Network design for heavy rainfall analysis

Théo Rietsch, P. Naveau, N. Gilardi, Armelle Guillou

JOURNAL OF GEOPHYSICAL RESEARCH: ATMOSPHERES, VOL. 118, 13,075–13,086, doi:10.1002/2013JD020867, 2013

Network design for heavy rainfall analysis

T. Rietsch,[1,2] P. Naveau,[1] N. Gilardi,[3] and A. Guillou[2]

Received 9 September 2013; revised 30 October 2013; accepted 20 November 2013; published 11 December 2013.

[1] The analysis of heavy rainfall distributional properties is a complex object of study in hydrology and climatology, and it is essential for impact studies. In this paper, we investigate the question of how to optimize the spatial design of a network of existing weather stations. Our main criterion for such an inquiry is the capability of the network to capture the statistical properties of heavy rainfall described by the Extreme Value Theory. We combine this theory with a machine learning algorithm based on neural networks and a Query By Committee approach. Our resulting algorithm is tested on simulated data and applied to high-quality extreme daily precipitation measurements recorded in France at 331 weather stations during the time period 1980–2010.

Citation: Rietsch, T., P. Naveau, N. Gilardi, and A. Guillou (2013), Network design for heavy rainfall analysis, *J. Geophys. Res. Atmos.*, *118*, 13,075–13,086, doi:10.1002/2013JD020867.

1. Introduction

[2] Weather, climate, and hydrological extremes have always been of importance in human history. With our changing climate, there has been a growing research effort to understand, model, and even predict extreme events at different time and spatial scales in atmospheric, hydrological, and statistical sciences [e.g., *Zwiers et al.*, 2013]. One driver for such a research endeavor resides in the increasing need of characterizing the frequency and intensity of extremes [see *Alexander et al.*, 2006; *Groisman et al.*, 2004]. Such probabilistic knowledge is paramount for impact studies, assessment methods, and adaptation strategies. In this context, high priority should be given to measuring uncertainties. This is directly linked to the issue, often overlooked in the statistical analysis of extreme events, of weather stations network design.

[3] On the one hand, statisticians and climatologists would like to work with dense networks that have been measuring long atmospherical time series in order to accurately capture the spatial and temporal variabilities of extreme values. On the other hand, economical, regulatory, and technical constraints demand a limited number of weather stations. Balancing those two opposing sides boils down to the question of how to design an optimal spatial network for extreme values.

[4] Depending on financial resources and regulatory requirements, the goal of designing a network can be either defined as *augmenting* a network size to improve the spatial coverage or *reducing* it while maintaining the best spatial coverage. In this paper, we discuss both cases but our application will only deal with the reduction size.

[5] To highlight some of the statistical difficulties about thinning spatial networks for extremes, we study daily rainfall recorded by the national French weather service, Météo-France. When analyzing rainfall measurements, there is always a tradeoff between spatial coverage and time series length. If we had opted to analyze daily precipitation with at least 50 years of instrumental data, then only a few dozen stations with high-quality recordings would have been available. In this case, the question of reducing the network size would have been only theoretical with no practical value. Météo-France has the mandate to keep and maintain those high-quality referenced stations. For this reason, we have decided to work on a shorter time period, 1980–2010, but with a large number of stations. With regard to extremes, this implies that fewer extremes in time will be analyzed and our design network has to take into account this higher-estimation variability due to the reduced sample sizes. Figure 1 displays the sites of our 331 weather stations. Three different types of sites can be identified on this map. The set of dots represents 147 weather stations with top quality measurements. The second group of 110 triangles has also high-quality recordings and rapid data quality checking. It will be kept for test and validation. The last group of 74 crosses is of lesser priority in the sense that data checking can take more than 48 h. If Météo-France wants to know the loss of information induced by removing a few weather stations, it would certainly tap into the group $\mathcal{X}_+$. In contrast, stations in $\mathcal{X}_\circ$ and in $\mathcal{X}_\triangle$ correspond to reference points (airports and such) and should not be candidates for removal. In Figure 1, we can observe that the sampling is not uniform in space. For example, the southern part along the Mediterranean Sea has already a higher density of stations than the northern part along the Belgium border. This can be explained by atmospherical and geographical reasons. The southern part can witness complex weather systems linked to

[1]Laboratoire des Sciences du Climat et de l'Environnement, CNRS, Gif-sur-Yvette, France.
[2]Institut Recherche Mathématique Avancée, Université de Strasbourg et CNRS, Strasbourg, France.
[3]Commissariat à Énergie atomique, CEA, Saint-Aubin, France.

Corresponding author: T. Rietsch, Institut Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS, 7 rue René Descartes, 67084 Strasbourg CEDEX, France. (theo.rietsch@math.unistra.fr)

2169-897X/13/10.1002/2013JD020867

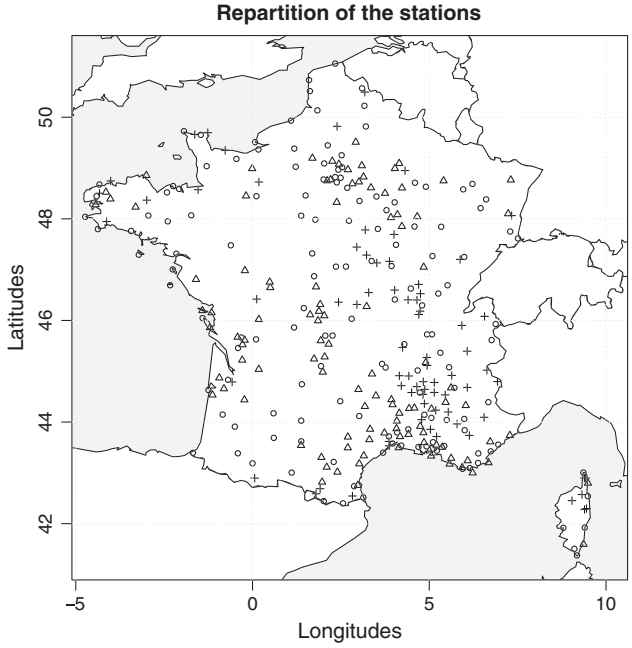

Figure 1. Locations of 331 weather stations in France (source: Météo-France). The set $\mathcal{X}_\circ$ made of dots represents 147 weather stations with top quality measurements. The second group $\mathcal{X}_\triangle$ of triangles has also 110 stations of high-quality recordings that we keep for test and validation. The last group $\mathcal{X}_+$ of 74 crosses is of lesser availability (data checking can take more than 48 h) and contains the site candidates for removal.

the pronounced orography (local effects) and coupled with a high population density (societal impacts). Extreme rainfall events there can be caused by southern winds, forcing warm and moist air to interact with mountainous areas, resulting in severe thunderstorms. Still, in terms of spatial design for extremes, is it better to remove stations in an already highly dense gauged region with very heavy rainfall or to discard an isolated weather station associated with less heavy precipitation? This inquiry puts into light the two statistical tasks needed in our study: modeling the distributional properties of extreme precipitation and developing an algorithm to decide which stations could be removed without much information loss in terms of those characteristics of extremes. The first task is directly related to Extreme Value Theory (EVT) and the second one to the field of spatial network design.

[6] For almost a century, extremal distribution features have been extensively studied by statisticians, climatologists, and hydrologists through two main statistical quantities. The distribution of maxima over a block size (e.g., a year or a season) has been classically approximated by the so-called Generalized Extreme Value distribution [*Fisher and Tippett*, 1928], and excesses over a fixed high threshold have often been fitted by a Generalized Pareto (GP) distribution [see *Pickands*, 1975]. In this work, we only focus on excesses above a threshold and we assume that they follow a GP distribution that is characterized by a scale and shape parameters; see equation (1) for details.

[7] In hydrology, the regional frequency analysis (RFA) described in *Hosking and Wallis* [1997] offers an interesting starting point to bring together GP distributions, heavy rainfall and spatial coherency. This RFA approach mainly consists in fitting a GP density at each location belonging to a climatically coherent region while the GP parameters are constrained over this region. In other words, a spatial structure is imposed on the GP parameters. This leads to the question on how to interpolate spatially GP parameters according to the network shown in Figure 1. Many avenues exist and we opt here for a well-known nonparametric regression method that will link the GP parameters with the geographical coordinates of the weather stations (the explanatory variables). The class of neural networks, adapted to spatial interpolation [*Bishop*, 2006; *Ceresetti et al.*, 2012], is a very flexible approach to model any continuous regression function. One essential reason for such a choice is that neural networks can be easily adapted to our problem of spatial network design.

[8] The theory of spatial network design has been extensively studied in the literature [see *Smith*, 2001, chapter 6]. At least two approaches can be distinguished. The first one is based on the theory of optimal design of experiments and has been initiated in the 1950s [*Kiefer and Wolfowitz*, 1959] and applied to spatial statistics in the 1980s [*Fedorov and Müller*, 1989]. It consists in fitting a model to the data and deciding to add stations at points where the variance of estimation is the highest (or to remove stations at which the variance of the model is the lowest). The second optimization method, based on the work of *Caselton and Zidek* [1984], is a Bayesian and information theoretic-based approach. In their work, they assume that observations come from an underlying random field with a multivariate normal distribution. The idea is to split the stations into two parts of fixed size. One of them corresponds to gauged sites and the other one to ungauged sites. The optimal design is the one for which the gauged sites bring the most information about the ungauged ones from an information-theoretical point of view. A Bayesian version of network design has been carefully investigated in environmental sciences, especially for dynamical monitoring networks [e.g., *Berliner et al.*, 1999; *Nychka et al.*, 1997; *Wikle and Royle*, 1999].

[9] Those past approaches cannot be directly applied to our problem for two reasons. First, we focus on extreme rainfall and their distributions may be skewed and heavy tailed. Second, the spatial patterns of precipitation are complex and it is difficult to impose an explicit dynamical model to describe them. This implies that linear models or linear approximations of design networks [*Cohn*, 1996] may not be appropriate in our context of heavy rainfall.

[10] To bypass the aforementioned limits, we will exploit the idea of Query by Committee (QBC) which is an algorithm that comes from machine learning. This algorithm was first introduced by *Seung et al.* [1992] for clustering problems and extended to the regression case by *Krogh and Vedelsby* [1995]. It was originally introduced to add weather stations. The QBC is an iterative process that gradually adds locations in order to improve the quality of the data set. It is particularly well suited in cases where new data are expensive to obtain and hence new observations have to be chosen carefully. We will adapt this method to handle the removal of weather stations.

[11] This paper is organized as follows. In section 2 we quickly describe the extremal behavior of heavy rainfall in France. Section 3 presents the QBC in our spatial

network context and its adaptation to extreme rainfall. The simulation study in section 4 highlights the advantages, drawbacks, and limitations of the QBC. In the last section, we apply our methodology to our French daily precipitation measurements.

2. Modeling of Heavy Rainfall

2.1. Extreme Value Theory

[12] As already mentioned in the introduction, we are interested by rainfall excesses above a high threshold. As explained in the book of *Coles* [2001], EVT states that if the threshold is high enough, the survival function, also called the tail distribution, can be well approximated by a GP tail defined as

$$P(Z > z) = \begin{cases} \left(1 + \frac{\xi z}{\sigma}\right)^{-1/\xi}, & \text{if } \xi \neq 0 \\ \exp\left(-\frac{z}{\sigma}\right), & \text{if } \xi = 0, \end{cases} \quad (1)$$

where the random variable Z represents thresholded rainfall excesses, $\sigma > 0$ is a scale parameter, and ξ is a shape parameter. The distribution is defined for $z \in \mathbb{R}_+$ if $\xi \geq 0$ and for $z \in [0, -\sigma/\xi]$ if $\xi < 0$. The shape parameter governs the heaviness of the tail and is usually found to be positive when dealing with precipitation. Among the different estimators for ξ that have been proposed in the literature, two are widespread in hydrology and climatology, namely the Maximum Likelihood (ML) method [*Smith*, 1985] and the probability-weighted moment (PWM) method [see *Hosking and Wallis*, 1987]. In this paper, the estimates are derived by the generalized probability-weighted moment (GPWM) method introduced by *Diebolt et al.* [2007] which is a refinement of the PWM one. Its use is motivated by a wider range of application ($\xi \in (-1, 3/2)$) compared to the PWM approach ($\xi \in (-1, 1/2)$). Unlike the ML estimator, it does not require any optimization and consequently, it is extremely fast and it does not provide divergent values (this can happen when maximizing the likelihood). The scale and shape parameters estimates are simply defined as

$$\widehat{\sigma} = \frac{2.5\widehat{\mu}_1\widehat{\mu}_{1.5}}{2\widehat{\mu}_1 - 2.5\widehat{\mu}_{1.5}} \text{ and } \widehat{\xi} = \frac{4\widehat{\mu}_1 - (2.5)^2\widehat{\mu}_{1.5}}{2\widehat{\mu}_1 - 2.5\widehat{\mu}_{1.5}}, \quad (2)$$

where $\widehat{\mu}_s$ represents the empirical estimator of the GPWM $\mu_s = \mathbb{E}\left[Z\overline{G}^s_{\sigma,\xi}(Z)\right]$ with $\overline{G}_{\sigma,\xi}(z) = P(Z > z)$.

[13] To illustrate the spatial variability of those estimates with regard to our French rainfall data, the upper and lower left panels of Figure 2 display the estimated values of $\widehat{\sigma}$ and $\widehat{\xi}$ for the measurements taken from the networks $\mathcal{X}_\circ$ and $\mathcal{X}_+$ shown in Figure 1. At each station, the threshold equals the 95% quantile after removing dry days. As expected for this country, the shape parameter roughly varies from 0 to 0.5, this latter value occurring in the southern part of France. Note that the estimated scale parameters are strongly positive and consequently, the constraint $\sigma > 0$ will be satisfied for this example. Besides the analogies and the few differences between the two interpolated maps, it is clear that the spatial pattern of heavy rainfall in France is not uniform and local and complex features are present. This rapid and exploratory analysis leads to at least three questions: What is the uncertainty estimation for the two GP parameters? Are they correlated? And how to model the spatial structures captured by each GP parameter?

[14] According to *Diebolt et al.* [2007], $\widehat{\xi}$ and $\widehat{\sigma}$ given by (2) are normally distributed with covariance matrix

$$\Sigma = \frac{1}{8(2\xi - 3)(4\xi - 7)} \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2} & \Sigma_{2,2} \end{pmatrix}, \quad (3)$$

with

$$\Sigma_{1,1} = -(\xi - 2)(8\xi^2 - 10\xi + 13)(2\xi - 5)^2,$$
$$\Sigma_{1,2} = -(2\xi - 5)(16\xi^3 - 92\xi^2 + 156\xi - 97)\sigma,$$
$$\Sigma_{2,2} = \frac{(-32\xi^4 + 328\xi^3 - 1220\xi^2 + 1918\xi - 1093)\sigma^2}{(\xi - 2)}.$$

The matrix Σ is not diagonal and this means that the two estimators $\widehat{\sigma}$ and $\widehat{\xi}$ are correlated. To visualize this correlation effect, we implement a small simulation study by randomly generating 500 GP distributed samples of size 200, which is roughly the temporal length of our excesses data set. Figure 3 (top, left) shows the corresponding cloud of points for $\sigma = 25$ and $\xi = 0.5$, possible values for our rainfall application. As shown by this panel, this negative correlation effect is far from being negligible and we will have to take it into account in our design strategy. We also draw the boxplots corresponding to the relative estimation error for $\hat{\xi}$ and $\hat{\sigma}$. As expected in EVT, the shape parameter estimate has a larger variability than the inferred scale parameter. To assess the influence of the shape parameter on the correlation strength, Figure 3 (bottom) displays the theoretical correlation between $\hat{\xi}$ and $\hat{\sigma}$ with $\sigma = 1$ and ξ varying from –0.5 to 1.5. Even the smallest absolute correlation value, at ξ around 1.0, is still large, around 0.5. In terms of spatial design, this implies that $\hat{\xi}$ and $\hat{\sigma}$ should be treated as a bivariate quantity.

2.2. Nonparametric Regressions

[15] As illustrated by the maps in Figure 2, it is not trivial to find some explanatory variables or parametric forms that could explain the spatial pattern of heavy rainfall in France. Hence, it makes sense to capture such complex spatial structures via a nonparametric approach. More precisely, we assume in the remaining of this paper that the GP scale and shape parameters vary in space and should be viewed as follows

$$\begin{cases} \mathbb{R}^2 \to & \mathbb{R}_+^* \times [-1, 1.5] \\ \mathbf{x} \mapsto f(\mathbf{x}) = & (\sigma(\mathbf{x}), \xi(\mathbf{x})), \end{cases} \quad (4)$$

where $\mathbf{x} = (\text{lat}, \text{long})$ represents the latitude and longitude coordinates of the weather station $\mathbf{x}$. Basically, *Diebolt et al.* [2007] tell us that we can start our investigations about spatial design with the simple nonparametric regression model

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x}),$$

where $\epsilon(\mathbf{x})$ corresponds to independent bivariate zero-mean Gaussian vectors with covariances defined from equation (3). In this context, the vector $y(\mathbf{x}) = (\hat{\xi}(\mathbf{x}), \hat{\sigma}(\mathbf{x}))$ represents the estimates obtained from (2) at some selected weather station location $\mathbf{x}$. At this stage, we would like to emphasize a few points.

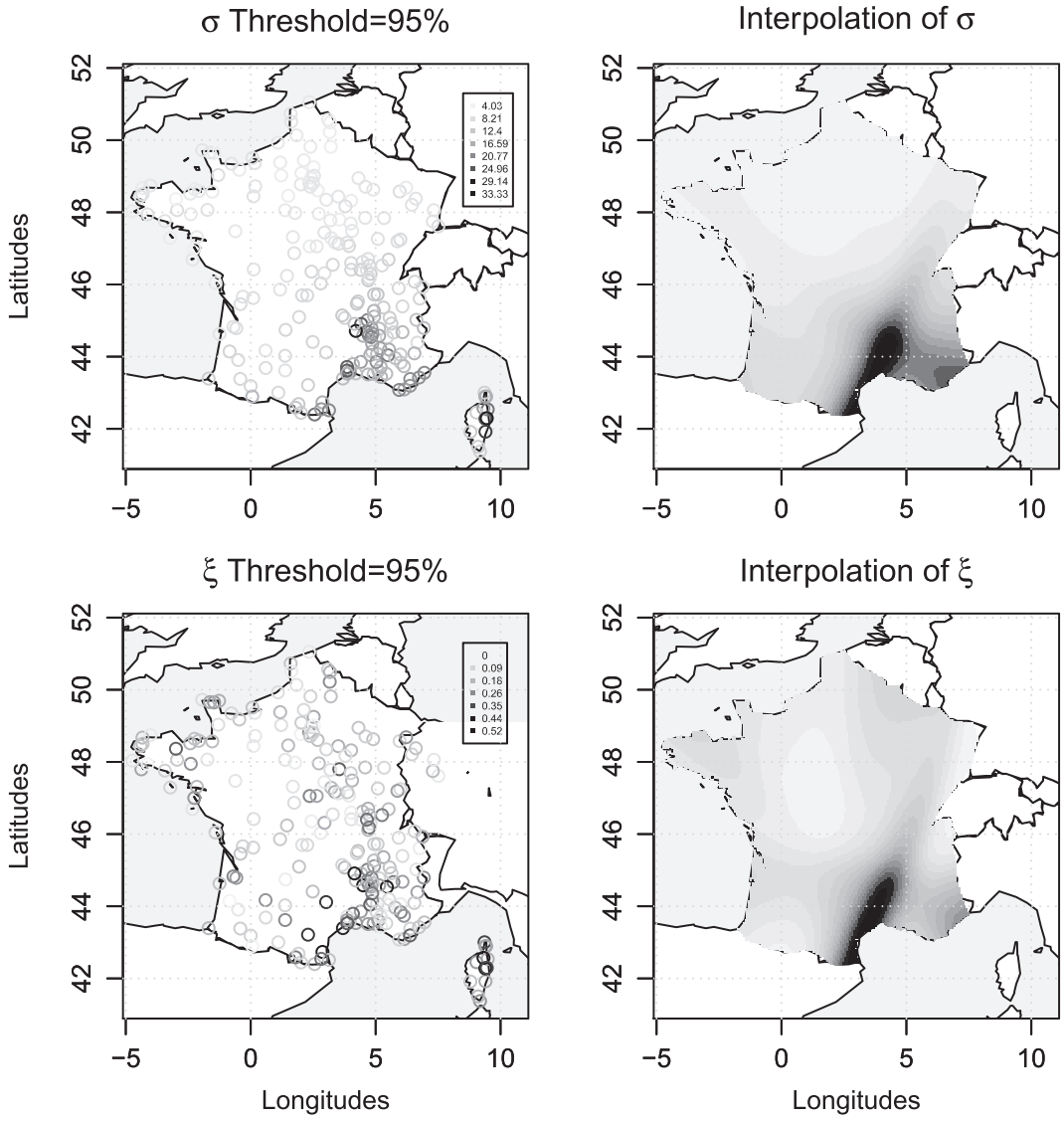

Figure 2. Exploratory analysis of heavy rainfall in France. The (left) individual and (right) interpolated GP parameters obtained with (2), respectively. Each station belongs to the networks $\mathcal{X}_\circ$ and $\mathcal{X}_+$ shown in Figure 1.

[16] In contrast to classical regression problems, our "observational" vector $y(\mathbf{x})$ is not observed anymore. It represents the estimated GP parameters, so the temporal dimension has been squeezed into a distributional form.

[17] As our heavy rainfall is assumed to be GP distributed, it is tempting to assume that those inferred GP parameters follow a bivariate Gaussian. But, as $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$ are allowed to vary in space, the two-dimensional covariance defined from equation (3) will also vary in space. This brings additional modeling complexity. Instead of imposing the asymptotic covariance blueprint described by equation (3), we prefer to integrate the uncertainty brought by the GP parameters estimation with resampling techniques; see section 3.3.

[18] It is also possible to add another level of complexity in the noise structure $\epsilon(\mathbf{x})$. For example, one could replace our noise independence assumption between $\epsilon(\mathbf{x}_j)$ and $\epsilon(\mathbf{x}_k)$ for $j \neq k$ by imposing a spatial covariance structure between the pair of stations (j, k). This is classically done in experimental design for linear models [e.g., *Berliner et al.*, 1999; *Nychka et al.*, 1997; *Wikle and Royle*, 1999]. As $f(\mathbf{x})$ is considered nonparametric here, identifiability issues will raise quickly and we prefer to keep the assumption of independence in $\epsilon(\mathbf{x})$. In other words, we allow for a complex mean behavior in $f(\mathbf{x})$ and a simple covariance structure in $\epsilon(\mathbf{x})$, instead of a simple spatial trend and a complex covariance.

[19] To estimate the unknown function $f(\mathbf{x})$ from $\{y(\mathbf{x}_i)\}$ with $i = 1, \ldots, n$, there exists a large variety of nonparametric regression methods in the statistical literature [e.g., *Hastie et al.*, 2001]. Keeping in mind our end goal (spatial design), we use one-hidden-layer neural networks with a hyperbolic tangent as activation function and two linear output neurons (two neurons because we jointly model the two GP parameters).

[20] In a nutshell, a neural network scheme expresses the observations as an additive combination of nonlinear

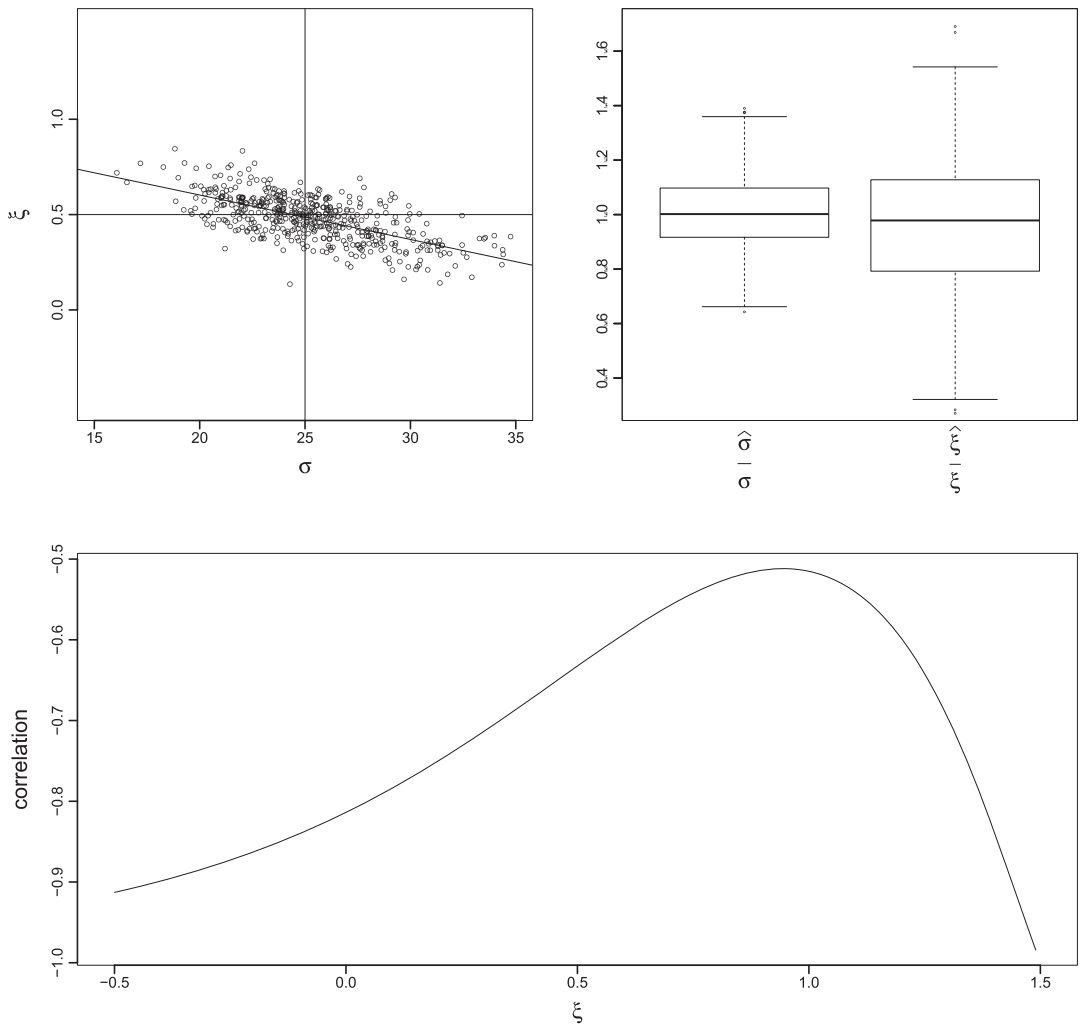

Figure 3. Basic properties of the GP estimates defined by (2). (top, left) The strong negative correlation between 500 estimates of $\sigma = 25$ and $\xi = 0.5$ and sample size 200. (top, right) The boxplots indicate the relative estimation error for $\hat{\xi}$ and $\hat{\sigma}$. (bottom) The theoretical correlation between $\hat{\xi}$ and $\hat{\sigma}$ with $\sigma = 1$ and ξ varying from –0.5 to 1.5.

blocks (here hyperbolic tangent functions that depend on the weather station latitude and longitude coordinates as explanatory variables), i.e.,

$$\beta_0^{(H+1)} + \sum_{j=1}^{H} \beta_j^{(H+1)} \tanh\left(\beta_0^{(j)} + \beta_1^{(j)}\text{lat} + \beta_2^{(j)}\text{long}\right),$$

where H is called the number of hidden neurons and β represents the vector of parameters in $\mathbb{R}^{4H+1}$ to be estimated. The networks are optimized by minimizing a classical sum of squares between observed and predicted values. Among the different optimization routines available, we choose the standard backpropagation algorithm [see *LeCun et al.*, 1998; *Rumelhart et al.*, 1986]. Concerning the number of hidden neurons H that characterizes the network complexity, we use a standard k-fold cross validation criterion [*Ceresetti et al.*, 2012, section 4.1]. As an example, the right panels of Figure 2 (right, top and bottom) show one possible set of two maps obtained by a neural network with two linear outputs and jointly fitted to the two GP parameter values displayed on the left panels (those parameters have been centered and renormalized when fitting the neural network).

[21] One of the features of neural networks that differentiates them from other interpolation techniques such as splines is that the optimization of the parameters converges to local minima [see *Rumelhart et al.*, 1986]. This means that a different initialization of the parameters will lead to a different network, probably with a different complexity. This drawback can be viewed as an asset in a spatial design context. It allows to generate a wide variety of regression functions for $f(\mathbf{x})$. The Query by Committee algorithm explained below takes advantage of this property.

3. Query by Committee and Spatial Design

3.1. Presentation of the Algorithm

[22] The Query by Committee (QBC) approach is a machine learning algorithm introduced by *Seung et al.* [1992]. It is based on neural networks and allows to create design of experiments. Its usefulness is emphasized in situations where it is expensive and time consuming to obtain

new data [see *Krogh and Vedelsby*, 1995]. We start by a presentation of its principle and theoretical basis.

[23] The key ingredient of the QBC resides in its so-called committee of experts. This term simply means that m different experts are capable to produce m different models of the unknown function $f(\mathbf{x})$. This is very similar to Bayesian Model Averaging techniques [see, e.g., *Hoeting et al.*, 1999; *Sabourin et al.*, 2013]. Then, the idea is to search where these models disagree the most with respect to a common opinion. Mathematically, this latter term can be viewed as a weighted ensemble average [see *Hansen and Salomon*, 1990; *Perrone and Cooper*, 1993] defined by

$$\overline{f}(\mathbf{x}) = \sum_{k=1}^{m} p_k \hat{f}^{(k)}(\mathbf{x}), \tag{5}$$

where $(\hat{f}^{(1)}, \ldots, \hat{f}^{(m)})$ represents our m models (experts) and the nonnegative weights p_k correspond to a priori knowledge about the quality of each expert. Here we set them all at $1/m$.

[24] A natural and important question is how to produce the committee of experts. As already mentioned, the optimization of neural networks converges to local minima and thus offers a simple way to build several different models with the same initial data set. Another avenue that we will also use resides in the possibility to draw samples from our GPWM estimates; see section 3.3.

[25] To measure the disagreement among experts, we simply compute

$$\Delta(\mathbf{x}) = \frac{1}{m} \sum_{k=1}^{m} d(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})), \tag{6}$$

where $d(.,.)$ is a distance between the ensemble average and each individual member at the site $\mathbf{x}$; see section 3.2 for details. The QBC algorithm then advices to choose new weather stations at the local maxima of the disagreement function and starts the routine again with the updated data sets.

[26] If s represents the number of stations that we want to add at each iteration, then the algorithm can be summed up as follows:

[27] 1. Take the initial training network $\mathcal{X}_\circ = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and its corresponding observational vector $\mathcal{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ with $\mathbf{y}_i = (\hat{\xi}(\mathbf{x}_i), \hat{\sigma}(\mathbf{x}_i))$.

[28] 2. Build a committee of m experts $(\hat{f}^{(1)}, \ldots, \hat{f}^{(m)})$ by regressing $\mathcal{Y}$ on $\mathcal{X}_\circ$ with m neural networks obtained by changing the initial conditions.

[29] 3. Find the s largest local maxima of the disagreement function $\Delta(\mathbf{x})$ among all experts over the whole territory and denote them by $\mathcal{X}' := (\mathbf{x}'_1, \ldots, \mathbf{x}'_s)$.

[30] 4. Use the estimates obtained at the stations of $\mathcal{X}'$ to get a new observational vector $\mathcal{Y}' = (\mathbf{y}'_1, \ldots, \mathbf{y}'_s)$.

[31] 5. Add the elements of $\mathcal{X}'$ (resp. $\mathcal{Y}'$) to the initial training network $\mathcal{X}$ (resp. $\mathcal{Y}$).

[32] 6. Restart from Step 1 with the updated data sets until a stopping criterion is reached.

[33] After each algorithmic loop, the average $\overline{f}(\mathbf{x})$ needed in Step 3 is recomputed with the updated data sets.

[34] As already stated in the introduction, our main goal is to propose a design algorithm to reduce the number of stations. In its current form, the QBC method only adds weather stations. This issue can be solved by noticing that rainfall data are already available at the stations that could be removed. More precisely, the candidates for removal belong to the set $\mathcal{X}_+$ shown in Figure 1. Hence, if we want to remove r stations from $\mathcal{X}_+$, we can progressively add stations from $\mathcal{X}_+$ to $\mathcal{X}_\circ$ until only r stations remain in $\mathcal{X}_+$. These last r stations will be considered the least informative and will be removed from the network. The choice of r, classically given by the decision maker, corresponds to our stopping criterion mentioned in Step 6.

[35] To assess the quality of the final fit, we have kept the set $\mathcal{X}_{\text{test}}$ shown in Figure 1 for validation. The generalization error can be derived as

$$\Delta^2_{\text{test}} = \sum_{\mathbf{x}_i \in \mathcal{X}_\triangle} d(\mathbf{y}_i, \overline{f}(\mathbf{x}_i)), \tag{7}$$

where $\mathbf{y}_i$ represents the estimate inferred at the location $\mathbf{x}_i$.

3.2. Choice of the Disagreement Function

[36] An essential feature of the QBC resides in the choice of the distance $d(.,.)$ to compute the disagreement function defined by (6). If one is only interested in the shape parameter ξ, a pertinent choice could be

$$d_\xi(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})) = \left(\hat{\xi}^{(k)}(\mathbf{x}) - \overline{\xi}(\mathbf{x})\right)^2.$$

Similarly, optimizing the network with respect to σ could be done by opting for

$$d_\sigma(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})) = \left(\hat{\sigma}^{(k)}(\mathbf{x}) - \overline{\sigma}(\mathbf{x})\right)^2.$$

In hydrology, estimating high return levels represents a classical output and, in this context, a valuable distance could be

$$d_p(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})) = \left(\widehat{q}_p^{(k)}(\mathbf{x}) - \overline{q}_p(\mathbf{x})\right)^2,$$

where $p \in (0, 1)$ and the GP quantile is obtained from

$$\widehat{q}_p^{(k)}(\mathbf{x}) = \left((1-p)^{-\widehat{\xi}^{(k)}(\mathbf{x})} - 1\right) \frac{\widehat{\sigma}^{(k)}(\mathbf{x})}{\widehat{\xi}^{(k)}(\mathbf{x})}.$$

[37] In the remainder of the paper, we focus on the three distances: $d_\sigma(.,.)$, $d_\xi(.,.)$ and $d_p(.,.)$. To conclude this section, we note that the function $d(.,.)$ is called at two different places, for computing $\Delta(\mathbf{x})$ and calculating Δ_{test}. In terms of interpolation quality, i.e., goodness of fit, the objects of interest are the scale and shape surfaces. Hence, $d_\sigma(.,.)$ and $d_\xi(.,.)$ should be used in Δ_{test}, but the distance $d_p(.,.)$ does not give direct information about the two surfaces. Comparing experts is a different task and $d_p(.,.)$ can be plugged in $\Delta(\mathbf{x})$.

3.3. GP Parameters Variability

[38] As noted in section 2.2, our "observational" vector $y(\mathbf{x}) = (\hat{\xi}(\mathbf{x}), \hat{\sigma}(\mathbf{x}))$ does not correspond to rainfall intensity but to estimated GP parameters. As GPWM are inferred quantities, it is possible to draw many realizations of shape and scale parameters in order to capture some variability inherent to transforming rainfall excesses into GP parameters. For each station $\mathbf{x}_i$, this means that we compute the GPWM estimates with the corresponding sample of excesses of length, say n_i. This gives us a pair $(\widehat{\sigma}(\mathbf{x}_i), \widehat{\xi}(\mathbf{x}_i))$ of estimates. Then we simulate ℓ samples of length n_i from a

$GP(\widehat{\sigma}(\mathbf{x}_i), \widehat{\xi}(\mathbf{x}_i))$ distribution. The GPWM method is applied on each sample to provide ℓ shape and scale estimates. Those ℓ shape and scale parameters available at each station will be used as the initial input of our QBC algorithm.

[39] Another approach to resample GP parameters could be to use the asymptotic Gaussian approximation proposed by *Diebolt et al.* [2007]; see equation (3). This will be faster but not much, the GPWM equations defined by (2) being explicit. By avoiding the asymptotic Gaussian approximation, we do not have to assume that the GP parameter estimates are perfectly normally distributed and consequently, we explore more accurately the uncertainties linked to the estimation with the GPWM method (the scatterplot in Figure 3 (top, left) may not be exactly Gaussian).

3.4. QBC for Heavy Rainfall

[40] Before implementing the main steps of our modified QBC algorithm, the practitioner needs to make a few decisions.

[41] One disagreement function $d(.,.)$ has to be selected. The total number of stations, say r, that should be removed from the set $\mathcal{X}_+$ has to be fixed. If we do not know how many sites have to be disregarded, then r has to be estimated by computing a stopping criterion defined by Δ_{test} [e.g., see *Gilardi*, 2004].

[42] The thinning of the network is based on adding progressively sites and, at the end of the process, removing the stations "that have not been added." Depending on the computational capacities available, the number of stations added at each QBC iteration, say s (or equivalently the number of iterations), has to be chosen. Adding only one station at each iteration, i.e., $s = 1$, is ideal, but it is the most expensive in terms of computational costs since the committee of experts has to be reconstructed by fitting m neural networks at each iteration.

[43] The selection of m, the number of experts, and of ℓ, the number of replicas for exploring the GPWM estimates variability, also corresponds to a tradeoff between computational constraints and sampling quality.

[44] All parameters being specified, we can summarize our modified QBC algorithm:

[45] 1. Compute the GP parameter estimates with (2) at each station $i \in \{1, \ldots, n\}$. Denote $\mathbf{y}_{i,1} := (\widehat{\xi}_1(\mathbf{x}_i), \widehat{\sigma}_1(\mathbf{x}_i))$.

[46] 2. For each station $i \in \{1, \ldots, n\}$, for each $j \in \{2, \ldots, \ell\}$, repeat the following resampling technique (explained in section 3.3) to obtain ℓ observational vectors $\mathcal{Y}_1, \ldots, \mathcal{Y}_\ell$ where $\mathcal{Y}_j = (\mathbf{y}_{1,j}, \ldots, \mathbf{y}_{n,j})$ and $\mathbf{y}_{i,j} = (\widehat{\xi}_j(\mathbf{x}_i), \widehat{\sigma}_j(\mathbf{x}_i))$:

a. Use the vector of estimates $\mathbf{y}_{1,i}$ to simulate a sample $\mathcal{Z}_{i,j} = (z_{i,j}^1, \ldots, z_{i,j}^{n_i})$ from a GP with the parameters of $\mathbf{y}_{i,1}$. We take n_i as the number of observation in the original sample of precipitation provided by station i.

b. Use the data of $\mathcal{Z}_{i,j}$ to obtain the vector of the resampled estimates of the parameters of the GP $\mathbf{y}_{i,j}$.

[47] 3. For $j = 1, \ldots, \ell$, take the initial training network as $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and its corresponding observational vector $\mathcal{Y}_j = (\mathbf{y}_{1,j}, \ldots, \mathbf{y}_{n,j})$ and implement the modified QBC, i.e.,

(i) Build a committee of m experts $(\hat{f}_j^{(1)}, \ldots, \hat{f}_j^{(m)})$ by regressing $\mathcal{Y}_j$ on $\mathcal{X}$ with m neural networks obtained by changing the initial conditions.

(ii) Find the s largest local maxima of $\Delta(\mathbf{x})$ that belong to $\mathcal{X}_+$ and denote them by $\mathcal{X}' := (\mathbf{x}'_1, \ldots, \mathbf{x}'_s) \subset \mathcal{X}_+$.

(iii) Use the estimates obtained at the stations of $\mathcal{X}'$ to get a new observational vector $\mathcal{Y}'_j = (\mathbf{y}'_{1,j}, \ldots, \mathbf{y}'_{s,j})$ by applying the resampling technique with $\ell = 1$.

(iv) Add the elements of $\mathcal{X}'$ (resp. $\mathcal{Y}'_j$) to the current training network $\mathcal{X}$ (resp. $\mathcal{Y}_j$).

(v) Restart from Step (i) with the updated data sets $\mathcal{X}$ and $\mathcal{Y}_j$ until no more than r stations can be added.

(vi) Store the sites in $\mathcal{X}_+$ that have not been added in Step (v).

[48] 4. Plot the frequency of the weather stations in $\mathcal{X}_+$ that have been stored in Step (vi).

[49] From a statistical perspective, the final product of this experiment can be understood as a binomial trial. We have a set of $dim(\mathcal{X}_+)$ individuals in $\mathcal{X}_+$ that can belong to two different groups of size r (stations removed) and $dim(\mathcal{X}_+) - r$ (stations kept), respectively. If all stations contain the same information (the null hypothesis), then the number of times, say $N(\mathbf{x})$, that a site $\mathbf{x}$ should be removed in ℓ trials follows a binomial distribution

$$P(N(\mathbf{x}) = i) = \binom{\ell}{i} (p(\mathbf{x}))^i (1 - p(\mathbf{x}))^{\ell - i}, \tag{8}$$

with

$$p(\mathbf{x}) = \frac{r}{dim(\mathcal{X}_+)} \text{ and } i = 1, \ldots, \ell.$$

For each station, we can identify if we fail to reject the null hypothesis ($p(\mathbf{x})$ does not vary with $\mathbf{x}$) for a given significant level, say 0.95.

4. Simulation Study

[50] In this simulation section, we mainly investigate the impacts of the disagreement functions, among four possibilities: $d_\sigma(.,.)$, $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a completely random disagreement.

[51] To simplify tables and figures, we work with one-dimensional simulations, i.e., $\mathbf{x} \in [-1, 1]$, and focus on two cases. The first case deals with a constant scale parameter and a varying shape parameter

$$\sigma(x) = 30 \text{ and } \xi(x) = \frac{1}{2}\left(1.3 - \exp\left(-16x^2\right)\right) - 0.1, \tag{9}$$

see the solid line in Figure 4. The second case corresponds to a dual situation, a constant shape parameter, and a varying scale parameter

$$\sigma(x) = 20\left(1.3 - \exp\left(-16x^2\right)\right) \text{ and } \xi(x) = 0.5, \tag{10}$$

see the solid line in Figure 5. In both cases, the varying functions are mostly constant but with a sharp drop around 0.5. If weather stations are uniformly distributed along the segment $[-1, 1]$, a well-adjusted optimal network should keep its stations around $x = 0.5$ and proposes to remove stations in regions where the parameters are fairly constant. Other setups with different varying shape parameters variations have also been studied and are available upon request (basically, our main conclusions remain identical).

[52] Concerning the other parameters, we consider a set of $n = 55$ stations randomly drawn on $[-1, 1]$ and divided

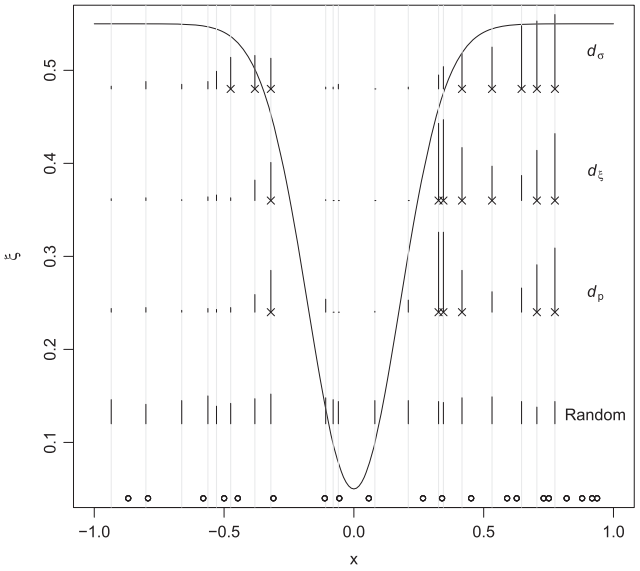

Figure 4. QBC outputs for GP parameters following (9). The solid line represents the hidden true shape parameter $\xi(x)$. The dots along the x axis correspond to the 20 sites in $\mathcal{X}_\circ$ and the locations candidate for removal are represented by the 20 vertical grey lines. The QBC is tuned to remove $r = 5$ stations out of 20. From top to bottom, the four horizontally distinct groups of black segments correspond to the x coordinates of stations removed by the QBC with respect to our four disagreement functions, $d_\sigma(.,.)$ and $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a complete random one, respectively. Crosses indicate sites where the binomial hypothesis with $p(\mathbf{x}) = 5/20$ and $\ell = 100$ trials, see equation (8), is rejected at the 0.95 significant level. Each black segment height is proportional to the number of times that a site is chosen to be removed.

into our three groups with $dim(\mathcal{X}_\circ) = dim(\mathcal{X}_+) = 20$ and $dim(\mathcal{X}_\triangle) = 15$. Our goal is to remove $r = 5$ stations, i.e., to add $20 - 5 = 15$ stations that belong to $\mathcal{X}_+$. We set $s = 3$ to add three stations at each iteration. The number of experts m and the resampling of GP estimates ℓ are set to $m = \ell = 100$.

[53] In Figure 4, the dots along the x axis correspond to the 20 sites in $\mathcal{X}_\circ$. For this particular random draw, we can see that there are a little bit more sites on the interval $[0, 1]$ than on $[-1, 0]$. Consequently, it would make sense to remove stations on already dense segments. The set of stations which are candidates for removal is represented by 20 vertical grey lines, i.e., points in $\mathcal{X}_+$. From top to bottom, the four horizontally distinct groups of black segments correspond to the x coordinates of stations removed by the QBC with respect to our four disagreement functions, $d_\sigma(.,.)$ and $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a completely random one, respectively.

[54] As expected, the completely random disagreement function produces the bottom group of black segments without any crosses. All grey lines are chosen with the same probability, and the shape parameter fluctuations are not taken into account. In comparison, the three other disagreement functions do a much better job. For instance, no significant grey lines marked with a cross are around $x = 0$ where there is a lot of change in $\xi(x)$ and few stations in $\mathcal{X}_\circ$, information is precious here and no stations should be removed in this vicinity. In addition, the asymmetrical distribution of circles belonging to $\mathcal{X}_\circ$ has been taken into account. Many more stations in $\mathcal{X}_+$ are removed on the right side of $x = 0$ than on the left one. Still, slight differences exist between the QBC outputs based on $d_\sigma(.,.)$, $d_\xi(.,.)$, or $d_p(.,.)$. In particular, $d_\xi(.,.)$ appears to capture well the two facts that a very high density of sites in $\mathcal{X}_\circ$ are located at $x \in [0.5, 1]$ and the shape parameter is constant over this region. If five stations have to be removed, the network should be thinned around this area. This feature is not as clear with $d_p(.,.)$ and to a lesser degree with $d_\sigma(.,.)$.

[55] The same type of interpretation can be undertaken if we replace model (9) by model (10); see Figure 5. One particular aspect of this graph is that the random draw of sites in $\mathcal{X}_\circ$ had put three sites in a tiny neighborhood around $x = 0.4$. But no station around this point was removed with the distance $d_\sigma(.,.)$ and $d_\xi(.,.)$. This phenomenon can be explained by the sharp gradient in $\sigma(x)$ and the low number of points in $\mathcal{X}_+$ in the interval $[0, 0.5]$. One can also compare with the neighborhood of also three sites around $x = -0.6$. This is reassuring because it emphasizes that the density network cannot be the sole criterion for removal. Our QBC algorithm attempts to balance three constraints: the smoothness of the GP parameters and two spatial densities, of the unmovable network $\mathcal{X}_\circ$ and of the group of potentially removable stations $\mathcal{X}_+$.

[56] Concerning the goodness of fit among distances, the values of Δ_{test} defined by (7) and computed at the first and last QBC iterations are displayed for our two models in Table 1. The bold values represent the best value within a block. The distance to find the optimal design via the QBC algorithm, see equation (6), can be different from the distance used for assessment in Δ_{test} defined by (7).

[57] For model (9), the expected result occurs, the smallest Δ_{test} happens when the distances to optimize, see (6), and to judge, see (7), are equal. As the number of added stations increases at each QBC step, it makes sense that Δ_{test} is smaller at the last iteration.

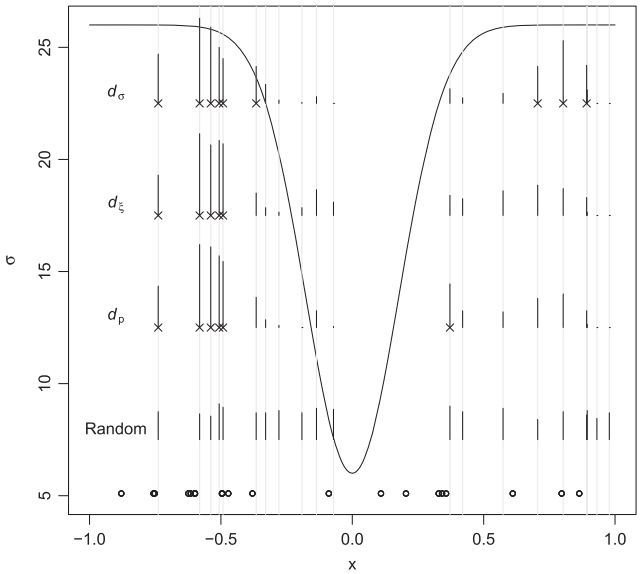

Figure 5. Same as Figure 4 but for GP parameters following (10) and the solid line corresponds to the hidden true scale parameter $\sigma(x)$.

Table 1. Values of Δ_{test} Defined by (7) at the First and Last QBC Iterations, for Each Distance Function and for Our Two Models[a]

Distance Chosen		Model (9)		Model (10)	
in (7)	in (6)	First Step	Last Step	First Step	Last Step
$d_\sigma(.,.)$	$d_\sigma(.,.)$	3.45	**3.30**	3.77	2.97
	$d_\xi(.,.)$	3.45	3.32	3.77	**2.95**
	$d_p(.,.)$	3.45	3.34	3.77	2.97
	Random	3.45	3.32	3.78	3.02
$d_\xi(.,.)$	$d_\sigma(.,.)$	0.159	0.147	0.142	**0.116**
	$d_\xi(.,.)$	0.160	**0.145**	0.142	**0.116**
	$d_p(.,.)$	0.160	0.146	0.142	**0.116**
	Random	0.160	0.150	0.143	**0.116**

[a]The bold values represent the best value within a block.

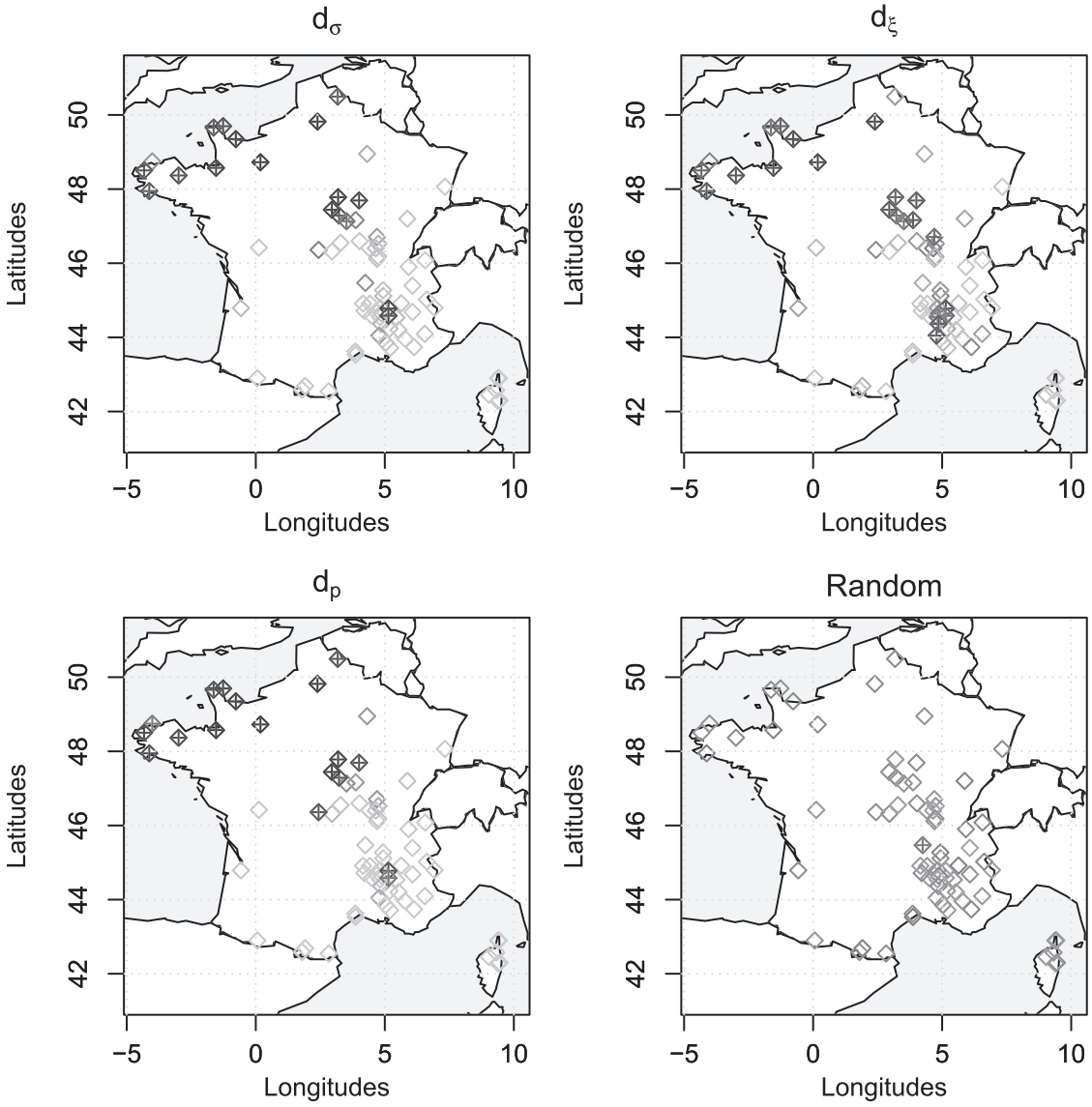

Figure 6. Candidate sites for removal from the original set of 74 stations in $\mathcal{X}_+$; see Figure 1. Each panel from top left to bottom right corresponds to a different disagreement function: $d_\sigma(.,.)$, $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$, and a completely random disagreement. Crosses indicate sites where the binomial hypothesis with $p(\mathbf{x}) = 14/74$ and $\ell = 100$ trials, see equation (8), is rejected at the 0.95 significant level. Grey diamonds correspond to sites where we fail to reject the null hypothesis (i.e., less than 26 out of 100) and the color intensity is proportional to the number of times that a station was selected for removal, the lighter the smaller.

Table 2. Same as Table 1 but for Our French Heavy Rainfall and for $d_\sigma(.,.)$[a]

Distance Chosen		QBC	
in (7)	in (6)	First Step	Last Step
$d_\sigma(.,.)$	$d_\sigma(.,.)$	4.44	**3.81**
	$d_\xi(.,.)$	4.44	3.83
	$d_p(.,.)$	4.44	**3.81**
	Random	4.45	3.89

[a]The bold values represent the best value within a block.

[58] For model (10), there is no clear winner with respect to $d_\xi(.,.)$ in (7). This may be due to a few causes. The true shape parameter being small, the quantity of the type $(\hat{\xi} - \overline{\xi})^2$ can be minuscule and more decimals are needed (or a renormalized version). A second reason is that the shape parameter is constant in model (10), so judging with $d_\xi(.,.)$ is not discriminatory enough. This is particularly relevant because one needs very large sample sizes to detect small changes around $\xi = 0.5$ [e.g., P. Naveau et al., A non-parametric entropy-based approach to detect changes in climate extremes, submitted to *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 2013; Table 2].

5. French Heavy Rainfall

[59] Coming back to our French rainfall network in Figure 1, the number of stations in $\mathcal{X}_+$ is written as $74 = 15 \times 4 + 14$ to run our QBC algorithm in reasonable time (i.e., a few hours on a desktop computer). With our notations, this means that we add at each iteration $s = 15$ sites and the $r = 14$ remaining stations will be tagged as strong candidates for removal. This strategy will be repeated $\ell = 100$ times with the ℓ resampled GP estimates that will be used as inputs for the $m = 100$ experts.

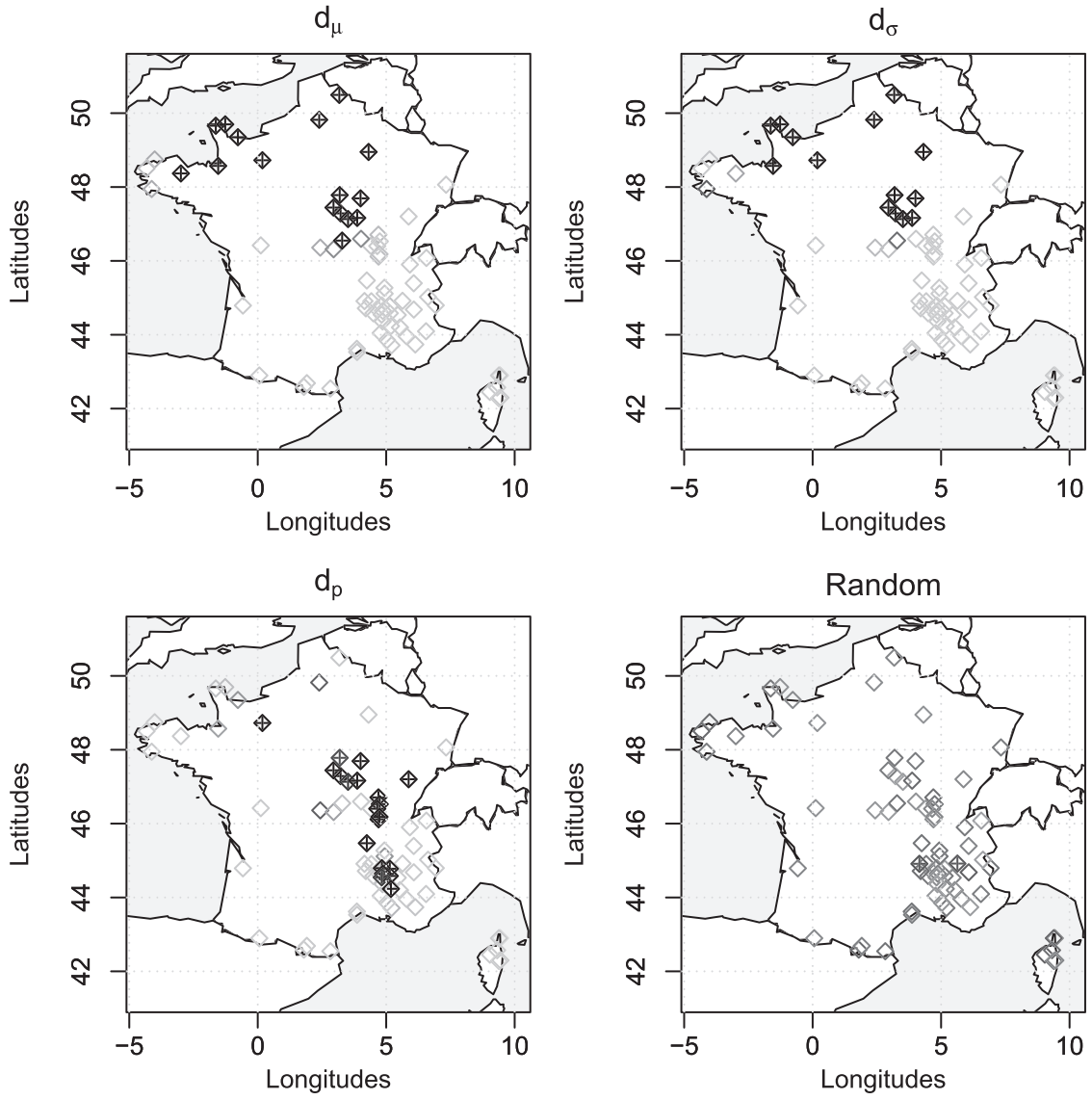

Figure 7. Same as Figure 6 but with the hypothesis that heavy rainfall follow a Gaussian distribution with mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$.

[60] From a computational point of view, it takes basically a little more than a couple of seconds to build a neural network with a desktop computer. Applying the algorithm 100 times, i.e. making 4 + 1 steps each time and building committees of 100 experts at each step, implies that 50,000 neural networks are constructed during the whole process. This roughly corresponds to 100,000 seconds which means a little more than one day of computation.

[61] Figure 6 provides four reduced networks, each one corresponding to a different disagreement function: $d_\sigma(.,.)$, $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$, and a completely random disagreement. Clearly, the random option cannot take into account the spatial variability of the GP parameters observed in Figure 2 and each station in $\mathcal{X}_+$ is selected for removal with the same probability.

[62] The other three distances produce similar maps and the statistically significant stations (with regard to the binomial hypothesis with $p(\mathbf{x}) = 14/74$ and $\ell = 100$ trials) are basically spread along the northern coastline of France with a few points in the center of France; see the crosses. No stations are removed in the Pyrenees and Corsica. This makes sense because the orography is complex and the network in $\mathcal{X}_\circ$ is not very dense there. A few stations in the southern region near Avignon have also been selected for removal. This can be explained by the very high density of stations in $\mathcal{X}_\circ$.

[63] The main message of Figure 6 appears to be that it is more reasonable to remove stations in the north, above the 46° latitude, (even if the network density is already low there) than discard sites in the south (even if the network is already dense there). This can be explained by statistical and atmospheric arguments. Intense heavy rainfall happen in the south, and the GP parameter is more difficult to estimate for heavier tails.

[64] Concerning the goodness of fit, we only focus on $d_\sigma(.,.)$ in Table 2, the values for $d_\xi(.,.)$ being basically all equal to 0.124. As expected, the fit improves when adding stations through Step 3-v of our QBC, between the first and last step. The disagreement functions $d_\sigma(.,.)$ and $d_p(.,.)$ appear to provide the best value of Δ_{test}.

[65] To close our rainfall example, we repeat our QBC algorithm but with the "business as usual" hypothesis, i.e., we assume in (4) that heavy rainfall excesses follow a Gaussian distribution with mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$, instead of our GP distribution assumption with parameters $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$. In the disagreement function d_p that is used here, the quantile corresponds to the quantile of a Gaussian distribution with parameters μ and σ^2. The normality assumption cannot be justified by EVT, but it is interesting for the decision maker to know if the classical Gaussian model leads to a different set of stations. Comparing Figures 6 and 7 answers positively to this inquiry. The tip of Brittany, the furthest Western part of France, is treated differently, no stations removed in Finistère with the Gaussian hypothesis. In addition, the disagreement function $d_p(.,.)$ in Figure 7 produces a very different network thinning from the ones obtained with $d_\mu(.,.)$ and $d_\sigma(.,.)$. This was not the case with the GP assumption, the QBC generated similar reduced networks. This French example illustrates that network thinning for extremes should be treated carefully and classical statistical tools may not be appropriate to make the right decisions.

6. Discussion

[66] Our main goal here was to identify weather stations that can be potentially removed from an existing network. A variety of aspects were not taken into account in this study and could open other research avenues. We can list at least five items.

[67] We have to keep in mind that this approach should not be seen as a decision criterion but as a help in the decision process. Other factors such as economical costs also have to be taken into consideration when it comes to decide which stations have to be removed. Also, it would be of great interest to use the economical cost of the suppression of a station as a weight to include in the decision process.

[68] In our application, we only focus on heavy rainfall but a weather station measures other atmospheric variables. For example, extreme winds and heat waves clearly are of interest for impact studies. Hence, a multivariate statistical approach that would integrate many atmospheric characteristics (and not only precipitation) should be implemented before taking any decisions concerning the removal of stations. This poses new mathematical challenges because multivariate EVT should be used to model the interactions between different types of extremes, especially for heat waves which have very different spatial patterns (large scale phenomena) than heavy rainfall (local scale).

[69] This example of heat waves also brings into light our assumption of spatial independence among our nonparametric regression residuals. For our case study, this is a reasonable hypothesis because we only want to remove a dozen of sites over France, i.e., the stations are apart by a few hundred kilometers. It is very unlikely, for heavy rainfall, that a strong spatial dependence is present in our residuals for this spatial range. But, it would have been different for heat waves and a covariance structure in $\epsilon(\mathbf{x})$ would be then necessary.

[70] From a more mathematical perspective, it would also be interesting to take advantage of spatial GP processes of excesses. This probabilistic object can handle infinite dimensions, and therefore, they are especially tailored to interpolate extreme values at locations without a station. This property is particularly useful for augmenting the network size. Although the probabilistic framework of such spatial processes is now well understood [e.g., *Ferreira and de Haan*, 2012], there is still a lot of open mathematical questions regarding the modeling and inference of such processes.

[71] Finally, combining EVT and machine learning appears to be a promising strategy to handle practical problems dealing with heavy rainfall modeling. One important idea of the QBC is the committee of experts. Here we have used neural networks, but other nonparametric (such as Kernel regression methods) or even parametric approaches could be used. Augmenting the number of ways of building experts will allow to capture additional uncertainties. In other words, other experts (i.e., statistical models) could be easily added to the QBC blueprint. Still, theoretical work on the mathematical side and practical considerations are needed to insure that such a strategy is optimal at removing weather stations in an operational context.

[72] **Acknowledgments.** Part of this work has been supported by the EU-FP7 ACQWA project (www.acqwa.ch) under contract 212250, by the

PEPER-GIS-ADEME project, by the ANR MOPERA project, by the ANR McSim project, and the MIRACCLE-GICC project. The authors thank Olivier Mestre for his expertise and help and Météo-France for providing a large data set of stations. The authors also would like to thank the reviewers for their interesting suggestions that allowed us to improve the final version of this paper.

References

Alexander, L. V., et al. (2006), Global observed changes in daily climate extremes of temperature and precipitation, *J. Geophys. Res.*, *111*, D05109, doi:10.1029/2005JD006290.

Berliner, M., Z.-Q. Lu, and C. Snyder (1999), Statistical design for adaptive weather observations, *J. Atmos. Sci.*, *56*, 2536–2552.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Science and Statistics, New York.

Caselton, W. F., and J. V. Zidek (1984), Optimal monitoring network designs, *Stat. Probab. Lett.*, *2*, 223–227.

Ceresetti, D., E. Ursu, J. Carreau, S. Anquetin, J. D. Creutin, L. Gardes, S. Girard, and G. Molinié (2012), Evaluation of classical spatial-analysis schemes of extreme rainfall, *Nat. Hazards Earth Syst. Sci.*, *12*, 3229–3240, doi:10.5194/nhess-12-3229-2012.

Cohn, D. A. (1996), Neural network exploration using optimal experiment design, *Neural Networks*, *9*, 1071–1083.

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

Diebolt, J., A. Guillou, and I. Rached (2007), Approximation of the distribution of excesses through a generalized probability-weighted moments method, *J. Stat. Plann.*, *137*, 841–857.

Fedorov, V. V., and W. Müller (1989), Comparison of two approaches in the optimal design of an observation network, *Statistics*, *20*, 339–351, doi:10.1080/02331888908802178.

Ferreira, A., and L. de Haan (2012), The generalized Pareto process; with application, *Tech. Rep.*, arXiv:1203.2551v1.

Fisher, R. A., and L. H. C. Tippett (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Math. Proc. Cambridge*, *24*, 180–190, doi:10.1017/S0305004100015681.

Gilardi, N. (2004), Design of experiments by committee of neural networks, *Tech. Rep.*, Institut Français du Pétrole.

Groisman, P. Y., R. W. Knight, D. R. Easterling, T. R. Karl, G. C. Hegerl, and V. N. Razuvaev (2004), Trends in intense precipitation in the climate record, *J. Clim.*, *18*, 1326–1350, doi:10.1175/JCLI3339.1.

Hansen, L. K., and P. Salomon (1990), Neural network ensembles, *IEEE Trans. Pattern Anal. Mach. Intell.*, *12*, 993–1001, doi:10.1109/34.58871.

Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*, 382–417, doi:10.1214/ss/1009212519.

Hosking, J. R. M., and J. R. Wallis (1987), Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics*, *29*, 339–349, doi:10.1080/00401706.1987.10488243.

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

Kiefer, J., and J. Wolfowitz (1959), Optimum designs in regression problems, *Ann. Math. Stat.*, *30*, 271–294, doi:10.1214/009053604000000382.

Krogh, A., and J. Vedelsby (1995), Neural network ensembles, cross validation, and active learning, in *Advances in Neural Information Processing Systems*, edited by G. Tesauro, D. S Touretzky, and T. K Leen, pp. 231–238, MIT Press, Cambridge,

LeCun, Y., L. Bottou, G. B. Orr, and K.-R. Müller (1998), Efficient backprop, in *Neural Networks: Tricks of the Trade*, edited by G. B. Orr and K. Müller, Lecture Notes in Computer Science, pp. 9–50, Springer, Berlin Heidelberg, doi:10.1007/3-540-49430-8_2.

Nychka, D., Q. Yang, and J. A. Royle (1997), Constructing spatial designs using regression subset selection, in *Statistics for the Environment-3: Pollution Assessment and Control*, pp. 131–154, Wiley, New York.

Perrone, M. P., and L. N. Cooper (1993), When networks disagree: Ensemble methods for hybrid neural networks, in *Artificial Neural Networks for Speech and Vision*, edited by R. J. Mammone, pp. 126–142, Chapman and Hall, London.

Pickands, J. (1975), Statistical inference using extreme order statistics, *Ann. Stat.*, *3*, 119–131.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986), Learning representations by back-propagating errors, *Nature*, *323*, 533–536, doi:10.1038/323533a0.

Sabourin, A., P. Naveau, and A.-L. Fougères (2013), Bayesian model averaging for multivariate extremes, *Extremes*, *16*, 325–350, doi:10.1007/s10687-012-0163-0.

Seung, H. S., M. Opper, and H. Sompolinsky (1992), Query by committee, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory,* COLT '92, pp. 287–294, ACM, New York, doi:10.1145/130385.130417.

Smith, R. L. (1985), Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, *72*, 69–90, doi:10.1093/biomet/72.1.67.

Smith, R. L. (2001), *Environmental Statistics*, Univ. of North Carolina, Chapel Hill.

Wikle, C. K., and J. A. Royle (1999), Space-time dynamic design of environmental monitoring networks, *J. Agric. Biol. Environ. Stat.*, *4*, 489–507.

Zwiers, F. W., L. V. Alexander, G. C. Hegerl, J. P. Kossin, T. R. Knuston, P. Naveau, N. Nicholls, C. Schär, S. I. Seneviratne, and X. Zhang (2013), Climate extremes: Challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events, in *Climate Science for Serving Society*, pp. 339–389, Springer, Netherlands.