A Bayesian Hierarchical Model for Spatial Extremes with Multiple Durations

Yixin Wang[a], Mike K. P. So[a]

[a] *The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*

Abstract

Bayesian spatial modeling of extreme values has become increasingly popular due to its ability to obtain relevant uncertainty measures for the estimates. This has implications for the problem of limited data on the study of extreme climatological events. Noticing the abundance of non-daily environmental records, 1-hour and 6-hour records in particular, we propose a Bayesian hierarchical model that can address multiple durations, in addition to the spatial effects, within a set of extreme records with multiple durations. The generalized Pareto distribution for threshold exceedances and the binomial distribution for exceedance frequencies are adopted for the top-most characterization of extreme data. The duration effect on spatial extremes is characterized by pooling the extremes with different durations and merging the duration into a latent spatial process structure as one of the covariates. Statistical inference is performed using Markov Chain Monte Carlo (MCMC) methods, for which an adaptive tuning algorithm is proposed. The methodology is applied to simulated datasets and real precipitation data for a region around Hong Kong.

Keywords: Bayesian analysis; Environmental data; Hierarchical model; Multiple durations; Spatial extreme; Spatial-temporal process.

1. Introduction

Investigation of the patterns of extreme climatological events plays a crucial role in environmental risk analysis and policy making. It informs the development of precautionary measures and sheds light on long-term regional planning. A careful study of these extreme events is, therefore, crucial. However, the limited availability of records of extreme events has always been a challenging aspect of such studies. Observations are only available from designated stations, where even the commonly studied daily records tend to be limited, let alone the records of extreme events. These limitations have motivated studies on the dependence structure of extreme events, which enable our estimations to borrow strength from data nearby.

Many previous studies have considered the dependence structure of extreme events as characterized by various covariates such as space, time, etc. de Haan (1985) first studied the high-dimensional multivariate dependence structure of extreme observations. Corresponding dependence measures were later developed for the study of multivariate structure of extreme events (Coles, Heffernan & Tawn, 1999; Schlather & Tawn, 2003; Cooley, Naveau & Poncet, 2006). In addition to many classical works using frequentist approach in spatial modeling, e.g., de Haan & Resnick (1977), de Haan (1984), Resnick (1987), Smith (1989), Tawn (1988), Tawn (1990), Coles (2001), and Davison, Padoan, & Ribatet (2012), the hierarchical Bayesian approach also receives attention in spatial and extreme value modeling. Wikle, Berliner, & Cressie (1998) described the monthly maximum temperature using an underlying state process with site-specific time series models characterizing the state variable; they incorporated both large-scale variabilities and small-scale space-time dynamics into the model. Cooley & Sain (2010) captured the spatial structure of meteorological extremes via a latent process under the point process representation, whereas Cooley, Nychka, & Naveau (2007) presented a similar idea with the threshold exceedance characterization. Ghosh & Mallick (2008) considered a spatial-temporal framework by incorporating spatial dependence into the likelihood while keeping the temporal dependence within the latent structure. Sang & Gelfand (2009) generalized the spatial

structure by adopting multivariate Markov random field models for the joint spatial dependence of location and scale parameters, while temporal dependence was allowed only for the location parameter. For spatial-temporal modeling, Sang & Gelfand (2010) further proposed continuous spatial process models with mean square continuous realizations to relax the common conditional independence assumption in the topmost hierarchies. More recently, Kunihama, Omori, & Zhang (2012) and Nakajima, Kunihama, Omori, & Frűhwirth-Schnatter (2012) proposed the state space approach to model temporal dependence in extremes and proposed several efficient algorithms for their corresponding Bayesian inferences.

Recall that setting up a dependence structure via a latent spatial process in hierarchical Bayesian spatial modeling helps in the study of pooled records from different locations. In particular, it helps to strengthen the parameter estimation at a particular location by borrowing strength from records at neighboring stations. Continuing with this thinking, we borrow additional strength from records with different durations. The use of multiple duration records is common in the study of intensity-duration-frequency relationships (Koutsoyiannis, Kozonis, & Manetas, 1998) and in constructing intensity-duration-frequency curves. Baldassarre, Castellarin, & Brath (2006) used extreme rainfall with durations ranging from 15 minutes to 24 hours to index storm estimation. Feng, Nadarajah, & Hu (2007) studied extreme precipitation in China using 2-day, 5-day, and 10-day moving sums of precipitation in addition to daily values. Back, Uggioni, & Vieira (2011) modeled precipitations of short durations. We pool the records from different locations and durations, and integrate duration as a covariate into a latent dependence structure.

The proposed model adopts a typical three-layer Bayesian hierarchy that is a similar in spirit to that of Cooley, Nychka, & Naveau (2007). The characterization of extremes is divided into two parts – the generalized Pareto distribution (GPD) for threshold exceedances and the binomial distribution for exceedance frequencies. A latent Gaussian process, involving duration and spatial covariates, is assumed for each distribution parameter. Essentially, if the spatial characterization includes longitude and latitude coordinates, for instance, then duration is weighted to serve as a third coordinate, where this weight is left as a parameter to estimate. This integration of duration into the model helps to avoid potentially nonsensical return-level estimates, such as a 6-hour return-level estimate that is higher than the 24-hour one, which can be the result when modeling data of different durations separately. In terms of statistical inference, we obtain approximate draws from the posterior distribution by adopting the adaptive random walk Metropolis-Hastings (MH) algorithm within a Gibbs sampler. An adaptive tuning algorithm is proposed to facilitate the mixing of Markov Chain Monte Carlo (MCMC) chains and hence the implementation of our approach.

There are two reasons for taking duration into consideration. First, the characterization of extreme records, especially those of precipitation, should always specify how long the time coverage is. In addition, most previous studies of meteorological extremes focus only on those of a particular duration, such as daily precipitation. However, in addition to daily records, there are usually many more records with other durations, namely 1-hour or 6-hour ones. Studying all of these records together can ease the data scarcity problem to a large extent and thus achieve a potential gain in model efficiency. In the Atmospheric & Environmental Real-time Database of the Institute for the Environment (IENV) at the Hong Kong University of Science and Technology (HKUST), for example, some stations only have 1-hour observations. Hence, over 96% of the records fall into the 1-hour category, whereas only around 2% are 6-hour, and less than 1% are 24-hour. Apart from this abundance of data, records with shorter durations are especially valuable for studying extremes as extreme events, precipitation in particular, seldom sustain their peak level over 12 hours. Moreover, the devastating damage caused by severe events is mostly due to extreme events over a relatively short period rather than cumulative events over a longer duration. Therefore, we propose a model to study the entire set of records with multiple durations and we believe the inclusion of duration in the dependence structure is worth investigating.

The proposed approach is illustrated by an application to a dataset of 1-hour, 6-hour, and 24-hour precipitation records in a small region around Hong Kong with 89 stations. In this application, we keep the shape parameter of the GPD constant throughout the study region for model simplicity. One striking outcome is that in the latent Gaussian process, the duration has a negative coefficient in the mean structure but a relatively large one in the covariance structure. This suggests that most extreme precipitation events occur with strong intensity but short durations. In addition, the subtlety of choosing multiple thresholds evolves naturally from the pooling of records with different durations. We compare a one-threshold model

(100 mm for 1/6/24-hour), a two-threshold model (50 mm for 1/6-hour and 100 mm for 24-hour), and a three-threshold model (25mm for 1-hour, 50mm for 6-hour and 100mm for 24-hour). The three models produce GPD scale and shape parameter estimates that fit similarly well to the data, which suggests a certain robustness of the model at least in this particular application. With the use of the proposed adaptive tuning algorithm,the addition of duration to the model does not impede a reasonably fast mixing of Markov Chain Monte Carlo (MCMC) chains when applied to both the real data and the simulated data.

The paper is organized as follows. Section 2 reviews extreme value statistics and discusses the spatial extremes. Section 3 describes the construction of the Bayesian hierarchical spatial model with multiple durations. Section 4 discusses the sampling scheme for Bayesian inference and sketches the proposed adaptive tuning algorithm. Section 5 demonstrates, via simulation studies, how the Bayesian estimator and the proposed adaptive tuning algorithm perform in general. Section 6 presents an application of the proposed model to real precipitation extremes. We conclude with a discussion in Section 7.

2. Extreme value statistics and spatial extremes

2.1. Extreme value distributions

Extreme value theory formulates statistical models for the tail of a probability distribution. There are two major univariate approaches: the block maxima approach and the exceedance over threshold approach. The block maxima approach considers the maxima in a series of equally sized finite blocks. Given a series of independently and identically distributed continuous data $Z_1, Z_2, \cdots$ and letting the block size be n, i.e., $M_n = \max(Z_1, Z_2, \cdots, Z_n)$, it follows from the Fisher-Tippett-Gnedenko theorem that if the normalized distribution of M_n converges to a non-degenerate function, i.e.,

$$\lim_{n\to\infty} P\{\frac{M_n - b_n}{a_n} \leq z\} = F(z), a_n > 0, b_n \in \mathbb{R},$$

where F is a non-degenerate distribution function, then the limit distribution lies in either the Gumbel, the Frèchet, or the Weibull family, which can be grouped into the generalized extreme value (GEV) distribution. Instead of following Gnedenko's approach to find the normalizing sequences a_n and b_n, it is possible to fit the GEV distribution directly to unnormalized data.

As the block maxima is not necessarily an extreme value, and an extreme value might not be a block maxima, the block maxima approach may, in some cases, suffer from some information loss. In this case, the exceedance over threshold approach can be adopted. Given that exceedances of a sufficiently high threshold are rare events to which the Poisson distribution applies, it is known from the Pickands-Balkema-de Haan theorem that the Generalized Pareto Distribution (GPD) is an asymptotic distribution that can be used to represent the conditional cumulative distribution of the excess $Y = Z - u$ of the variable Z over the threshold u, given $Z > u$ for sufficiently large u, i.e. $P(Z - u \leq x | Z > u)$ (Pickands, 1975; Davison & Smith, 1990). The cumulative distribution function of the GPD is

$$F(x; \sigma_u, \xi) = \begin{cases} 1 - (1 + \xi \frac{x}{\sigma_u})^{-1/\xi} & \text{for } \xi \neq 0, \\ 1 - \exp(-\frac{x}{\sigma_u}) & \text{for } \xi = 0, \end{cases} \tag{1}$$

for $x \geq 0$ when $\xi \geq 0$, and $0 \leq x \leq -\sigma/\xi$ when $\xi < 0$, where $\sigma_u > 0$ is the scale parameter and $\xi \in \mathbb{R}$ is the shape parameter.

2.2. Return levels

The t-observation return level – denoted as z_t here – that is exceeded on average once every t observations is convenient for interpreting extreme value models rather than individual parameter values. We illustrate its calculation with the exceedances over threshold characterization of extremes. Following the notation in Section 2.1, we denote Z as the variable of interest and u as the chosen threshold. The definition of return level then implies

$$P\{Z > z_t\} = \frac{1}{t}. \tag{2}$$

However, given that the GPD is an asymptotic distribution for exceedances over high thresholds, equation (1) gives, when $\xi \neq 0$,

$$P\{Z > z_t\} = \zeta_u[1 + \xi(\frac{z_t - u}{\sigma_u})]^{-1/\xi}, \tag{3}$$

where σ_u is the scale parameter, ξ is the shape parameter, and $\zeta_u = P\{Z > u\}$ can be seen as a rate parameter. Equating equations (2) and (3) solves z_t, i.e.,

$$z_t = u + \frac{\sigma}{\xi}[(t\zeta_u)^{\xi} - 1], \tag{4}$$

provided that t is sufficiently large to ensure $z_t > u$ and $\xi \neq 0$. If we further assume the number of observations taken in a year to be n, then the r-year return level l_r when $\xi \neq 0$ becomes

$$l_r = u + \frac{\sigma}{\xi}[(rn\zeta_u)^{\xi} - 1], \tag{5}$$

and the estimation of return levels then proceeds with the substitution of parameters by their estimates.

Consider σ_u, ξ, and ζ_u as functions of locations and durations, i.e., they are actually $\sigma_u(\boldsymbol{x}, d)$, $\xi(\boldsymbol{x}, d)$, and $\zeta_u(\boldsymbol{x}, d)$. Then, even if we fix the location $\boldsymbol{x}$, associated with different durations – d's – are probably different values of the parameters $\sigma_u(\boldsymbol{x}, d)$, $\xi(\boldsymbol{x}, d)$, and $\zeta_u(\boldsymbol{x}, d)$. Hence, different durations yield correspondingly different return levels via the resulting different scale, shape, and rate parameters.

2.3. Spatial extremes

There is a pressing need to understand spatial extremes because devastating extremes of natural processes such as precipitation or temperature are inherently spatial. The modeling challenges mainly lie in two areas: the overall dependence structure and the site-wise marginal distributions. Classical extreme value modeling can be done via point process representation (Resnick, 1987; 2007). The approaches to the dependence structure can be based on spatial max-stable processes using the spectral representation of de Haan (1984). Various classes of max-stable processes have been proposed in the literature such as Brown & Resnick (1977), Smith (1990), Schlather (2002), and Kabluchko, Schlather, & de Haan (2009). An application of max-stable processes can be found in Coles (1993). Davis, Klüppelberg, & Steinkohl (2013a) studied max-stable processes using space-time correlation functions. Likelihood inference was considered in Padoan, Ribatet, & Sisson (2010) and Genton, Ma, & Sang (2011). Efficient inference via log-Gaussian random functions (Wadsworth & Tawn, 2014) and by composite likelihood (Huser & Davison, 2013a; Erhardt & Smith, 2013) have been proposed. Davis, Klüppelberg, & Steinkohl (2013b) and Huser & Davison (2013b) presented new ideas for space-time modeling in the max-stable framework. Recently, Ferreira & de Haan (2014) proposed a generalized Pareto process method that extends the classical GPD idea to continuous functions. Other approaches are built on copulas, as in Sang & Gelfand (2010), or on latent variables, as in Cooley, Naveau, & Poncet (2007), allowing a better fit for marginal distributions. Our work is an extension of the latent variable modeling framework for spatial extremes cumulative over time.

3. Model

The Bayesian hierarchy in the proposed model has a typical three-layer structure composed of a data layer, a process layer, and a prior layer. The first data layer fits data to their corresponding distributions as a measurement process. The second process layer describes the latent process – in this case a spatial plus duration process – that drives the behavior of the distribution parameters in the first layer. This formulation concludes with a prior layer that incorporates the prior information on the base parameters that control the latent spatial process. Given the three layers constructed above, the Bayesian hierarchical formulation, with each piece of π as a conditional probability density function, is

$$\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{y}) \propto \underbrace{\pi(\boldsymbol{y}|\boldsymbol{\theta}_1)}_{data}\underbrace{\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)}_{process}\underbrace{\pi(\boldsymbol{\theta}_2)}_{prior},$$

where $\boldsymbol{\theta}_j$ is the collection of parameters of the jth model layer.

3.1. Model formulation

Let $Z(\boldsymbol{x}, d)$ denote the data record, for example total precipitation, over a d-hour duration and at a location $\boldsymbol{x}$, which is defined by its longitude and latitude coordinates, where d usually takes the values of 1, 3, 6, 12, 15, 18, 24, etc. To study the extreme records and infer the return levels afterwards, one must first be able to infer all of the parameters in equation (5), i.e., $\sigma_u(\boldsymbol{x}, d)$, $\zeta_u(\boldsymbol{x}, d)$, and $\xi(\boldsymbol{x}, d)$, by fitting the model to the $Z(\boldsymbol{x}, d)$'s. In particular, the inference of $P(\boldsymbol{Z}(\boldsymbol{x}, d) > z + u)$ can be naturally decomposed into two parts, namely threshold exceedances and exceedance frequencies, as follows.

$$P(\boldsymbol{Z}(\boldsymbol{x}, d) > z + u) = P(\boldsymbol{Z}(\boldsymbol{x}, d) - u > z | \boldsymbol{Z}(\boldsymbol{x}, d) > u) P(\boldsymbol{Z}(\boldsymbol{x}, d) > u) \tag{6}$$

$$= (1 + \xi(\boldsymbol{x}, d) \frac{z}{\sigma_u(\boldsymbol{x}, d)})^{-1/\xi(\boldsymbol{x}, d)} \zeta_u(\boldsymbol{x}, d) \text{ with } \zeta_u(\boldsymbol{x}, d) = P(\boldsymbol{Z}(\boldsymbol{x}, d) > u). \tag{7}$$

For each part, the hierarchical Bayesian model is constructed to integrate the spatial and duration effects. The data layers of the two hierarchical models are characterized by the GPD and the binomial distribution, respectively, where each of $(\sigma_u(\boldsymbol{x}, d), \xi(\boldsymbol{x}, d))$ and $\zeta_u(\boldsymbol{x}, d)$ constitutes a collection of data layer parameters $\boldsymbol{\theta}_1$.

A latent process with a parameter collection $\boldsymbol{\theta}_2$ is then assumed for $\sigma_u(\boldsymbol{x}, d)$ and $\zeta_u(\boldsymbol{x}, d)$ each to drive the spatial and duration dependence structures. Building on the three-layer Bayesian hierarchical structure, the posterior distribution of the parameters can therefore be expressed as

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{Z}(\boldsymbol{x}, d)) \propto p_1(\boldsymbol{Z}(\boldsymbol{x}, d) | \boldsymbol{\theta}_1) p_2(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2) p_3(\boldsymbol{\theta}_2), \tag{8}$$

where $p_j, j = 1, 2, 3$, is the probability density function of the jth model layer conditional on the parameter collection $\boldsymbol{\theta}_j$.

3.2. Threshold exceedances - a Bayesian hierarchical formulation

3.2.1. Data layer

The data layer is characterized by a generalized Pareto distribution,with its original scale parameter re-parameterized as $\phi(\boldsymbol{x}, d) = \log \sigma_u(\boldsymbol{x}, d)$, which helps to remove the positiveness restriction on the parameter domain. This reparameterization of the scale parameter was also adopted in other studies, e.g., Cooley, Nychka, & Naveau (2007), Sang & Gelfand (2009) and Mendes, de Zea Bermudez, Turkman, Pereira, & Vasconcelos (2010) for modeling spatial extremes. Let $Z_k(\boldsymbol{x}_i, d_{ij})$ be the kth recorded amount at location $\boldsymbol{x}_i$, $i = 1, 2, \cdots, s$, with duration d_{ij} hours, $j = 1, 2, \cdots, D_i$, where s denotes the total number of stations and D_i denotes the number of all possible durations at station $\boldsymbol{x}_i$. Let $\phi(\boldsymbol{x}_i, d_{ij})$ and $\xi(\boldsymbol{x}_i, d_{ij})$ denote the log-of-scale and shape parameters at location $\boldsymbol{x}_i$ with duration d_{ij}. The conditional distribution of exceedances over the threshold u can then be characterized as

$$P(\boldsymbol{Z}_k(\boldsymbol{x}_i, d_{ij}) - u > z | \boldsymbol{Z}_k(\boldsymbol{x}_i, d_{ij}) > u) = (1 + \frac{\xi(\boldsymbol{x}_i, d_{ij}) z}{\exp \phi(\boldsymbol{x}_i, d_{ij})})^{-1/\xi(\boldsymbol{x}_i, d_{ij})}.$$

An important assumption here is that the observations after preprocessing are independent, conditional on the GPD parameters; i.e., the dependence among the observations can be completely captured by the GPD and its parameters. The observations with different durations do not usually overlap in time. Even if they do overlap, they are not simple sums of the corresponding records with shorter durations, but are measured separately. The first probability density in equation (8) therefore becomes

$$p_1(\boldsymbol{Z}(\boldsymbol{x}, d) | \boldsymbol{\theta}_1, \boldsymbol{Z}_k(\boldsymbol{x}, d) > u, \ \forall k) = \prod_{i=1}^{s} \prod_{j=1}^{D_i} \prod_{k=1}^{n_{ij}} \frac{1}{\exp \phi(\boldsymbol{x}_i, d_{ij})} (1 + \frac{\xi(\boldsymbol{x}_i, d_{ij}) z_{ijk}}{\exp \phi(\boldsymbol{x}_i, d_{ij})})^{-1/\xi(\boldsymbol{x}_i, d_{ij}) - 1}, \tag{9}$$

where n_{ij} denotes the number of records available at location $\boldsymbol{x}_i$ with duration d_{ij} hours. z_{ijk} denotes the exceedance over threshold of the realization of $Z_k(\boldsymbol{x}_i, d_{ij})$, i.e., $z_{ijk} = z_k(\boldsymbol{x}_i, d_{ij}) - u$, and $\boldsymbol{\theta}_1 = [\boldsymbol{\phi}, \boldsymbol{\xi}]^T$, where $\boldsymbol{\phi}$ is the vector of the log-of-scale parameters $\phi(\boldsymbol{x}, d)$'s and $\boldsymbol{\xi}$ is the vector of shape parameters $\xi(\boldsymbol{x}, d)$'s.

3.2.2. Process layer

Adopting the standard geo-statistical methods, we model the log-of-scale parameter $\phi(\boldsymbol{x}, d)$ as a Gaussian process. This process is assumed to be isotropic and stationary if no significant non-stationarity or anisotropy is detected. The mean structure of the Gaussian process can easily be changed to allow different covariates. We illustrate the mean structure here with a linear function of duration, although we are certainly free to add other spatial covariates. The covariance function is set by a structure corresponding to the exponential variogram without nugget effects. In this model, different axes of point measures, i.e., latitude, longitude, and duration, are given different weights and the weight is left as a parameter to estimate. Therefore, the mean $\boldsymbol{\mu_\phi}$ of the latent Gaussian process is a vector of length $\sum_{i=1}^{s} D_i$, where D_i denotes the number of observations that belong to the station i, $i = 1, 2, ..., s$. The $(\sum_{i=1}^{i_1-1} D_i + j_1)$th entry of $\boldsymbol{\mu_\phi}$ which corresponds to the j_1th observation of the i_1th station is

$$\mu_{\phi,(i_1,j_1)} = E[\phi(\boldsymbol{x_{i_1}}, d_{i_1 j_1})] = \alpha_{\phi,0} + \alpha_{\phi,1} \times d_{i_1 j_1}, \tag{10}$$

and the covariance matrix $\boldsymbol{\Sigma_\phi}$ is the matrix whose $(\sum_{i=1}^{i_1-1} D_i + j_1, \sum_{i=1}^{i_2-1} D_i + j_2)$ entry is

$$\Sigma_{\phi,(i_1,j_1),(i_2,j_2)} = cov(\phi(\boldsymbol{x_{i_1}}, d_{i_1 j_1}), \phi(\boldsymbol{x_{i_2}}, d_{i_2 j_2})) \tag{11}$$

$$= \beta_{\phi,0} \times \exp\left(-\beta_{\phi,1} \times \sqrt{||\boldsymbol{x_{i_1}} - \boldsymbol{x_{i_2}}||^2 + \beta_{\phi,2}^2 \times (d_{i_1 j_1} - d_{i_2 j_2})^2}\right), \tag{12}$$

where $||\boldsymbol{x_{i_1}} - \boldsymbol{x_{i_2}}||$ represents the earth distance between locations $\boldsymbol{x_{i_1}}$ and $\boldsymbol{x_{i_2}}$ calculated by longitude and latitude coordinates. We treat the difference in durations, $(d_{i_1 j_1} - d_{i_2 j_2})^2$, in equation (12) as a proximity. This proximity concept has been widely used in multidimensional scaling (Cox & Cox, 2001). We expect that this extra proximity (on top of $||\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2}||$) can help us to pool information from close locations and similar durations to make statistical inferences. This idea is like the geostatistical weighted regression (Fotheringham, Brunsdon & Charlton, 2002) that weighs local observations to estimate spatial regression parameters. We assume in this paper that the "effect" of duration is proportional to the difference in the duration, i.e., the effect is linear and is explained by the term $\beta_{\phi,2}^2 \times (d_{i_1 j_1} - d_{i_2 j_2})^2$ in equation (12). An alternative parameterization is to allow discrete durations and account each pairwise duration difference separately with a single parameter. This parameterization will lead to a more complicated model setting.

Both $||\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2}||^2$ and $(d_{i_1 j_1} - d_{i_2 j_2})^2$ can be regarded as proximities as in multidimensional scaling (Cox & Cox, 2001). Adding the two proximities directly to characterize the spatial correlation may not adequately reflect the relative importance of the proximities. As such, $\beta_{\phi,2}$ in equation (12) is used to re-scale $(d_{i_1 j_1} - d_{i_2 j_2})^2$ to have a suitable balance between the two proximities. In addition to multidimensional scaling, the re-scaling concept for creating distance measures can be found in Mahalanobis distance (Madia, Kent & Bibby, 1979), which weighs different random variables using a precision matrix. Our approach is to search for the best combination of the proximities using the likelihood inference of $\beta_{\phi,2}$ and other spatial parameters. The estimation of $\beta_{\phi,2}$ shares the same spirit in estimating a norm-dependent distance in Curriero (2007, page 917) or identifying isometric embedding metric space by multidimensional scaling (Schabenberger & Gotway, 2005, chapter 4). Essentially, this weighted duration can be regarded as a third coordinate, in addition to longitude and latitude, that describes the data "location".

Equation (11) is the usual exponential covariance function (Diggle & Ribeiro, 2007, chapter 3). This exponential function is a sensible choice as it is a valid covariance function applied to any negative definite distance measure (Curriero, 2007). We take the usual assumption of isotropy in (11) to set up a parsimonious model to incorporate durations in the statistical inference of other model parameters. We estimate all of the spatial parameters via a likelihood inference so that the isotropy property is best approximated.

The shape parameter $\xi(\boldsymbol{x}, d)$ is known to be very sensitive to the regionally independent variables. If the region considered is not large, a single value assumed for $\xi(\boldsymbol{x}, d)$ over the region might not cause a serious underfitting of the model. For simplicity, therefore, it can be modeled as a constant throughout the whole region, i.e.,

$$\xi(\boldsymbol{x}, d) = \xi. \tag{13}$$

The above characterization of the log-of-scale and shape parameters in the GPD gives the second probability in (8) as

$$p_2(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) = \frac{1}{\sqrt{(2\pi)^s|\boldsymbol{\Sigma}|}} \exp[-\frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu})] \times p_\xi(\xi), \tag{14}$$

where $\boldsymbol{\theta_2} = [\boldsymbol{\alpha}_\phi, \boldsymbol{\beta}_\phi]^T$, $\boldsymbol{\phi}$ is the vector of the log-of-scale parameters $\phi(\boldsymbol{x}, d)$'s, ξ is the value of shape parameters for all of the locations and durations, and $p_\xi(\xi)$ is the prior assumed for ξ, which is taken to be the uninformative prior here.

3.2.3. Prior layer

The third prior layer incorporates the priors of $\boldsymbol{\alpha_\phi}$ and $\boldsymbol{\beta_\phi}$. The priors assigned to the variables are assumed to be independent. No prior information on ϕ is available, so uninformative priors are assigned to the regression parameters, i.e., $\alpha_{\phi,l} \sim Unif(-\infty, \infty), l = 0, 1$. The priors of $\beta_{\phi,l}, l = 0, 1, 2$, are known to be very sensitive to relevant covariates, so informative priors must be adopted. Following the same approach as in Cooley et al. (2007), the prior distribution of $\beta_{\phi,1}$ is set by considering the sill and range of the variogram model, whereas the prior distribution of $\beta_{\phi,0}$ is set by the maximum likelihood estimate obtained by fitting a GPD to each station separately and then fitting an empirical variogram to the $\hat{\phi}'s$. $\beta_{\phi,2}$ serves as a regression parameter here, so an uninformative prior is assigned. The priors in this study are set very loosely because of the unavailability of relevant numerical information or meteorological guidance. A preliminary sensitivity analysis shows that the model is most sensitive to the lower bound of $\beta_{\phi,1}$ that controls the upper bound of the variogram's range, which generally agrees with the results obtained by Cooley et al. (2007). Given the above characterization, and the prior layer probability in (8), the final hierarchical model structure of threshold exceedances can be written as follows.

$$p_3(\boldsymbol{\theta_2}) = p_{\alpha_\phi}(\boldsymbol{\alpha_\phi}) \times p_{\beta_\phi}(\boldsymbol{\beta_\phi}) \propto 1 \times p_{\beta_\phi}(\boldsymbol{\beta_\phi}). \tag{15}$$

3.3. Exceedance frequencies

The exceedance rate ζ_u is another crucial parameter to estimate in equations (6) and (7). The exceedance frequency model is assumed to be independent of the threshold exceedance model. The exceedance rate parameter ζ_u – referring to the probability that an observation is indeed an extreme one (i.e., the cluster maximum after the de-clustering step) – is modeled in the same way as we model the log-of-scale parameter ϕ; i.e., by a hierarchical model with data, process, and prior layers.

The data layer of the exceedance frequency model fits the exceedances frequencies to the binomial distribution with m total trials and a success probability of $\zeta(\boldsymbol{x}, d)$, which is re-parameterized by a logit transformation to remove the [0,1] restriction on its domain; i.e., $\eta(\boldsymbol{x}, d) = \log\{\zeta(\boldsymbol{x}, d)/(1-\zeta(\boldsymbol{x}, d))\}$. The process layer then characterizes the re-parameterized $\eta(\boldsymbol{x}, d)$. Similar to the case of $\phi(\boldsymbol{x}, d)$, we model $\eta(\boldsymbol{x}, d)$ by a Gaussian process with a linear mean structure $\boldsymbol{\mu_\eta}$. The $(\sum_{i=1}^{i_1-1} D_i + j_1)$th entry of $\boldsymbol{\mu_\eta}$, corresponding to the j_1th observation of the D_{i_1} observations in total that belong to the station i_1, is

$$\mu_{\eta,(i_1,j_1)}(\boldsymbol{x_i}, d_{i_1 j_1}) = \alpha_{\eta,0} + \alpha_{\eta,1} \times d_{i_1 j_1}. \tag{16}$$

The covariance structure $\boldsymbol{\Sigma_\eta}$ corresponds to the exponential variogram with respect to covariates such as duration and longitude/latitude. It has its $(\sum_{i=1}^{i_1-1} D_i + j_1, \sum_{i=1}^{i_2-1} D_i + j_2)$ entry as

$$\Sigma_{\eta,(i_1,j_1),(i_2,j_2)} = cov(\eta(\boldsymbol{x_{i_1}}, d_{i_1 j_1}), \eta(\boldsymbol{x_{i_2}}, d_{i_2 j_2})) \tag{17}$$

$$= \beta_{\eta,0} \times \exp\left(-\beta_{\eta,1} \times \sqrt{||\boldsymbol{x_{i_1}} - \boldsymbol{x_{i_2}}||^2 + \beta_{\eta,2}^2 \times (d_{i_1 j_1} - d_{i_2 j_2})^2}\right). \tag{18}$$

The prior layer assigns uninformative priors to $\alpha_{\eta,0}, \alpha_{\eta,1}$, and $\beta_{\eta,2}$, and informative priors to $\beta_{\eta,0}$ and $\beta_{\eta,1}$, following the procedures described in Section 3.2.3.

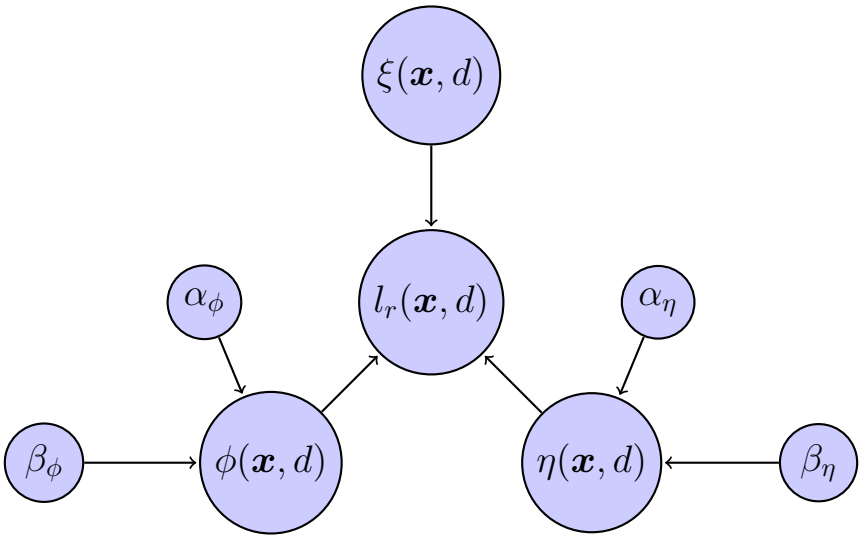


Figure 1: Schematic diagram of the hierarchical model

4. Bayesian inference

4.1. The MCMC sampling scheme

Based on the schematic diagram in Figure 1 for the return level estimation and the posterior distribution (equations 8, 9, 14, 15) obtained in the preceding sections, the inference on the parameters can be performed by the Markov Chain Monte Carlo (MCMC) method, using Metropolis-Hastings (MH) steps within a Gibbs sampler to update the parameters. Following the notation in Section 3.2, we illustrate the detailed posteriors in the threshold exceedances model, where $\theta_{\_}$ denotes all of the parameters in the model except the one of interest.

$$f(\boldsymbol{\phi}(\boldsymbol{x}, d)|\theta_{\_}) \propto \prod_i \prod_j \prod_k g(Z_k(\boldsymbol{x}_i, d_{ij}); \phi(\boldsymbol{x}_i, d_{ij}), \xi) \times N(\boldsymbol{\phi}(\boldsymbol{x}, d); \mu_\phi(\boldsymbol{x}, d), \Sigma_\phi(\boldsymbol{x}, d)),$$

$$f(\alpha_\phi|\theta_{\_}) \propto N(\boldsymbol{\phi}(\boldsymbol{x}, d); \mu_\phi(\boldsymbol{x}, d), \Sigma_\phi(\boldsymbol{x}, d)),$$
$$f(\beta_\phi|\theta_{\_}) \propto N(\boldsymbol{\phi}(\boldsymbol{x}, d); \mu_\phi(\boldsymbol{x}, d), \Sigma_\phi(\boldsymbol{x}, d)) \times p_{\beta_\phi}(\boldsymbol{\beta_\phi}),$$
$$f(\xi|\theta_{\_}) \propto \prod_i \prod_j \prod_k g(Z_k(\boldsymbol{x}_i, d_{ij}); \phi(\boldsymbol{x}_i, d_{ij}), \xi),$$

where $g(\cdot)$ is the probability density function of the GPD, $N(\cdot)$ is the probability density function of the normal distribution, μ_ϕ and Σ_ϕ are specified by equations (10) and (11), and p_{β_ϕ} is the prior assigned to $\boldsymbol{\beta_\phi}$.

The exceedance frequency model can be handled in a similar manner, with η, α_η, and β_η updated successively in a loop.

4.2. An adaptive tuning algorithm

In the threshold exceedance model, the samplings for ϕ and ξ are partitioned for the Gibbs sampler, within which the random walk MH steps are adopted for the updates of α_ϕ, β_ϕ, and ϕ. The shape parameter ξ can also be updated afterwards using random walk MH steps. Within each update, two steps are performed to update θ at the ith iteration of the random walk MH algorithm:

Step 1: Generate a point from θ_i^* from $\theta_i^* = \theta^{[i-1]} + \epsilon$, where $\epsilon \sim N(0, a\Sigma)$.

Step 2: Accept θ_i^* as $\theta^{[i]}$ with the probability $p = \min\{1, f(\theta_i^*)/f(\theta^{[i-1]})\}$. Otherwise, set $\theta^{[i]} = \theta^{[i-1]}$, where f is the target density.

Both the step size a and the covariance structure Σ need to be specified in advance. However, some model parameters are often high-dimensional; for example, the parameter $\boldsymbol{\phi}$ here has a dimension of all possible location-duration combinations. In this case, it is usually very difficult, if not impossible, to optimize each dimension to obtain proper tuning parameters, while inappropriate tuning parameters will make the mixing of MCMC chains very slow. Therefore, adaptive algorithms (Haario, Saksman & Tamminen, 2001; Roberts & Rosenthal, 2009) may be used.

To facilitate the implementation of our approach, we propose an adaptive tuning algorithm to avoid too much overhead in the adaptation. We first run M iterations, $M = 5000$ for instance, which we divide into multiple mini-batches of m iterations, $m = 50$ for instance. At each iteration of this initial trial, the update of a c-dimensional vector is partitioned into independent updates of each of its c dimensions.

Before presenting the adaptive tuning algorithm, let us start with some definitions. Let $b = (b_1, b_2, ..., b_c)$ be the step size parameter to be tuned and let m be the size of a mini-batch. Define l as a pre-specified constant in $[0, 1]$, serving as a factor for discounting or enlarging the step size parameter. Denote θ_k as the kth dimension of θ and $q_{(i-1)/m,k}$ as the acceptance rate of θ_k in the $((i-1)/m)$th mini-batch, i.e., from the $(i-m)$th to the $(i-1)$th iteration. $\theta^*_{k,i}$ is the ith proposed candidate of θ_k and $\theta_k^{[i]}$ is the ith draw of θ_k. Let f refer to the target density function and let R represent the optimal acceptance rate of the algorithm; for example, the optimal acceptance rate of univariate random walk MH steps is approximately 0.44 (Rosenthal, 2011).

The detailed algorithm for updating a c-dimensional vector θ at the ith ($i \neq 1$) iteration then works as follows.

Step 1: If $i \equiv 1 \mod m$, then update b. For each $k \in \{1, 2, 3, ..., c\}$, if $q_{(i-1)/m,k} \leq R$, then $b_k \leftarrow b_k \cdot l$; if $q_{(i-1)/m,k} > R$, then $b_k \leftarrow b_k / l$. Otherwise, go to step 2.

Step 2: For each $k \in \{1, 2, 3, ..., c\}$,

Substep 2.1: generate a point $\theta^*_{k,i}$ from $\theta^*_{k,i} = \theta_k^{[i-1]} + b_k\epsilon$, where $\epsilon \sim N(0, 1)$, and

Substep 2.2: accept $\theta^*_{k,i}$ as $\theta_k^{[i]}$ with probability $p = \min\{1, f(\theta^*_{k,i})/f(\theta_k^{[i-1]})\}$. Otherwise, set $\theta_k^{[i]} = \theta_k^{[i-1]}$.

The idea of this algorithm is to tune the step size b so that the acceptance rate is approximately R: we decrease b when the acceptance rate in the last mini-batch is too low and we increase b when the last acceptance rate is too high. The idea behind this algorithm's adaptation is similar to that of Roberts & Rosenthal (2009), but differs in its method of adaptation. We adjust the step size b by discounting or enlarging it by a factor l, whereas they update it by adding or subtracting an adaptation amount δ.

The enlarging and discounting adjustment is preferable here because it enables b to reach a proper value much faster than addition and subtraction. It is also more sensitive to those proposed values for the upper layer parameters that deviate too largely from the target distribution, generating null values in computing the probabilities of acceptance and rejection. The enlarging and discounting adjustment responds swiftly to these proposals, avoiding the chain of the upper layer parameters getting stuck at a value for a long time. Hence, this algorithm is particularly suitable for inferences in hierarchical structured models.

After M iterations – M usually equal to 5000 – we can often see a sign indicating good mixing from the time series plots of the draws, the sample autocorrelation function of the draws, and the adaptation record of b. In particular, if we spot that b_k is mainly alternating between two values, say b_0 and $b_0 \cdot l$, it has probably reached a certain steady state that will allow us to revise the proposal distribution to avoid too much overhead caused by adaptation. This is another instance where the multiplication and division adjustment works better than the addition and subtraction adjustment, probably because the adaptation amplitude of b is relatively larger in the former than in the latter, and hence the former adaptation is coarser but renders the sign of stability easier to spot.

To reduce the overhead of adaptation in addition to updating the c dimensions of θ one at a time, we revise the sampling scheme from the $(M+1)$th iteration by applying the random walk kernel MH algorithm with a newly extracted Σ_M in place of the original Σ, as follows.

Step 1: Generate a point from θ_i^* from $\theta_i^* = \theta^{[i-1]} + \epsilon$, where $\epsilon \sim N(0, a_M\Sigma_M)$.

Step 2: Accept θ_i^* as $\theta^{[i]}$ with probability $p = \min\{1, f(\theta_i^*)/f(\theta^{[i-1]})\}$. Otherwise, set $\theta^{[i]} = \theta^{[i-1]}$, where f is again the target density.

Setting	$\beta_{\phi,1}$	$\beta_{\phi,2}$	ξ	Setting	$\beta_{\phi,1}$	$\beta_{\phi,2}$	ξ
1	0.01	0.2	0.1	7	0.01	0.2	-0.1
2	0.01	0.3	0.1	8	0.01	0.3	-0.1
3	0.01	0.4	0.1	9	0.01	0.4	-0.1
4	0.02	0.2	0.1	10	0.02	0.2	-0.1
5	0.02	0.3	0.1	11	0.02	0.3	-0.1
6	0.02	0.4	0.1	12	0.02	0.4	-0.1

Table 1: True parameters of the simulation settings

The new Σ_M is set as the sample covariance of the first M draws or only the draws after we see a sign of good mixing, as mentioned above. As for the value of a_M, it is known from Roberts, Gelman, & Gilks (1997) and Roberts & Rosenthal (2001) that the proposal $N(\theta, (2.38)^2/c \cdot \Sigma)$ is optimal in a large-dimensional context, so it is advisable to set the a_M as $(2.38)^2/c$ and, if necessary, it can be further tuned in a similar adaptive way to the tuning algorithm proposed above. By switching back to this multi-dimensional update algorithm, we can save on the computational costs caused by updating the c dimensions of θ one at a time. This idea of revising the sampling scheme after M iterations is similar in spirit to So & Chan's (2014) method, but differs in the way it captures the covariance structure. Although adaptively updating each dimension of θ one at a time bears some computational costs in the initial iterations, it enables us to obtain a covariance structure closer to the target density and facilitates the mixing of the later random walk kernel MH steps to a large extent.

The effective gains of this adaptive tuning algorithm are further illustrated in Section 5.2 where a comparison with the Roberts & Rosenthal (2009) algorithm is presented.

5. Simulation studies

5.1. Performance of the Bayesian estimator

To examine the performance of our Bayesian estimator, we perform the following simulation study. With hints from the real precipitation dataset in Section 6, we consider 12 sets of $(\beta_{\phi,1}, \beta_{\phi,2}, \xi)$ while keeping the same regression parameters $(\alpha_{\phi,0}, \alpha_{\phi,1}, \beta_{\phi,0})$ in the model, i.e., $\alpha_{\phi,0} = 4.2, \alpha_{\phi,1} = -.01$, and $\beta_{\phi,0} = .05$. The true parameters are shown in Table 1. Settings 1-6 have positive ξ's and settings 7-12 have negative ξ's. This setting design allows us to assess the performance of the parameter estimation under different scales and shapes of the exceedance-over-threshold distribution. The sample size of $n = 4000$ is adopted. For each of the 12 settings, 100 replications are generated. We apply the MCMC sampling scheme described in Section 4. Three parallel chains of 25000 iterations were run for each model with the first 5000 iterations regarded as the burn-in time.

Table 2 summarizes the posterior inference results from fitting the models for each of the settings listed in Table 1. Specifically, it presents the grand means of the posterior means of the parameters obtained in the 100 replications and the corresponding Bayesian coverage rates under each setting. Based on the results in Table 2, our sampling scheme generally performs very well, with the posterior mean estimates of the parameters very close to the true parameters in general and the Bayesian coverage rates falling below 90% only occasionally.

5.2. Performance of the adaptive tuning algorithm

The inclusion of multiple durations in our model, which increases the model complexity and the processing of more observations, poses challenges for the mixing of MCMC chains. However, the adaptive tuning algorithm proposed in Section 4 relieves the issue to a large extent: If we look at the first 5000 draws immediately following the revision of the sampling scheme at the 5000th iteration, the autocorrelation functions of the MCMC chains decrease rather quickly as the lag lengths increase. Autocorrelation functions generally fall below the dashed line delimiting the significant and insignificant correlations before the lag length increases to 15, which implies a reasonable convergence rate.

Setting	$\alpha_{\phi,0}$	$\alpha_{\phi,1}$	$\beta_{\phi,0}$	$\beta_{\phi,1}$	$\beta_{\phi,2}$	ξ
1	4.190804	-0.009882	0.042450	0.011698	0.222013	0.101481
	(97%)	(98%)	(95%)	(98%)	(95%)	(91%)
2	4.229490	-0.010343	0.042257	0.011685	0.293042	0.098483
	(94%)	(95%)	(98%)	(95%)	(97%)	(93%)
3	4.197297	-0.009950	0.045819	0.010816	0.365163	0.102796
	(95%)	(93%)	(94%)	(94%)	(92%)	(89%)
4	4.222669	-0.009676	0.044230	0.021850	0.215137	0.101827
	(94%)	(96%)	(91%)	(87%)	(95%)	(90%)
5	4.198273	-0.009921	0.047597	0.019821	0.268527	0.099629
	(93%)	(89%)	(95%)	(91%)	(93%)	(90%)
6	4.204666	-0.010081	0.048481	0.018891	0.389439	0.101701
	(95%)	(88%)	(93%)	(92%)	(96%)	(96%)
7	4.219011	-0.009553	0.042179	0.011687	0.202524	-0.098491
	(95%)	(98%)	(96%)	(98%)	(95%)	(90%)
8	4.190688	-0.009790	0.049523	0.009767	0.284785	-0.100964
	(94%)	(98%)	(94%)	(92%)	(95%)	(92%)
9	4.201665	-0.009798	0.045266	0.009747	0.378498	-0.103005
	(90%)	(95%)	(89%)	(96%)	(92%)	(90%)
10	4.180387	-0.010167	0.047371	0.021808	0.194642	-0.100115
	(91%)	(93%)	(90%)	(93%)	(98%)	(90%)
11	4.164973	-0.010648	0.043078	0.020855	0.277216	-0.100914
	(91%)	(87%)	(96%)	(93%)	(92%)	(89%)
12	4.182385	-0.009883	0.046238	0.019912	0.399140	-0.100216
	(92%)	(88%)	(91%)	(89%)	(98%)	(92%)

Table 2: Means of the posterior means of unknown parameters over 100 replications for all of the settings with Bayesian coverage rates in paratheses

Variable	Proposed		Roberts & Rosenthal (2009)	
	ACT	Avr Sqr Dist	ACT	Avr Sqr Dist
$\alpha_{\phi,0}$	6.1778	3.14E-02	6.7697	3.24E-02
$\alpha_{\phi,1}$	6.6185	6.30E-06	11.8497	1.49E-06
$\beta_{\phi,0}$	6.0038	2.98E-05	7.0011	2.35E-05
$\beta_{\phi,1}$	3.8631	2.02E-06	4.8690	1.70E-06
$\beta_{\phi,2}$	4.4370	6.35E-03	10.9606	5.54E-03
ξ	7.3913	1.34E-05	19.3609	1.13E-05

Table 3: The integrated autocorrelation times (ACT) and average squared jumping distances (abbreviated above as *Avr Sqr Dist*), both taken after discarding the first fifth of the run as burn-in and then averaged among the parallel chains under all of the settings, for both the proposed adaptive tuning algorithm and the Roberts & Rosenthal (2009) adaptive algorithm.

Moreover, we compare the performance of our adaptive tuning algorithm to the one proposed in Roberts & Rosenthal (2009). For each simulated dataset, we ran 10 parallel chains of 25000 iterations, keeping the same the initial values, initial random walk step sizes, and initial adaptation amounts for both algorithms. Table 3 presents, averaged among the parallel chains under all of the parameter settings, the integrated autocorrelation times (ACT) and the average squared jumping distances (abbreviated in the table as *Avr Sqr Dist*), for the six base parameters after discarding the first fifth of the run as burn-in for both algorithms. Both the integrated autocorrelation times and the average squared jumping distances are criteria for the efficiency of the sampling algorithms and were used in Roberts & Rosenthal (2009) as well. The ACT estimates the average number of iterations needed for an independent sample to be drawn and thus controls the statistical efficiency of our Monte Carlo estimates (Sokal, 1996). The average squared jumping distance, on the other hand, serves as an empirical estimate of the expected squared jumping distance (ESJD). For a one-dimensional target distribution, for instance,

$$ESJD \equiv E||\theta_{t+1} - \theta_t||^2 = 2(1 - \rho_1)\text{Var}(\theta_t), \tag{19}$$

where θ_t is the sample value drawn at the tth iteration, $\text{Var}(\theta_t)$ denotes the variance of our drawn samples under independent sampling, and ρ_1 is the first-order autocorrelation of the Markov chain for our drawn samples (Pasarica & Gelman, 2003). A larger average/expected squared jumping distance is therefore preferred in the same way as a smaller first-order autocorrelation of the Markov chain is preferred for an efficient sampling algorithm.

Table 3 shows that the integrated autocorrelation times are generally smaller in the chains generated by the proposed algorithm. In particular, the autocorrelation times of the $\beta_{\phi,2}$ chain and the ξ chain generated by the proposed adaptive tuning algorithm are 60% smaller (better) than the chains generated by the adaptive algorithm proposed in Roberts & Rosenthal (2009). The average squared jumping distances in the proposed algorithm are also mostly larger, except in the chain of $\alpha_{\phi,0}$ where the two algorithms performed similarly. Thus adapting the MCMC algorithm, i.e., finding the appropriate proposal scalings for each coordinate of each parameter, by enlarging or discounting by a factor instead of adding or subtracting an amount is generally able to increase the speed of the mixing of the chains.

Looking beyond the chains of the base parameters, in more than 30% of the parallel chains generated by the Roberts & Rosenthal (2009) algorithm, multiple coordinates of the log-of-scale parameter vector became stuck at a value for a long time. This value is usually very far from the target distribution, generating null values in computing the probabilities of acceptance and rejection, but is accepted due to a low probability event earlier. This renders the whole chain of log-of-scale parameter vector insufficiently informative to the inference and calls for additional tuning of the adaptation amount by hand. In contrast, none of the parallel chains generated by the proposed algorithm suffered from this issue. This shows that the proposed algorithm's faster rescaling response to the upper layer parameter proposals – faster in the sense that exponential increase/decrease is faster than linear increase/decrease with respect to the number of steps – also facilitates the mixing of the upper layer parameters and hence the model parameters in all of the layers in general, while minimizing the need for user interventions in the adaptive MCMC algorithm.

Finally, the algorithm seeks to scale the proposal standard deviations properly and leave it mostly unchanged when it reaches the optimal value. Therefore, the application of the random walk kernel MH algorithm in the proposed adaptive tuning is capable of largely reducing the overhead of adaptation after the algorithm has found the optimal step sizes, usually before the end of the 5000 burn-in steps. The computational costs of the inference can thus be largely reduced, even with a big dataset.

6. Application to precipitation extremes

6.1. Study region and data structure

Our study region, Hong Kong, is located in the South China Sea at the mouth of the Pearl River Delta. It is a mountainous area with steep slopes and lowlands, especially in the northwest. It lies between 113°and 115°E and 21.5°and 23.5°N. The principal peaks of Hong Kong range from around 550 m to around 950 m in height. A narrow body of water, Victoria Harbor, separates Hong Kong Island and Kowloon Peninsula,

Duration (hours)	1	6	24
Mean precipitation (mm)	1.161	2.319	13.024

Table 4: Mean precipitation by duration

and is one of the deepest natural maritime ports in the world. The climate of Hong Kong is subtropical and monsoonal, tending towards temperate for nearly half of the year. The mean annual rainfall ranges from around 1400 mm to more than 3000 mm, about 80% of which occurs between May and September. The temperature occasionally drops below 10°C (50°F) and the lowest temperature recorded by Hong Kong Observatory is 0°C (32°F). Therefore, the precipitation in Hong Kong mainly consists of convective rainfall while snow and tornadoes are rare.

The data are from the Atmospheric and Environmental Real-time Database of the Institute for the Environment (IENV) at the Hong Kong University of Science and Technology (HKUST). The database includes 89 local stations in the chosen region with the earliest precipitation data available from January 1982. Figure 2 shows a map of the stations. There are certain periods of time for which no observations are available, but most stations have relatively complete records from January 1992. All of the station records have missing values and invalid data. Here, we use the data up to October 31, 2012, with 5,439,120 observations in total including 441,368 nonzero observations. Each station is given with its elevation, longitude, and latitude. Each precipitation observation is accompanied by the time it was recorded and the duration of its coverage. The elevations of the stations range from 5 m to 450 m, with over 90% under 150 m. The durations of available observations are 1, 3, 6, 12, and 24 hours. Around 97.27% of the observations fall into the 1-hour category, 1.78% are 6-hour, 0.69% are 24-hour, 0.16% are 12-hour, and 0.1% are 3-hour. Among non-zero observations, 84.60% fall into the 1-hour category, 7.47% are 24-hour, 7.43% are 6-hour, and 0.50% are 12-hour. Given the above, we include only 1-hour, 6-hour, and 24-hour data in our illustration.

The number of 1-hour, 6-hour, and 24-hour records here does not correspond exactly to the ratio of 24:4:1 due to some intricacies in the system's measurement. One possible reason is that a single missing 1-hour observation will make both the corresponding 6-hour and 24-hour observations missing. Some stations, moreover, only have 1-hour or 6-hour observations. Also the 6-hour and 24-hour records in this database are not simple sums of the corresponding 1-hour observations and hence are properly measured separately. Due to these multiple intricacies in the data structure, we did not try to merge the 1-hour observations manually to produce 6-hour and 24-hour observations.

The data are rounded to the first decimal place. Table 4 and Figures 3 and 4a give the mean precipitation by duration and the estimated density functions of both the total precipitation and the exceedances over a 100 mm threshold, which together offer a comprehensive picture of precipitation distributions by duration. The table and figures show that the extreme precipitation observations of different durations do not generally share the same distribution but differ in scale. Figure 3, in particular, suggests that the 24-hour records are generally larger than 6-hour ones and the 6-hour ones are also generally larger than the 1-hour ones. If we look only at the extreme records, however, the 1-hour, 6-hour, and 24-hour records do not always follow the same sequence. The significance of the duration effect in the mean or covariance structure of extreme values is thus of special interest. This is further quantitatively studied in Section 6.4.

Our model, specifically equation (11), assumes isotropy and stationarity of the latent process. However, it is hard to detect any nonstationarity with the current dataset of around 25 years. As for isotropy, Erhardt & Smith (2014) used a Smith process to check the assumption. In this paper, we turn to check for directional dependence in the empirical semivariogram of the process. In particular, because the log-of-scale parameters ϕ's are not observed, we fit a generalized Pareto distribution to each station-duration pair independently and then fit an empirical variogram to the maximum likelihood (ML) estimates $\hat{\phi}$'s. Figure 5a shows the empirical variogram fitted to the $\hat{\phi}$'s as well as the four covariance models – spherical, exponential, Gaussian, and linear – fitted to the empirical variogram. The exponential model, giving the smallest root mean squared error (RMSE), fits the sample variogram the best. Therefore, Figure 5b, together with this omnidirectional exponential model (in solid lines), shows the semivariograms in the directions of N 0°E, N 45°E, N 90°E, and N 135°E with angular tolerance of 22.5 and no bandwidth limit. The directional semivariograms are

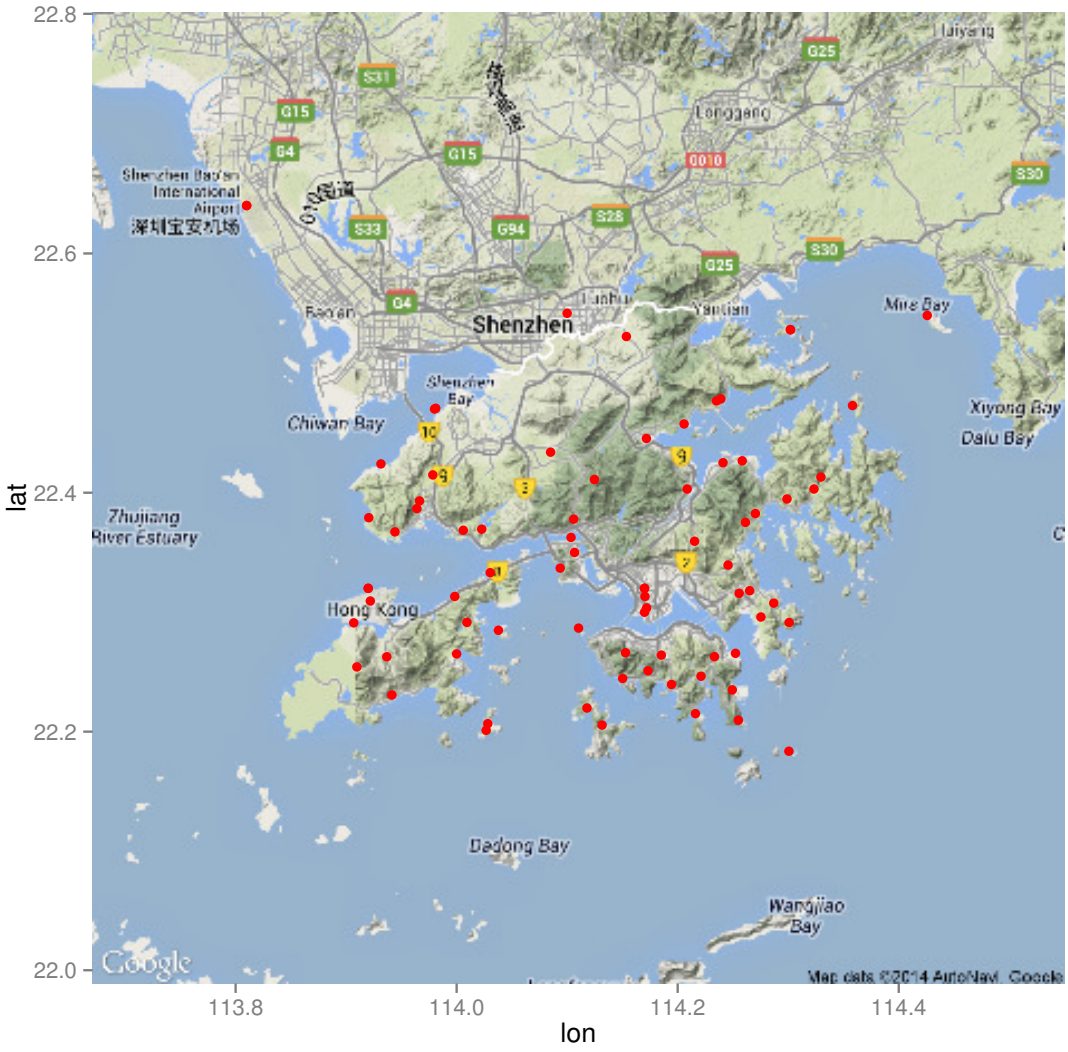
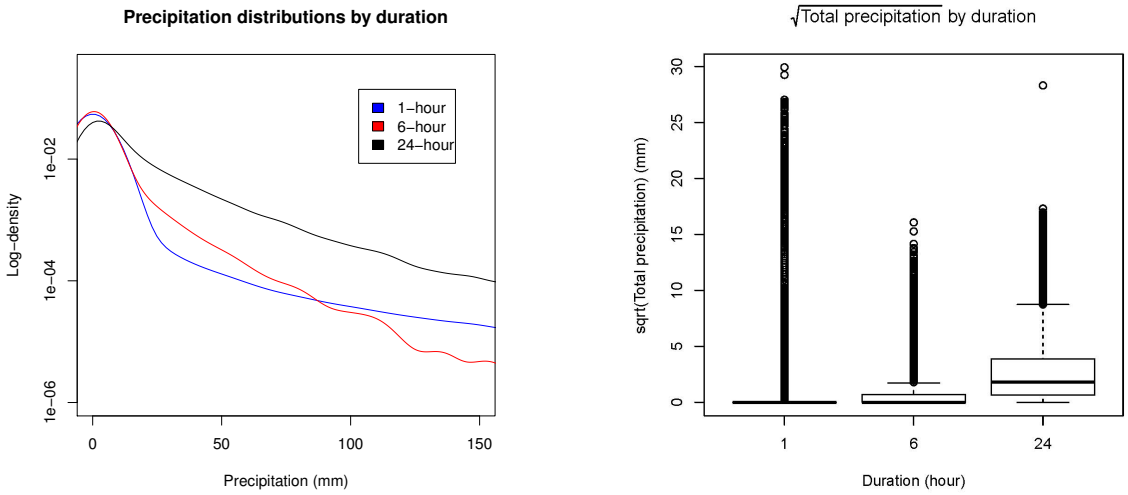


Figure 2: Stations in the chosen region around Hong Kong



(a) Log-density plot of total precipitation by duration

(b) Box plot of $\sqrt{\text{total precipitation}}$ by duration

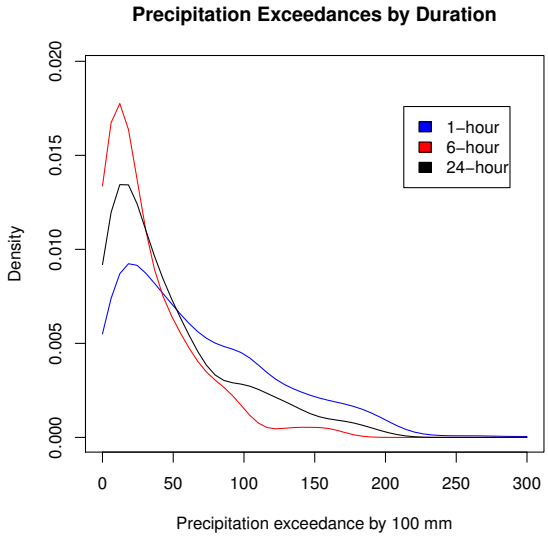
Figure 3: Precipitation distributions by duration

noisier since a limited number of data pairs can be admitted for estimation in each direction. However, they do not show overwhelming evidence of anisotropy. Dealing with possible anisotropy or nonstationarity as in Blanchet & Davison (2011) and Huser & Genton (2014) with multiple durations can be an extension for further research.

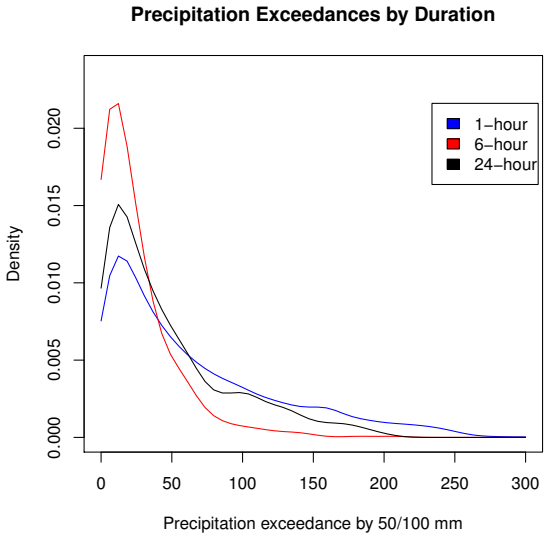
6.2. Threshold selection

The issue of threshold choice is an intricate balance between bias and variance. The standard practice is to adopt as low a threshold as possible, subject to the condition that the model provides a reasonable approximation. Two methods are available for this purpose: an exploratory technique carried out prior to model estimation and another that assesses the stability of the parameter estimates based on the fitting of models across a range of different thresholds. After an initial inspection of the data, we first consider a threshold of 100 mm, which is meteorologically meaningful in terms of extreme precipitation under the subtropical monsoonal climate, rendering it an attractive candidate. Moreover, an initial sensitivity analysis shows that taking thresholds of 100 mm, 110 mm, 120 mm, 130 mm, 140 mm, and 150 mm gives similar point estimates for the GPD parameters, implying that the threshold choice of 100 mm is sufficiently high for only a small bias and adequately low for a reasonable estimation. With the threshold set at 100 mm, all of the observations exceeding this threshold account for 1.5% of the data, including the zero observations.

As we are considering a dataset that allows multiple durations, one potential problem is that only one threshold is applied to all of the durations. The single threshold of 100 mm may be too high for precipitation records with a shorter duration, so that only a few particularly extreme observations are kept in the dataset while the majority of the extremes are lost; this would be particularly noticeable in the extreme frequency counts. In this case, the model might fail to capture all of the characteristics of precipitation with shorter durations. The option of selecting multiple thresholds for different durations therefore follows naturally, hence we also apply our approach to a multiple-threshold setting. In particular, given the limitations on the data structure, we consider what is essentially a double-threshold model for illustration, a 50-mm threshold for 1/6-hour observations and a 100-mm threshold for 24-hour observations, the former of which is selected in a similar way to that described above. Figure 4b shows the estimated density function of exceedances in the two-threshold setting in comparison with Figure 4a. These two figures show that the empirical density functions under the two settings have similar relative locations but differ slightly in scale. Given this, it is possible that the model possesses a certain robustness regardless of how many thresholds we choose to include in the model. This is further quantitatively studied in Section 6.4.

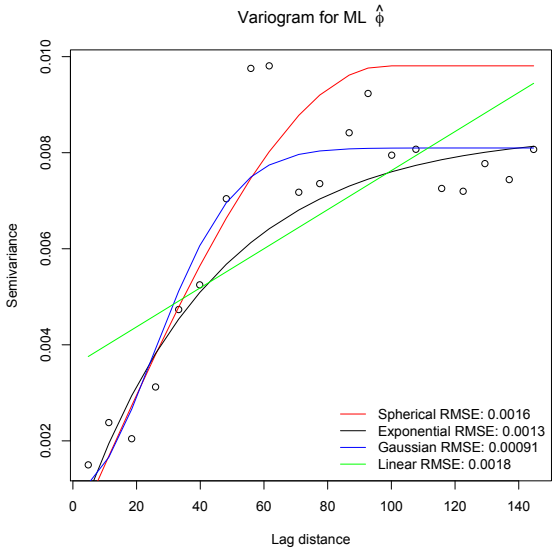


(a) Exceedance distributions by duration with a single 100 mm threshold

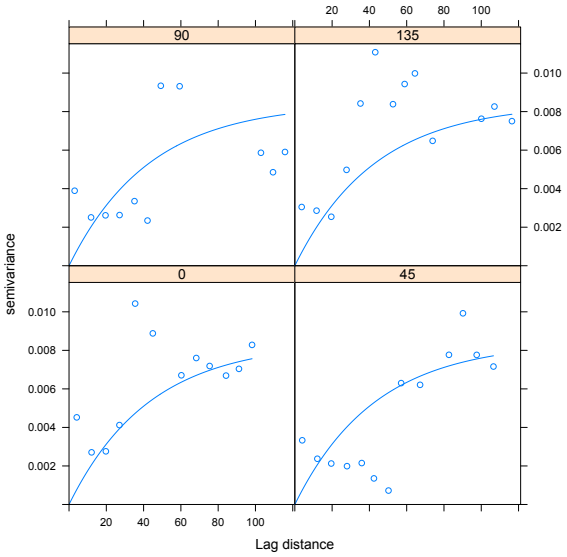


(b) Exceedance distributions by duration with 50/100 mm thresholds

Figure 4: Exceedance distributions by duration



(a) Empirical variogram fitted to ML estimates of ϕ



(b) Directional semivariograms for ML estimates of ϕ

Figure 5: Anisotropy detection

	$D(\bar{\theta})$	$\overline{D(\theta)}$	p_D	DIC	IC1	IC2
S_1 : full model	148377.7	148444.5	66.8	148511.3	148578.1	148610.5
$S_2 : \alpha_{\phi,1} = 0$	148386.1	148458.0	71.9	148529.9	148601.8	148622.0
$S_3 : \alpha_{\phi,1} = \beta_{\phi,2} = 0$	148457.3	148518.6	61.3	148579.9	148641.2	148658.6

Table 5: DIC, IC1, and IC2 scores of the three GPD hierarchical models tested

6.3. Data preprocessing

The data layer of the threshold exceedance model needs, conditional on the process and prior layers below, independently and identically distributed (i.i.d) data, or other conditions such as stationarity plus long-term asymptotic independence. To ensure long-term asymptotic independence, the data are de-clustered by deleting all of the extreme clusters except for the highest measurement. As is discussed in Section 6.1, although the time coverage of the measurements with different durations might overlap, the measurements are taken separately and are not simple sums of the corresponding 1-hour or 6-hour observations. In this study, seasonal effects are not taken into consideration because we did not detect any apparent seasonal effects in the dataset. Note also that temporal trends are left out of the model as the time span of the available data is probably not long enough to characterize long-term trends. The spatial independence is ensured by testing the annual maximum residuals of the stations using a first-order variogram (Cooley, Naveau & Poncet, 2006). The results show no significant dependence beyond a distance of 16 km and a low dependence within the distance. Most stations in the region have distances larger than 16 km and the effect of a slight violation of spatial independence is temporarily ignored.

6.4. Numerical results and model comparisons

We apply our approach to compare the three models under the one-threshold setting: S_1 (full model), S_2: $\alpha_{\phi,1} = 0$, and S_3: $\alpha_{\phi,1} = \beta_{\phi,2} = 0$. Specifically, S_1 considers the duration effect in both the mean and the covariance structures, S_2 has the duration effect only in the covariance structure, and S_3 leaves the duration effect out of both structures. Following the procedures described in Sections 3.2.3 and 3.3, the priors of β_ϕ, chosen as simple truncations of the parameter space, are set as follows: $\beta_{\phi,0} \sim [.001, .07]$ and $\beta_{\phi,1} \sim Unif[.004, .018]$. The comparison is performed using Bayesian information criteria. The deviance information criterion (DIC) (Spiegelhalter, Best, Carlin & Van Der Linde, 2002) takes into account both the model fit $\bar{D}$ and the model complexity $p_D = \overline{D(\theta)} - D(\bar{\theta})$, where θ denotes the model parameters and $D(\theta) = -2\log(f(\text{data}|\theta)) + \text{constant}$. The overall DIC score is the sum of $\bar{D}$ and p_D, and the models with smaller DIC scores are preferred. Ando (2012) generalized the Bayesian predictive information criterion in (Ando, 2007) into another two information criteria, IC1 and its asymptotic version IC2, namely IC1= $\bar{D}+2p_D$ and IC2 = $\bar{D}+2p$, where p is the dimension of θ. The two criteria can be regarded as bias-corrected versions of the DIC for coping with the overfitting tendency of the DIC. The DIC, IC1, and IC2 all give relative measures for model comparison rather than absolute measures such as the posterior probability of a model. The DIC, IC1, and IC2 scores of the three settings S_1, S_2, and S_3 are presented in Table 5. The DIC, IC1, and IC2 scores of the three model settings all differ by more than 10, which helps to rule out S_2 and S_3. This implies that the duration effect is significant in both the mean and covariance structures.

To further understand how much we have gained or lost from adopting the proposed model with multiple durations in the place of the contemporary models that handle only the data with a single duration, we compare the fit of the data by including in our inference (1) only the 1-hour data, (2) only the 1-hour and 6-hour data, and (3) the 1-hour, 6-hour, and 24-hour data.

The measures of DIC, IC1, and IC2 for the model fit, however, no longer apply here. A major component of these measures is $\overline{D(\theta)} = -2\overline{\log(f(\text{data}|\theta))}$, which relies very much on the number of observations included, and the three scenarios to be compared indeed involve quite different numbers of observations. Therefore, we do the comparison here by the fit of the data to the estimated scale and shape parameters, that is, wherever applicable, $-2\log(f(\text{1-hour data}|\hat{\boldsymbol{\phi}}, \hat{\xi}))$, $-2\log(f(\text{6-hour data}|\hat{\boldsymbol{\phi}}, \hat{\xi}))$, and

Setting	Data	# of thresholds	1-hour data fit	6-hour data fit	24-hour data fit
S_1	1-hour	1	142589.50	/	/
S_2	1/6-hour	1	142591.70	678.08	/
S_3	1/6/24-hour	1	142568.72	670.60	5625.47
S_4	1-hour	2	142600.90	/	/
S_5	1/6-hour	2	142596.39	688.72	/
S_6	1/6/24-hour	2	142583.04	682.97	5684.20
S_7	1-hour	3	142597.70	/	/
S_8	1/6-hour	3	142595.98	688.73	/
S_9	1/6/24-hour	3	142567.87	689.24	5659.28

Table 6: The data fit of nine model settings tested

$-2\log(f(\text{24-hour data}|\hat{\boldsymbol{\phi}},\hat{\xi}))$. Fixing the number of thresholds we use in the data, the number of observations falling into each of the 1-hour, 6-hour, and 24-hour categories should remain the same across the three scenarios.

Across the settings with different numbers of thresholds, the comparison of the data fit is only done for the data that exceeds the maximum threshold in the category, that is the part of the data in each category shared across the different settings. Here the comparison is done for (1) taking 100 mm as a threshold for all the 1-hour, 6-hour, and 24-hour data, (2) taking 50 mm as a threshold for both the 1-hour and 6-hour data and 100 mm for the 24-hour data, and (3) taking 25 mm as a threshold for the 1-hour data, 50 mm for 6-hour data, and 100 mm for the 24-hour data. The data fit is compared only for those 1-hour, 6-hour, and 24-hour observations exceeding the 100 mm threshold. The thresholds of 25 mm, 50 mm and 100 mm are chosen in the same way as in Section 6.2.

The comparison of all the $3 \times 3 = 9$ scenarios is presented in Table 6. Under the one-threshold setting, it shows that we lost a little bit of the 1-hour data fit by adding the 6-hour data to the 1-hour data in our inference. However, by including the 24-hour data, we witnessed a substantial gain in both the 1-hour data fit and the 6-hour data fit. Under the two-threshold setting, in contrast, both the inclusion of the 6-hour data and the 24-hour data improve the fit of the data present in the previous model. Under the three-threshold setting, the inclusion of the 6-hour data improved slightly the 1-hour data fit, and the further inclusion of the 24-hour data made another substantial improvement to the 1-hour data fit but kept the 6-hour data fit almost unchanged.

In general, the inclusion of the additional data with a different duration, even if this gives as few as 4% more extreme observations as in the case of adding the 24-hour data under the one-threshold setting, generally improves the fit of the data that are already included in the inference. The addition of the 6-hour and 24-hour data under the two-threshold and three-threshold settings always increases the amount of extreme observations by more than 5%. Therefore, the earlier loss caused by the addition of the 6-hour data under the one-threshold setting is probably due to the too few new observations supplied by the additional data – the additional 6-hour data provided only 0.5% more extreme observations – and thus may not be able to serve as sufficiently representative of the phenomenon associated with 6-hour observations. This has probably rendered the estimation rather unstable and incurred a modicum of loss in the 1-hour data fit afterwards.

Considering the one-threshold, two-threshold, and three-threshold settings, we see from Table 6 that the three models produce GPD scale and shape parameter estimates that fit similarly well to the part of the data they share, which suggests certain robustness of the model at least in this particular application. The best 1-hour data fit occurs in settings S_3 and S_9 while both the best 6-hour data fit and the best 24-hour data fit occur in S_3. Hence, we may conclude that the one-threshold setting is preferred here.

Apart from looking at the model selection scores and the data fit, in Table 7 we also present the posterior estimates (mean, standard deviation, and 95% credible interval) of the base parameters α_ϕ, β_ϕ, and ξ in S_1 (full model). $\beta_{\phi,2}$ – the regressed weight of duration in the covariance structure – and the magnitude of the covariates show that the duration effect has a larger share of total effects than longitude and latitude,

	$\alpha_{\phi,0}$	$\alpha_{\phi,1}$
Mean	4.312	-0.007871
2.5% Quantile	3.895	-0.015856
97.5% Quantile	4.750	-0.000024
SD	0.216	0.004024

(a) Posterior estimates of α_ϕ

	$\beta_{\phi,0}$	$\beta_{\phi,1}$	$\beta_{\phi,2}$
Mean	0.049027	0.013134	0.265092
2.5% Quantile	0.008264	0.005693	0.034132
97.5% Quantile	0.069251	0.017814	0.488506
SD	0.016308	0.003458	0.137221

(b) Posterior estimates of β_ϕ

Mean	2.5% Quantile	97.5% Quantile	SD
-0.060583	-0.069846	-0.050289	0.004925

(c) Posterior estimates of ξ

Table 7: Posterior estimates of S_1 (the full model)

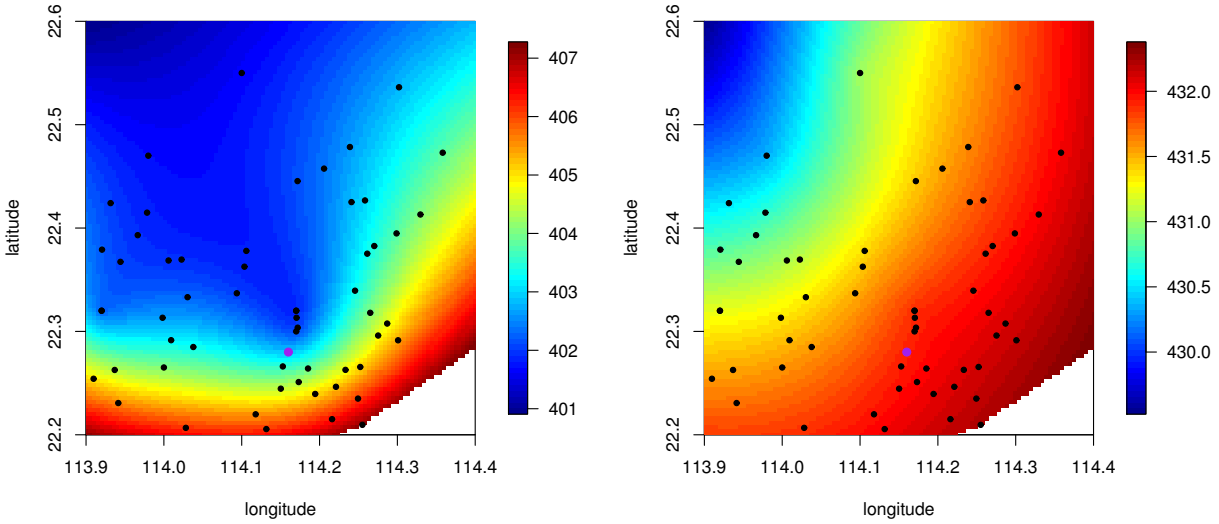
indicating its significance. Surprisingly, we observe in the posterior estimates that the coefficient $\alpha_{\phi,1}$ of duration in the mean structure takes a negative value. This is, at first sight, counter-intuitive because 24-hour precipitation exceedances should in principle be larger than the 6-hour ones. However, given the small magnitude of the $\alpha_{\phi,1}\times$ duration compared with $\alpha_{\phi,0}$, this indicates that most extreme precipitations are strong in intensity but short in duration, and therefore this extreme behavior does not always accumulate or intensify over time. This also identifies a potentially interesting direction for future studies of extreme precipitation with short duration alone.

6.5. Return level maps

Site-specific return level estimates can be obtained by plugging the MCMC estimates into equation (5). The further interpolation of return-level estimates can be performed by interpolating the relevant parameters, i.e., ϕ, α_ϕ, β_ϕ, and η, α_η, β_η. The sampling distribution is produced by the standard conditional draws from the Gaussian process – given that the mean and covariance structure determined by the values of α's and β's, ϕ, and η are sampled separately from its conditional distribution given its current values. The sampling is performed for each iteration of the MCMC. The entire collection of draws from each variable ϕ, ξ, and η gives the posterior samples of return levels via equation (5) and hence produces the corresponding return level maps.

We illustrate in Figures 6, 7 and 8 the 25-year 6-hour and 24-hour return level maps . The black dots in the plots represent the stations in the region and the purple dot signifies Central district, Hong Kong. The figures show that the variabilities captured by the return-level maps are generally similar if restricted to a particular duration, which may be due partially to the covariance structure assumed in the latent Gaussian process. In particular, both 6-hour and 24-hour plots show heavier precipitation in the southeastern region and lighter precipitation in the northwestern region. This subtle difference in the pattern of variation suggests, however, that the central region experiences intense 6-hour precipitation slightly more frequently than the rest of the region. After all, the estimated volumes of 6-hour and 24-hour extreme precipitation do not differ by a large margin. Seemingly surprising, this is consistent, if we restrict our attention to extreme values only, with the precipitation distributions by duration presented in Figure 3 earlier. One possible explanation can be that most extreme precipitations occur within a 6-hour duration, which is supported by the typical rainfall intensity-duration-frequency relationships in the hydrology literature (Koutsoyiannis, Kozonis, & Manetas, 1998).

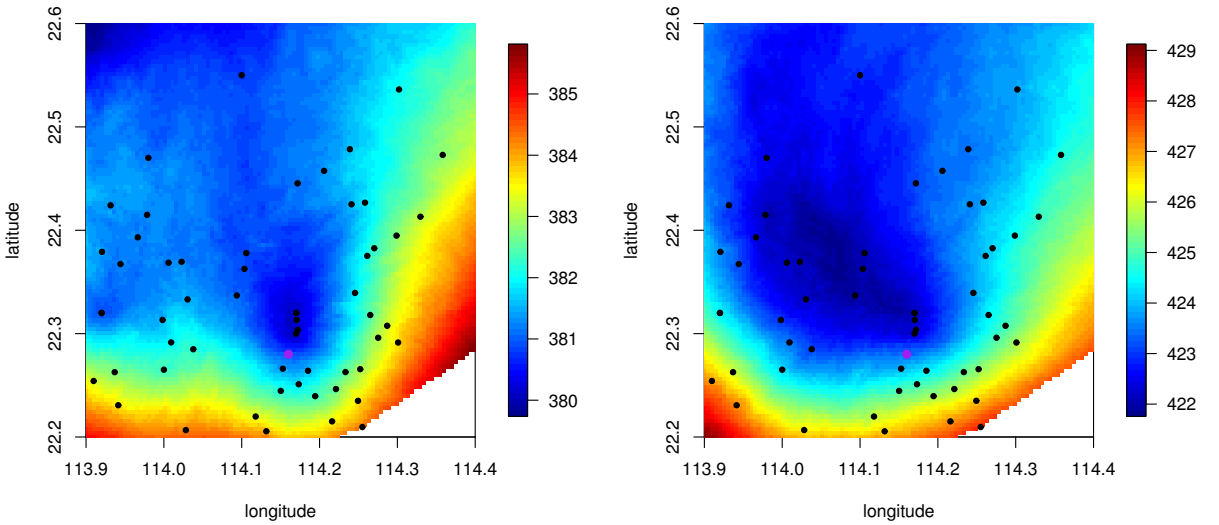
Looking across the three figures, we can also see that the sharpest change of 6-hour return levels occurs at Victoria Harbor, an area transitioning from the Kowloon Peninsula (North) to the Hong Kong Island (South). On the contrary, the sharpest change of 24-hour return levels occurs in northwestern Hong Kong, an area transitioning from the Pearl River Delta Plain (North) to the mountainous New Territories (South). We may thus conjecture that 6-hour precipitation records are more sensitive to differences between sea and land while 24-hour precipitation records are more sensitive to differences between plain and mountain, at least in this region around Hong Kong. The validity of this conjecture remains an interesting question to answer in climatology.



(a) 6-hour mean (unit: mm)

(b) 24-hour mean (unit: mm)

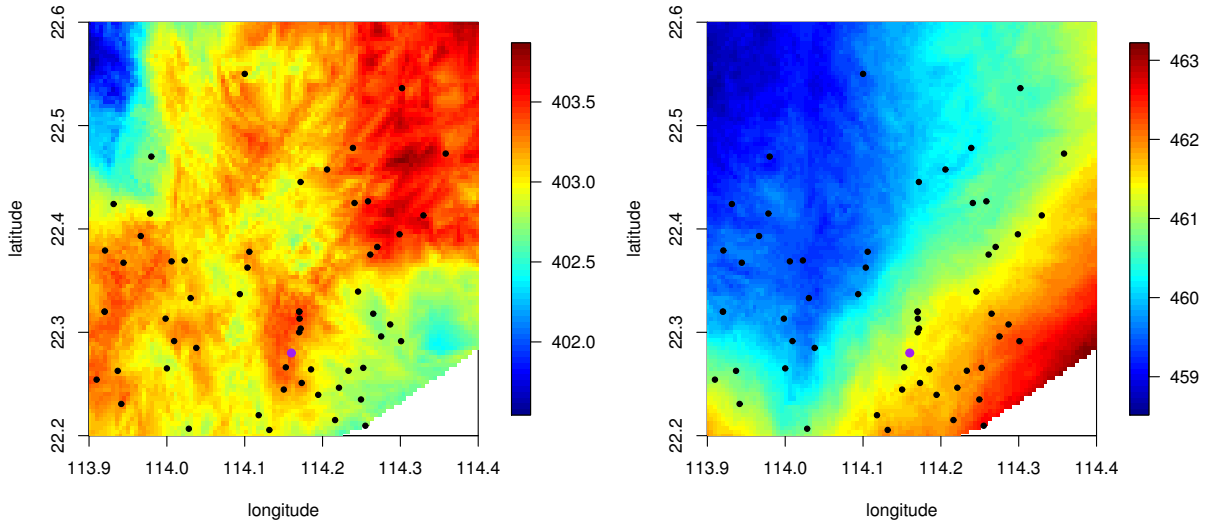
Figure 6: Posterior mean estimates of 25-year 6/24-hour return levels



(a) 6-hour 2.5% quantile (unit: mm)

(b) 6-hour 97.5% quantile (unit: mm)

Figure 7: 95% credible intervals of 25-year 6-hour return levels



(a) 24-hour 2.5% quantile (unit: mm)

(b) 24-hour 97.5% quantile (unit: mm)

Figure 8: 95% credible intervals of 25-year 24-hour return levels

6.6. Posterior predictive checking

As a self-consistency check of whether the observed data is plausible under the posterior predictive distribution, we performed posterior predictive checks on the mean and standard deviation of threshold exceedances as well as marginal predictive checks on each single data points in the data set. Based on the posterior distributions of base parameters, we produced 200 simulated replications of the threshold exceedances data set and looked at the posterior predictive tail area probabilities (Gelman et al., 2013)

$$\int\int I_{T(y^{rep},\omega)\geq T(y,\omega)} p(y^{rep}|\omega)p(\omega|y)dy^{rep}d\omega,$$

where I is the indicator function, y^{rep} is a simulated data set of threshold exceedances, y is the observed data set, ω is the set of base parameters of our model, $T(y^{rep},\omega)$ is in general a function of a data set and the base parameters, most often a statistic, taken to be mean and SD here. Table 8 presents the summary of our posterior predictive checks.

Marginal predictive checks concern tail area probabilities of marginal predictions $Pr(y_i^{rep} > y_i|y)$, where y_i^{rep} is a single data point in a simulated data set. Based on the 200 simulated data sets produced earlier, we computed p-values for each observed data point: The p-values have mean 0.5149 and standard deviation 0.2200. No overwhelmingly significant systematic discrepancies between observed and simulated posterior predictive data set have been detected in terms of mean and standard deviation as well as individual predictions. However, whether there are systematic discrepancies in other aspects of the data is unknown. Moreover, what test statistics are optimal for posterior predictive checks of extreme value distributions remains a very interesting question to pursue, especially when only a very limited number of extreme data points can be observed so that p-values tend to be large.

7. Conclusions and discussion

The proposed model extends the traditional use of two-dimensional locations in the spatial structure to three-dimensional locations. In particular, the third dimension is the weighted duration, whose weight is regressed via Bayesian inference and is incorporated to accommodate datasets with observations that have

Test statistic (T)	Observed $T(y)$	95% credible interval by on $T(y^{rep})$	p-value
Mean	69.51145	[67.50203, 70.67999]	0.265
SD	64.14785	[63.70134, 66.95148]	0.899

Table 8: Summary of posterior predictive checks for means and SDs of threshold exceedances

different durations. This weighted proximity / distance measure helps to assess the influence exerted by independent variables of different natures. Merging duration into the original spatial process also allows better use of limited extreme data and ensures sensible estimates for the return levels obtained for multiple durations.

Despite the rich structure of the proposed model, a tuning algorithm is proposed so that this rich structure will pose fewer challenges in implementation, especially in terms of the slow mixing of MCMC chains. This proposed tuning algorithm is not restricted to this particular scenario but is also applicable to the general use of Metropolis-Hastings algorithms.

The spatial-duration model remains a very flexible framework for further extensions. A straightforward extension is to consider variables other than duration, and hence more scenarios. Models other than the Gaussian process or the exponential covariance function can also be studied. We may also consider a further spatial-temporal extension to model the temporal and spatial variability simultaneously. However, a major challenge here is the steep, almost exponential, increase in computational cost when dealing with rich spatial-temporal frameworks. This also brings up the higher-than-usual variability in the parameters, which potentially makes the tuning of MCMC algorithms a particularly demanding task. Therefore, further development of structures among the parameters would be interesting extensions, among which more compact representations would be most welcome due to their ability to reduce the computational cost. Multiple candidate spatial-temporal structures are interesting to investigate using additive, multiplicative, or composite models, for instance. The inference, however, will probably remain a hard, or become an even harder task.

Acknowledgements

The authors would like to thank the editor, an associate editor, and two referees for their helpful suggestions and comments. This work is supported by a Hong Kong RGC General Research Fund (604313).

References

[1] Ando, T., 2007. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika 94 (2), 443–458.
[2] Ando, T., 2011. Predictive bayesian model selection. American Journal of Mathematical and Management Sciences 31 (1-2), 13–38.
[3] Back, A. J., Uggioni, E., Vieira, H. J., 2011. Modeling precipitation of short duration by means of the modified Bartlett-Lewis rectangular pulse model. Revista Brasileira de Meteorologia 26, 461–472.
[4] Baldassarre, G. D., Castellarin, A., Brath, A., 2006. Relationship between statistics of rainfall extremes and mean annual precipitation: an application for design-storm estimation in northern central Italy. Hydrology and Earth System Sciences 10, 589–601.
[5] Blanchet, J., Davison, A. C., 2011. Spatial modelling of extreme snow depth. Annals of Applied Statistics 5, 1699–1725.
[6] Brown, B., Resnick, S., 1977. Extreme values of independent stochastic processes. Journal of Applied Probability 14, 732–739.
[7] Coles, S., Heffernan, J., Tawn, J., 1999. Dependence measures for extreme value analyses. Extremes 2 (4), 339–365.
[8] Coles, S. G., 1993. Regional modelling of extreme storms via max-stable processes. Journal of the Royal Statistical Society. Series B (Methodological) 55 (4), 797–816.
[9] Coles, S. G., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer London.
[10] Cooley, D., Naveau, P., Poncet, P., 2006. Variograms for spatial max-stable random fields. Vol. 187 of Lecture Notes in Statistics. Springer New York,, pp. 373–390.
[11] Cooley, D., Nychka, D., Naveau, P., 2007. Bayesian spatial modeling of extreme precipitation return levels. Journal of the American Statistical Association 102 (479), 824–840.
[12] Cooley, D., Sain, S. R., 2010. Spatial hierarchical modeling of precipitation extremes from a regional climate model. Journal of Agricultural, Biological, and Environmental Statistics 15 (3), 381–402.
[13] Cox, T. F., Cox, M. A. A., 2001. Multidimensional scaling. Chapman & Hall/CRC.

[14] Curriero, F. C., 2007. On the use of non-Euclidean distance measures in geostatistics. Mathematical Geology 38, 907–926.
[15] Davis, R. A., Klüppelberg, C., Steinkohl, C., 2013a. Max-stable processes for modeling extremes observed in space and time. Journal of the Korean Statistical Society 42, 399–414.
[16] Davis, R. A., Klüppelberg, C., Steinkohl, C., 2013b. Statistical inference for max-stable processes in space and time. Journal of the Royal Statistical Society, Series B 75, 791–819.
[17] Davison, A. C., Padoan, S. A., Ribatet, M., 2012. Statistical modeling of spatial extremes (with discussion). Statistical Science 27, 161–186.
[18] Davison, A. C., Smith, R. L., 1990. Models for exceedances over high thresholds (with discussion). Journal of the Royal Statistical Society, Series B 52, 393–442.
[19] de Haan, L., 1984. A spectral representation of max-stable processes. The Annals of Probability 12, 1194–1204.
[20] de Haan, L., 1985. Extremes in higher dimensions: the model and some statistics. In: Proceedings of the 45th Session of the I.S.I. 26.3.
[21] de Haan, L., Resnick, S. I., 1977. Limit theory for multivariate sample extremes. Z. Wahrscheinlichkeitstheor. Verw. Gebiete 40, 317–337.
[22] Diggle, P. J., Ribeiro, P. J., 2007. Model-based geostatistics. Springer New York.
[23] Erhardt, R. J., Smith, R. L., 2012. Approximate Bayesian computing for spatial extremes. Computational Statistics and Data Analysis 56, 1468–1481.
[24] Erhardt, R. J., Smith, R. L., 2014. Weather derivative risk measures for extreme events. North American Actuarial Journal 18 (3), 379–393.
[25] Feng, S., Nadarajah, S., Hu, Q., 2007. Modeling annual extreme precipitation in China using the generalized extreme value distribution. Journal of the Meteorological Society of Japan 85, 599–613.
[26] Ferreira, A., de Haan, L., 2014. The generalized pareto process: with a view towards application and simulation. Bernoulli 20 (4), 1717–1737.
[27] Fotheringham, A. S., Brunsdon, C., Charlton, M., 2002. Geographically weighted regression: the analysis of spatially varying relationships. Wiley.
[28] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.
[29] Genton, M. G., Ma, Y., Sang, H., 2011. On the likelihood function of Gaussian max-stable processes. Biometrika 98, 481–488.
[30] Ghosh, S., Mallick, B. K., 2011. A hierarchical Bayesian spatio-temporal model for extreme precipitation events. Environmetrics 22 (2), 192–204.
[31] Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive metropolis algorithm. Bernoulli 7 (2), pp. 223–242.
[32] Huser, R., Davison, A. C., 2013a. Composite likelihood estimation for the Brown-Resnick process. Biometrika 100, 511–518.
[33] Huser, R., Davison, A. C., 2013b. Space-time modelling of extreme events. Journal of the Royal Statistical Society, Series B 76, 439–461.
[34] Huser, R., Genton, M. G., 2014. Non-stationary dependence structures for spatial extremes. Working paper.
[35] Kabluchko, Z., Schlather, M., de Haan, L., 2009. Stationary max-stable fields associated to negative definite functions. The Annals of Probability 37, 2042–2065.
[36] Koutsoyiannis, D., Kozonis, D., Manetas, A., 1998. A mathematical framework for studying rainfall intensity-duration-frequency relationships. Journal of Hydrology 206, 118–135.
[37] Kunihama, T., Omori, Y., Zhang, Z., 01 2012. Efficient estimation and particle filter for max-stable processes. Journal of Time Series Analysis 33 (1), 61–80.
[38] Mardia, K. V., Kent, J. T., Bibby, J. M., 1979. Multivariate analysis. Academic Press London.
[39] Mendes, J. M., de Zea Bermudez, P. C., Turkman, K. F., Pereira, J., Vasconcelos, M. J. P., 2010. Spatial extremes of wildfire sizes: Bayesian hierarchical models for extremes. Environmental and Ecological Statistics 17, 1–28.
[40] Nakajima, J., Kunihama, T., Omori, Y., Frűhwirth-Schnatter, S., 2012. Generalized extreme value distribution with time-dependence using the ar and ma models in state space form. Computational Statistics & Data Analysis 56 (11), 3241 – 3259, 1st issue of the Annals of Computational and Financial Econometrics Sixth Special Issue on Computational Econometrics.
[41] Padoan, S. A., Ribatet, M., Sisson, S. A., 2010. Likelihood-based inference for max-stable processes. Journal of the American Statistical Association 105, 263–277.
[42] Pasarica, C., Gelman, A., 2003. Adaptively scaling the metropolis algorithm using expected squared jumped distance. Tech. rep.
[43] Pickands III, J., 1975. Statistical inference using extreme order statistics. The Annals of Statistics 3 (1), 119–131.
[44] Resnick, S. I., 1987. Extreme Values, Regular Variation, and Point Processes. Springer New York.
[45] Resnick, S. I., 2007. Heavy-tail phenomena: Probabilistic and statistical modeling. Springer New York.
[46] Roberts, G., Gelman, A., Gilks, W., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. Annals of Applied Probability 7 (1), 110–120.
[47] Roberts, G. O., Rosenthal, J. S., 2001. Optimal scaling for various metropolis-hastings algorithms. Statistical Science 16 (4), 351–367.
[48] Roberts, G. O., Rosenthal, J. S., 2009. Examples of adaptive mcmc. Journal of Computational and Graphical Statistics 18 (2), 349–367.
[49] Rosenthal, J. S., 2011. Optimal proposal distributions and adaptive MCMC, 1st Edition. Handbooks of Modern Statistical Methods. Chapman & Hall, CRC, Florida, USA.
[50] Sang, H., Gelfand, A. E., 2009. Hierarchical modeling for extreme values observed over space and time. Environmental

and Ecological Statistics 16 (3), 407–426.
[51] Sang, H., Gelfand, A. E., 2010. Continuous spatial process models for spatial extreme values. Journal of Agricultural, Biological, and Environmental Statistics 15 (1), 49–65.
[52] Schabenberger, O., Gotway, C. A., 2005. Statistical methods for spatial data analysis. Chapman & Hall/CRC.
[53] Schlather, M., 2002. Models for stationary max-stable random fields. Extremes 5 (1), 33–44.
[54] Schlather, M., Tawn, J. A., 2003. A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika 90 (1), 139–156.
[55] Smith, R. L., 1989. Extreme value analysis of environmental time series: an application to trend detection in ground level ozone (with discussion). Statistical Science 4, 367–393.
[56] Smith, R. L., 1990. Max-stable processes and spatial extremes.
[57] So, M. K. P., Chan, R. K. S., 2014. Bayesian analysis of tail asymmetry based on a threshold extreme value model. Computational Statistics & Data Analysis 71, 568 – 587.
[58] Sokal, A. D., 1996. Monte carlo methods in statistical mechanics: Foundations and new algorithms.
[59] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4), 583–639.
[60] Tawn, J. A., 1988. Bivariate extreme value theory: Models and estimation. Biometrika 75, 397–415.
[61] Tawn, J. A., 1990. Modeling multivariate extreme value distributions. Biometrika 77, 245–253.
[62] Wadsworth, J. L., Tawn, J. A., 2014. Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. Biometrika 101, 1–15.
[63] Wikle, C. K., Berliner, L. M., Cressie, N., 1998. Hierarchical Bayesian space-time models. Environmental and Ecological Statistics 5 (2), 117–154.