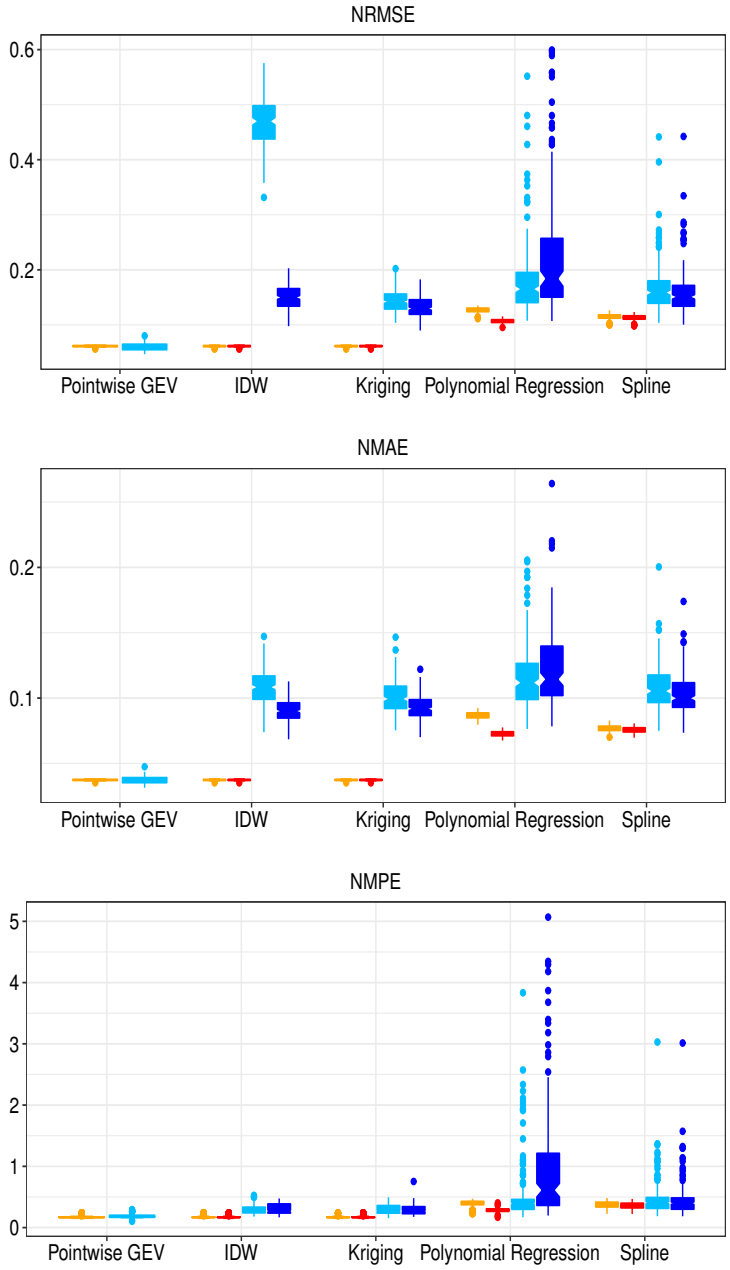

Figure 23: Boxplots of the scores obtained from Table 2 for 200 combinations between fitting and validation stations for each interpolation technique. Models with three geographical covariates are displayed in orange boxplots for fitting stations and in sky-blue boxplots for validation ones. Models with the 75% quantile precipitation in addition to the three geographical covariates are displayed in red boxplots for fitting stations and in dark -blue boxplots for validation ones.

We also consider the scores in Table 2 obtained by replacing $\tilde{q}_{p_m,s_i}$ by the L-moments quantile estimators in Appendix D at level probability p_m for each station s_i (see *Pointwise GEV* row in

Table 8 and Figure 23). 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 20 of Appendix D, for a pointwise return level map associated to these Pointwise GEV estimators. Obviously, since IDW and considered kriging techniques provide exact interpolations, their results exactly correspond with the ones from Pointwise GEV parameters estimation.

Table 8 suggests that when using only longitude, latitude and elevation as covariates, kriging performs better, since almost all considered scores are lower. Spline and polynomial regression perform similarly. However, the less performing models seems to be the polynomial regression ones, both in terms of median values (see Table 8) and of sensitivity of the combinations between fitting and validation stations (see boxplots in Figure 23). 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 23) are relatively poor compared to those for the fitting stations (orange and red ones). This considerations is true in particular for NRMSE and NMAE scores. 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 23 is considerably large. In Figure 23, it can be observed that the behaviour of models with three geographical covariates and 75% quantile precipitation as covariates slightly improve models with only the three geographical covariates.