Modeling Spatio-Temporal Extreme Events Using Graphical Models

Yu, Hang; Dauwels, Justin

2016

Yu, H., & Dauwels, J. (2016). Modeling Spatio-Temporal Extreme Events Using Graphical Models. IEEE Transactions on Signal Processing, 64(5), 1101-1116.

https://hdl.handle.net/10356/89365

https://doi.org/10.1109/TSP.2015.2491882

Modeling Spatio-Temporal Extreme Events using Graphical Models

Hang Yu, *Student Member, IEEE,* and Justin Dauwels, *Senior Member, IEEE*

Abstract—**We propose a novel statistical model to describe spatio-temporal extreme events. The model can be used, for instance, to estimate extreme-value temporal pattern such as seasonality and trend, and further to predict the distribution of extreme events in the future. Such model usually involves thousands or even millions of variables in the spatio-temporal domain, whereas only one single observation is available for each location and time point. To address this challenge, previous works usually employ learning and inference methods that are computationally burdensome, and therefore are prohibitive for large-scale data. Moreover, they assume that the shape and scale parameters of the extreme-value distributions are constant across the spatio-temporal domain, which is often too restrictive in practice. In this paper, we break through these limitations by exploring graphical models to capture the highly structured dependencies among the parameters of extreme-value distributions. Furthermore, we develop an efficient stochastic variational inference (SVI) algorithm to learn the parameters of the resulting non-Gaussian graphical model. The computational complexity of the SVI algorithm is sublinear in the number of variables, thus enabling the proposed model to tackle large-scale spatio-temporal data in real-life applications. Results of both synthetic and real data demonstrate the effectiveness of the proposed approach.**

Index Terms—**extreme events, spatio-temporal, graphical models, thin-plate models, variational inference, expectation maximization, stochastic optimization, Kronecker product**

I. Introduction

ANALYSIS of multiple extreme-value time series has found applications and permeated the literature in a wide variety of domains, ranging from finance to climatology. For example, extreme precipitation can characterize climate change [2] and cause flood or flash-flood related hazards [3]. Therefore, assessing the spatial and temporal pattern of such events and making reliable predictions of future trends is crucial for risk management and disaster prevention.

For stationary data, one of the most common approaches for describing their extreme events is the block maximum approach, which models the maxima of a set of contiguous temporal blocks of observations using the Generalized Extreme Value (GEV) distribution [4]. It has been shown that block maxima of stationary time series that are sufficiently separated (e.g., annual maxima) are almost independent [4]. However, increasing the block size to well separate the block maxima often leads to a small sample size, and thus, the resulting estimates of GEV parameters are unreliable. The problem can be alleviated by considering the dependence between multiple GEV distributed variables (e.g., annual maximum precipitation at multiple locations) [5], [6]. As an alternative, instead of using one single sample (i.e., the maximum) in a large block, we can shrink the size of the blocks and introduce temporal dependence to the extreme-value samples so as to utilize the data more effectively. Combination of both approaches leads to GEV distributions whose parameters varying smoothly across both space and time. Apart from stationary time series, nonstationarity in the underlying process such as seasonality, trend, regime changes and dependence on external factors are often the rule rather than the exception. Hence, when modeling block maxima of this process, such covariate effects also need to be taken into account, and this again results in a spatio-temporal model. Unfortunately, under both settings, only one sample is available at each measuring site and time point, whereas the corresponding GEV distribution has three parameters to be estimated.

Due to the abovementioned challenge, there is only a handful of spatio-temporal models for extreme events at present. In the following, we review the literature on spatio-temporal extreme events. A clear account of temporal dependence is presented in [4], which defines the temporal changes of GEV parameters through deterministic functions, such as linear, log-linear and quadratic functions. However, the restrictive functional forms pose a serious limitation in practice. A more satisfactory approach is to replace the deterministic function with a linear combination of suitable basis functions, such as splines, and add a penalty to guard against overfitting [7]. The smoothness parameters (i.e., penalty parameters) are chosen through cross validation or Akaike Information Criterion. The tuning process is usually computationally burdensome since numerous candidate values of the smoothness parameters have to be tested before the proper amount of smoothness is determined. Moreover, the computational complexity increases exponentially with the number of smoothness parameters. To overcome this deficiency, a Bayesian approach is proposed in [8], where the smoothness parameters are regarded as random variables and Gamma priors are imposed on them. Such Bayesian models are often inferred by the Monte Carlo Markov chain (MCMC) algorithm, and the algorithm can be unacceptably slow for large-scale problems. On the other hand, Neville *et al.* [9] apply the mean field variational Bayes

Part of the material in this paper was presented at ISIT, Honolulu, HI, July 2014 [1].

H. Yu is with School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798 Singapore (e-mail: HYU1@e.ntu.edu.sg).

J. Dauwels is with School of Electrical and Electronic Engineering and School of Physical and Mathematical Sciences, Nanyang Technological University, 50 Nanyang Avenue, 639798 Singapore (e-mail: JDAUWELS@ntu.edu.sg).

method to learn the model. Similarly as in [10], the GEV distributions are approximated by Gaussian mixtures for a fixed set of shape parameters, in order to avoid the complex functional form of the GEV distributions. Closed-form update rules for the parameters of the variational distributions are derived. As a consequence, the algorithm needs to be run once for each possible value of the shape parameter in the predefined set (which may be large) before the one associated with the largest likelihood is selected.

Another line of research investigates the application of dynamic linear models (DLM) to extreme-value time series, cf. [11]-[15]. DLMs relate the present GEV parameters to the historical estimates while embedding the spatial dependence implicitly in the evolution matrix. Such models are often estimated via MCMC methods [11]-[13], [15] or generalized expectation maximization [14]. Unfortunately, the computational cost of learning a DLM is $\mathcal{O}(NP^3)$, where P is the number of measuring stations and N is the number of block maxima observed at each station (i.e., length of time series). As a result, such models are prohibitive for the cases where P is large. Furthermore, DLMs use directed acyclic graphs to represent the dependence from past to present, and hence, the estimates of GEV parameters at time t only depend on extreme-value samples up to t. Obviously, it is more tempting to take full advantage of all the observed samples, both before and after t, to yield more reliable estimates of GEV parameters. This indicates that GEV parameters at different time points depend on each other, thus constituting an undirected cyclic graph.

Apart from the above mentioned issues, the previous works [7]-[14] often restrict the shape and scale parameters of the GEV distributions to be constant across the spatio-temporal domain, and only allow the location parameters to vary so as to capture the non-stationarity in the data. This assumption stems from the analysis on spatial extremes in [16], which shows that varying the shape parameter across the space only slightly improves the model fitting, but the resulting score of the deviance information criterion (DIC) [17] is larger than that with a constant shape parameter. Moreover, it also simplifies the corresponding learning algorithms. However, as pointed out in [15], treating the shape and scale parameter as a constant is inappropriate for modeling monthly maxima. This point is also demonstrated by our numerical results on both synthetic and real data.

In this paper, we propose to exploit undirected graphical models (i.e., Markov random fields) [18] to capture the highly structured spatial and temporal dependencies among GEV parameters. We aim to estimate the temporal pattern of extreme events, such as the trend or seasonality of the data in time. Furthermore, we intend to predict the distribution of extreme events in the future based on the current trend. In the example of extreme rainfall, forecasting whether the size of extremes will increase in the future is the key for flood warning and strategic planning.

To move forward to this goal, we first assume that the single block maximum at each site and time position (i.e., each month) follows a GEV distribution. We further stipulate that each of the three GEV parameters (the shape, scale, and location parameter) corresponding one site and time point can be decomposed into the sum of two components: a spatial and a temporal component. The proposed model is therefore similar in spirit to the generalized additive models that are popular in the literature of extreme events modeling [7], [9], [19]. Next, we impose Gaussian graphical model priors, particularly, thin-plate model priors, respectively on the spatial components across space and the time components across time. The amount of dependence is then determined by the smoothness parameters of the thin-plate models. In order to infer all the parameters, we follow the empirical Bayes approach; we generate point estimates of the smoothness parameters while inferring the posterior distribution of the GEV parameters. Specifically, we approximate the posterior distribution of the GEV parameters by a multivariate Gaussian distribution with a diagonal covariance matrix, and exploit efficient stochastic optimization methods [20], [21] to learn both the smoothness parameters and the parameters of the variational distribution. As only noisy gradients (rather than the exact ones) are required in each iteration of the stochastic variational inference algorithm, the computational complexity can be reduced to be sublinear in the number of variables. Numerical results show that the proposed model can automatically recover the underlying pattern of GEV parameters across both space and time, given one single sample observed at each location and time point. Moreover, it also provides an effective tool to predict the future distribution of extreme events.

The rest of the paper is organized as follows. In Section II, we review thin-plate models, since those models play a central role in our approach. In Section III, we present the proposed spatio-temporal model for extreme events in detail. The efficient stochastic variational inference algorithm is then developed in Section IV. We also explain how to predict the distribution of extreme events in the future in this section. In Section V, we assess the proposed model and benchmark it with other models by means of synthetic and real data. We conclude in Section VI with a discussion and an outlook.

II. Thin-Plate Models

In this section, we first give a short description of graphical models, and then we discuss the special case of thin-plate models. In particular, we introduce thin-plate models with zero curvature and zero gradient boundary conditions respectively. The former is useful to predict future trend while the latter can deal with multi-dimensional data. As a result, we utilize the former to capture temporal dependence among GEV parameters and the latter for modeling spatial dependence.

In an undirected graphical model (i.e., a Markov random field), the probability distribution is represented by an undirected graph $\mathcal{G}$ which consists of P nodes $\mathcal{V}$ and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Each node i is associated with a random variable z_i. An edge (i,j) is absent if the corresponding two variables z_i and z_j are conditionally independent: $P(z_i, z_j|\boldsymbol{z}_{\mathcal{V}|i,j}) = P(z_i|\boldsymbol{z}_{\mathcal{V}|i,j})P(z_j|\boldsymbol{z}_{\mathcal{V}|i,j})$, where $\mathcal{V}|i,j$ denotes all the nodes in the set $\mathcal{V}$ except i and j.

In particular, for Gaussian distributed $\boldsymbol{z} = [z_i]^T$, the resulting graphical model is called a Gaussian graphical model or a Gauss-Markov random field (GMRF). Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{m}, \Sigma)$ with mean vector $\boldsymbol{m}$ and positive-definite covariance matrix Σ. The Gaussian graphical model can be equivalently parameterized as $\mathcal{N}(K^{-1}h, K^{-1})$ with a precision matrix $K = \Sigma^{-1}$ and a potential vector $\boldsymbol{h} = K\boldsymbol{m}$. The resulting PDF can be expressed as:

$$p(\boldsymbol{z}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{z}^T K\boldsymbol{z} + \boldsymbol{h}^T\boldsymbol{z}\right\}. \tag{1}$$

Interestingly, the graph $\mathcal{G}$ is characterized by the precision matrix (the inverse covariance) K, i.e., $K_{ij} \neq 0$ if and only if the edge $(i, j) \in \mathcal{E}$ [22].

The thin-plate model [23], [24] is a GMRF that is commonly used as smoothness prior as it penalizes the second-order difference. In other words, we model the second-order differences as a Gaussian distribution:

$$\Delta^2 z_i \sim \mathcal{N}(0, \alpha_z^{-1}). \tag{2}$$

For a one-dimensional problem where the variables z_i's are evenly located on a chain, the second-order difference at z_i can be defined as $\Delta^2 z_i = z_{i-1} - 2z_i + z_{i+1}$. As a result, the density function of a thin-plate model with a chain structure can be written as [23]:

$$p(\boldsymbol{z}) \propto \exp\left\{-\frac{\alpha_z}{2}\sum_{i=2}^{P-1}(z_{i-1} - 2z_i + z_{i+1})^2\right\} \tag{3}$$

$$= \exp\left\{-\frac{\alpha_z}{2}\boldsymbol{z}^T K_{\text{tp}}\boldsymbol{z}\right\}, \tag{4}$$

where the smoothness parameter α_z controls the curvature, and K_{tp} has the following form:

$$K_{\text{tp}} = A^T A, \tag{5}$$

$$A = \begin{bmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{bmatrix}. \tag{6}$$

Note that A is a $P-2\times P$ matrix. Apparently, the precision of a thin-plate model $K = \alpha_z K_{\text{tp}}$. It is easy to tell from (3) that the thin-plate model is invariant to the addition of a constant, and more importantly, a linear function along the Markov chain. In other words, $K_{\text{tp}}\boldsymbol{1} = \boldsymbol{0}$ and $K_{\text{tp}}\boldsymbol{s}_1 = \boldsymbol{0}$, where $\boldsymbol{1}$ is a column vector of all ones, and $\boldsymbol{s}_1 = [1, 2, ..., P]^T$. As such, this prior can accommodate the linear trends without penalty. We can also conclude that K_{tp} is rank deficient with two zero eigenvalues. As a result, the improper density is often used in practice [23], [25], that is,

$$p(\boldsymbol{z}) \propto |K|_+^{0.5}\exp\left\{-\frac{1}{2}\boldsymbol{z}^T K\boldsymbol{z}\right\}, \tag{7}$$

where $|K|_+$ denotes the product of the positive eigenvalues of the precision matrix K. We can also read from (5) that the conditional mean of one variable z_i conditioned on other variables $\boldsymbol{z}_{\mathcal{V}|i}$ is [23]:

$$E(z_i|\boldsymbol{z}_{\mathcal{V}|i}) = \frac{4}{6}(z_{i+1} + z_{i-1}) - \frac{1}{6}(z_{i+2} + z_{i-2}), \tag{8}$$

which can be regarded as second-order polynomial interpolation based on four nearby variables z_{i-2}, z_{i-1}, z_{i+1}, and z_{i+2} without an overall level. Therefore, the thin-plate model allows the deviation from any overall mean level without having to specify the overall mean level itself. Such property is often favored in practice. Furthermore, the zero curvature boundary condition of (3) (i.e., $\Delta^2 z_1 = \Delta^2 z_P = 0$) aids in predicting (or extrapolating) future values [23], i.e.,

$$E(z_{P+1}|z_1, \cdots, z_P; \alpha_z) = z_P + (z_P - z_{P-1}). \tag{9}$$

Therefore, the conditional mean of z_{P+1} is simply the linear extrapolation based on the last two observations z_{P-1} and z_P. Such models will be exploited to model temporal trend of GEV parameters.

A zero gradient boundary condition is also often applied in thin-plat models, i.e., $z_0 = z_1$ and $z_{P+1} = z_P$. Consequently, we can simplify the second-order difference at the boundary variable z_1 and z_P respectively as:

$$\Delta^2 z_1 = z_0 - 2z_1 + z_2 = z_2 - z_1, \tag{10}$$

$$\Delta^2 z_P = z_{P-1} - 2z_P + z_{P+1} = z_{P-1} - z_P. \tag{11}$$

Hence, the resulting thin-plate model with constant boundary condition is [24]:

$$p(\boldsymbol{z}) \propto \exp\left\{-\frac{\alpha}{2}\sum_{i=1}^{P}(|N(i)|z_i - \sum_{j\in N(i)} z_j)^2\right\}, \tag{12}$$

where $N(i)$ denotes the neighboring nodes of z_i and $|N(i)|$ is the number of neighbors z_i have. In (12), each node is modeled to be close to the average of its neighbors. Note that the resulting K_{tp} has rank $P - 1$. This model can be easily extended to the case of multiple dimensions, and coincides with the boundary conditions proposed in [26] to address the problem of extending (3) to a two-dimensional spatial domain. Moreover, it is shown in [27] that such models perform better when modeling spatial dependence of spatial data than thin-membrane models [28] that penalize gradient. As a result, we will employ this type of thin-plate models to capture spatial dependence among GEV parameters in the sequel.

III. Spatio-Temporal Models for Extreme Events

In this section, we present the proposed spatio-temporal graphical model for extreme events. Suppose that we have N block maxima x_{ij} at each of the P locations, where $i = 1, \cdots, N$ and $j = 1, \cdots, P$. The resulting number of dimensions of the spatio-temporal model is $D = NP$. We further assume the observations are missing at random, in order to demonstrate that the proposed model is capable of dealing with missing data. The set of observed spatio-temporal indices is denoted as $\mathcal{V}_O$.

A. Likelihood: Generalized Extreme Value Distributions

Motivated by the extreme value theory, we assume that each observed x_{ij} follows a Generalized Extreme Value (GEV) distribution with cumulative distribution function (CDF) [4]:

$$F(x_{ij}|\xi_{ij}, \sigma_{ij}, \mu_{ij}) = \exp\left\{-\left[1 + \xi_{ij}\left(\frac{x_{ij} - \mu_{ij}}{\sigma_{ij}}\right)\right]^{-\frac{1}{\xi_{ij}}}\right\}, \tag{13}$$

where $\mu_{ij} \in \mathbb{R}$ is the location parameter, $\sigma_{ij} > 0$ is the scale parameter and $\xi_{ij} \neq 0$ is the shape parameter. We further use

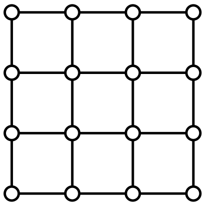


Fig. 1: Neighborhood structure of the graphical model for spatial components.

$\zeta_{ij} = \log \sigma_{ij}$ to parameterize the GEV distribution such that $\zeta_{ij} \in \mathbb{R}$ and a Gaussian prior can be imposed on ζ_{ij}. Taking derivatives of (13) with regard to x_{ij} yields the probability density function (PDF):

$$f(x_{ij}|\xi_{ij}, \zeta_{ij}, \mu_{ij}) = \frac{1}{\exp(\zeta_{ij})} \left[1 + \xi_{ij} \left(\frac{x_{ij} - \mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}} - 1} \cdot \exp\left\{-\left[1 + \xi_{ij} \left(\frac{x_{ij} - \mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}}}\right\}. \tag{14}$$

Note that the support of both (13) and (14) is $\{x_{ij} : 1 + \xi_{ij}(x_{ij} - \mu_{ij})/\exp(\zeta_{ij}) > 0\}$. For $\xi_{ij} = 0$, the above functions are undefined, and thus are replaced by the results obtained by taking the limit as $\xi_{ij} \to 0$. In this case, the resulting CDF and PDF are:

$$F(x_{ij}|0, \zeta_{ij}, \mu_{ij}) = \exp\left[-\exp\left(-\frac{x_{ij} - \mu_{ij}}{\exp(\zeta_{ij})}\right)\right], \tag{15}$$

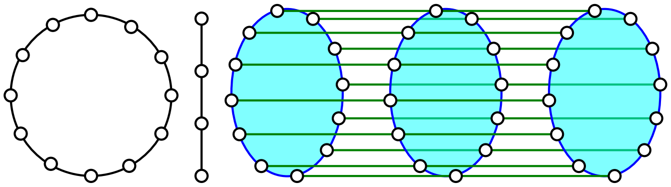
$$f(x_{ij}|0, \zeta_{ij}, \mu_{ij}) = \frac{1}{\exp(\zeta_{ij})} \exp\left(-\frac{x_{ij} - \mu_{ij}}{\exp(\zeta_{ij})}\right) \cdot \exp\left\{-\exp\left(-\frac{x_{ij} - \mu_{ij}}{\exp(\zeta_{ij})}\right)\right\}, \tag{16}$$

where x_{ij} is defined in $\mathbb{R}$.

Note that the key parameter of GEV distributions is the shape parameter, which determines the subfamily. Specifically, $\xi_{ij} = 0$ yields Gumbel distributions with light upper tails, $\xi_{ij} > 0$ corresponds to Fréchet distributions with heavy upper tails, while $\xi_{ij} < 0$ corresponds to Weibull distributions with bounded upper tails.

B. Prior: Thin-plate Models

We now turn our attention to the prior distributions of the GEV parameters. Since the three parameters share the same dependence structure, we present them in a unified form. Let z_{ij} denote either ξ_{ij}, ζ_{ij} or μ_{ij}. We assume that each parameter z_{ij} at time instant i and site j can be decomposed as $z_{ij} = z_{Ti} + z_{Sj}$, where z_{Ti} is the temporal component at time instant i and z_{Sj} is the spatial component at site j. Note that we construct the model in a similar fashion to the generalized additive models that have seen broad applications in the literature of extreme value analysis due to their simplicity, flexibility and utility, cf. [7], [9], [19] and references therein. However, different from the generalized additive models that are often decomposed as the deterministic (spline) functions for the spatial and the temporal component plus a noise term, we only put smoothness priors on the spatial and temporal component without specifying their functional form. Thus, the resulting model can be more flexible. In addition, we include the noise term implicity in $\boldsymbol{z}_S$ and $\boldsymbol{z}_T$.



(a) (b) (c)

Fig. 2: Neighborhood structure of the graphical model for temporal components: (a) cycle graph; (b) chain graph; (c) neighborhood structure for temporal dependence.

In the following, we describe the priors on $\boldsymbol{z}_S = [z_{S1}, \cdots, z_{SP}]^T$ and $\boldsymbol{z}_T = [z_{T1}, \cdots, z_{TN}]^T$ individually. Such priors are constructed according to the highly structured dependence between the GEV parameters.

Without loss of generality, we assume that the measuring stations are deployed on a regular lattice as shown in Fig. 1. As a result, we employ the thin-plate model with zero gradient boundary condition (12) to capture the spatial dependence among the sites (variables) in the spatial component $\boldsymbol{z}_S$:

$$p(\boldsymbol{z}_S|\alpha_z) \propto \exp\left\{-\frac{\alpha_z}{2} \sum_{i=1}^{P} (|N(i)| z_{Si} - \sum_{j \in N(i)} z_{Sj})^2\right\} \propto |\alpha_z K_S|_+^{0.5} \exp\left\{-\frac{\alpha_z}{2} \boldsymbol{z}_S^T K_S \boldsymbol{z}_S\right\}. \tag{17}$$

Here, $N(i)$ typically includes four neighbors (two vertical and two horizontal) of node i.

Next, we consider the temporal dependence. We first deal with the periodicity and the trend separately, and then integrate them together to construct the temporal graphical model. More specifically, we partition $\boldsymbol{z}_T$ according to the period as $(z_{T1}, z_{T2}, \cdots, z_{T\tau}), (z_{T\tau+1}, z_{T\tau+2}, \cdots, z_{T2\tau}), \cdots$, where τ is the period. Since we focus on block maxima, τ is automatically determined by the block size. For example, if we analyze monthly maxima, $\tau = 12$, whereas for seasonal maxima, $\tau = 4$. For variables in each group, e.g., $(z_{T1}, z_{T2}, \cdots, z_{T\tau})$, we couple them together via a cycle graph as shown in Fig. 2a to accommodate the periodicity in the time series. In this case, the thin-plate model with zero curvature and zero gradient boundary condition takes on the same form; we use $\beta_z K_{pr}$ to denote the precision matrix of the thin-plate model, where β_z is the smoothness parameter. On the other hand, we capture the trend by coupling $z_{Ti}, z_{Ti+\tau}, z_{Ti+2\tau}, \cdots$ together via a chain graph (see Fig. 2b), for $i = 1, \cdots, \tau$. Here, we utilize the thin-plate model with zero curvature boundary condition for the sake of future forecast, and represent the corresponding precision matrix as $\gamma_z K_{tr}$. As a consequence, the neighborhood structure of the temporal model can be specified as in Fig. 2c, and the overall precision matrix is given by:

$$K_T = (\gamma_z K_{tr}) \oplus (\beta_z K_{pr}) \tag{18}$$

$$= \gamma_z K_{tr} \otimes I_{pr} + \beta_z I_{tr} \otimes K_{pr} \tag{19}$$

$$= \gamma_z \tilde{K}_{tr} + \beta_z \tilde{K}_{pr}, \tag{20}$$

where $\oplus$ and $\otimes$ denote the Kronecker sum and Kronecker product respectively, I_* is an identity matrix with the same dimension as K_*, $\tilde{K}_{pr} = I_{pr} \otimes K_{pr}$ characterizes the dependence within each period (corresponding to the blue edges in Fig. 2c), and $\tilde{K}_{tr} = K_{tr} \otimes I_{tr}$ characterizes the dependence between contiguous periods (corresponding to the green edges in Fig. 2c).

As mentioned in Section II, thin-plate models do not specify the overall mean level for $\boldsymbol{z}_S$ and $\boldsymbol{z}_T$, and thus, there can be an infinite number of combinations (z_{Ti}, z_{Sj}) with $z_{ij} = z_{Ti} + z_{Sj}$ unchanged. To remedy the problem, we explicitly add a constraint in the prior of $\boldsymbol{z}_T$ that $\sum_i z_{Ti} = 0$. Taken together, the prior density of $\boldsymbol{z}_T$ can be written as:

$$\begin{aligned} p(\boldsymbol{z}_T|\beta_z, \gamma_z) &\propto \left(N|\gamma_z \tilde{K}_{tr} + \beta_z \tilde{K}_{pr}|_+\right)^{0.5} \\ &\cdot \exp\left\{-\frac{1}{2}\boldsymbol{z}_T^T \left(\gamma_z \tilde{K}_{tr} + \beta_z \tilde{K}_{pr} + \mathbf{1}\mathbf{1}^T\right)\boldsymbol{z}_T\right\}, \end{aligned} \tag{21}$$

where $\mathbf{1}$ is a column vector of all ones. Recall that the eigenvalue of K_T corresponding to eigenvector $\mathbf{1}$ is 0 (see Section II). By adding $\mathbf{1}\mathbf{1}^T$ to K_T, we only modify the eigenvalue to N with other eigenvalues unchanged. Therefore, $|\gamma_z \tilde{K}_{tr} + \beta_z \tilde{K}_{pr} + \mathbf{1}\mathbf{1}^T|_+ = N|\gamma_z \tilde{K}_{tr} + \beta_z \tilde{K}_{pr}|_+$.

C. Spatio-Temporal Graphical Models for Extreme Values

The joint PDF of the overall spatio-temporal model can be written as:

$$\begin{aligned} &p(\boldsymbol{x}, \boldsymbol{\xi}_S, \boldsymbol{\xi}_T, \boldsymbol{\zeta}_S, \boldsymbol{\zeta}_T, \boldsymbol{\mu}_S, \boldsymbol{\mu}_T|\alpha_\xi, \beta_\xi, \gamma_\xi, \alpha_\zeta, \beta_\zeta, \gamma_\zeta, \alpha_\mu, \beta_\mu, \gamma_\mu) \\ &= \prod_{\{i,j\}\in\mathcal{V}_O} f(x_{ij}|\xi_{Ti} + \xi_{Sj}, \zeta_{Ti} + \zeta_{Sj}, \mu_{Ti} + \mu_{Sj}) \\ &\cdot \prod_{z\in\{\xi,\zeta,\mu\}} p(\boldsymbol{z}_S|\alpha_z)p(\boldsymbol{z}_T|\beta_z, \gamma_z), \end{aligned} \tag{22}$$

where $f(x_{ij}|\xi_{Ti}+\xi_{Sj}, \zeta_{Ti}+\zeta_{Sj}, \mu_{Ti}+\mu_{Sj})$ is the GEV density function (i.e., the likelihood of the GEV parameters, cf. (14) and (16)) introduced in Subsection III-A.

Since the GEV densities are non-Gaussian, the resulting overall graphical model is non-Gaussian as well. Therefore, we exploit variational inference methods to estimate both the GEV parameters and the smoothness parameters given observed extreme values $\boldsymbol{x} = [x_{ij}]^T$, which is explained in the next section.

IV. Learning and Inference

In this section, we first elaborate on how to learn both the GEV and smoothness parameters given the extreme-value observations. We then employ the model to predict future GEV distributions.

A. Learning GEV and Smoothness Parameters

As mentioned in Section I, we estimate all the parameters through an empirical Bayes approach [29]. Specifically, we infer the smoothness parameters by maximum likelihood estimation. Let $\boldsymbol{y} = [\boldsymbol{\xi}_S; \boldsymbol{\xi}_T; \boldsymbol{\zeta}_S; \boldsymbol{\zeta}_T; \boldsymbol{\mu}_S; \boldsymbol{\mu}_T]$ and $\boldsymbol{\theta} = [\alpha_\xi; \beta_\xi; \gamma_\xi; \alpha_\zeta; \beta_\zeta; \gamma_\zeta; \alpha_\mu; \beta_\mu; \gamma_\mu]$. The likelihood is given by:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \int_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y} = \int_{\boldsymbol{y}} p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}, \tag{23}$$

where $p(\boldsymbol{x}|\boldsymbol{y})$ are the GEV densities, and $p(\boldsymbol{y}|\boldsymbol{\theta})$ are the thin-plate model priors. Since maximizing $p(\boldsymbol{x}|\boldsymbol{\theta})$ (23) directly is intractable, we instead find $q(\boldsymbol{y})$ and $\boldsymbol{\theta}$ to maximize the lower bound of $\log p(\boldsymbol{x}|\boldsymbol{\theta})$:

$$L = \int_{\boldsymbol{y}} q(\boldsymbol{y}) \log \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{\theta})}{q(\boldsymbol{y})}d\boldsymbol{y} \leq \log p(\boldsymbol{x}|\boldsymbol{\theta}). \tag{24}$$

Ideally, we choose $q(\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ such that the lower bound L is maximized. However, in our case, we cannot obtain the closed-form expression of the posterior distribution $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ due to the complicated functional form of the GEV densities. Alternatively, we resort to the variational inference algorithm [30], in which we find the variational distribution $q(\boldsymbol{y})$ with a fixed but tractable functional form that maximizes L. More precisely, we set $q(\boldsymbol{y})$ to be a multivariate Gaussian distribution with a diagonal covariance matrix. Specifying $q(\boldsymbol{y})$ to such a simple function can dramatically speed up the learning process. At the same time, each factor $q(y_i)$ can reliably approximate the corresponding marginal posterior distribution $p(y_i|\boldsymbol{x}, \boldsymbol{\theta})$ given by Gibbs sampling, especially the mean, as shown in our numerical experiments.

We estimate both the smoothness parameters and the parameters of $q(\boldsymbol{y})$ via stochastic optimization [32], which is discussed below at length. Since the variational distribution $q(\boldsymbol{y})$ is given by $\mathcal{N}(\boldsymbol{y}; \boldsymbol{m}, CC^T)$, where $\boldsymbol{m}$ is the mean vector, and C is a diagonal matrix with the standard deviation vector $\boldsymbol{\nu}$ on the diagonal, $\boldsymbol{y}$ can be equivalently parameterized as [20]:

$$\boldsymbol{y} = C\boldsymbol{y}_e + \boldsymbol{m}, \tag{25}$$

$$\boldsymbol{y}_e \sim \phi(\boldsymbol{y}_e), \tag{26}$$

where $\phi(\boldsymbol{y}_e) = \mathcal{N}(\boldsymbol{y}_e; \mathbf{0}, I)$ is a multivariate Gaussian distribution with zero mean and unit variance. By changing variables according to $\boldsymbol{y}_e = C^{-1}(\boldsymbol{y} - \boldsymbol{m})$, L can be expressed as [20]:

$$\begin{aligned} L = &\int_{\boldsymbol{y}_e} \phi(\boldsymbol{y}_e) \log \frac{p(\boldsymbol{x}|C\boldsymbol{y}_e + \boldsymbol{m})p(C\boldsymbol{y}_e + \boldsymbol{m}|\boldsymbol{\theta})|C|}{\phi(\boldsymbol{y}_e)}d\boldsymbol{y}_e \\ =&E_{\phi(\boldsymbol{y}_e)}\left[\log p(\boldsymbol{x}|C\boldsymbol{y}_e + \boldsymbol{m})\right] + E_{\phi(\boldsymbol{y}_e)}\left[\log p(C\boldsymbol{y}_e + \boldsymbol{m}|\boldsymbol{\theta})\right] \\ &+ \log|C| + c, \end{aligned} \tag{27}$$

where c is a constant that summarizes all irrelevant terms. Since $p(C\boldsymbol{y}_e + \boldsymbol{m}|\boldsymbol{\theta})$ corresponds to a Gaussian graphical model, we can obtain the closed-form expression of the second term in (27):

$$\begin{aligned} &E_{\phi(\boldsymbol{y}_e)}\left[\log p(C\boldsymbol{y}_e + \boldsymbol{m}|\boldsymbol{\theta})\right] \\ &= \sum_{z\in\{\xi,\zeta,\mu\}} E_{\phi(\boldsymbol{z}_{Se})}\left[\log p(\boldsymbol{z}_S|\alpha_z)\right] + E_{\phi(\boldsymbol{z}_{Te})}\left[\log p(\boldsymbol{z}_T|\beta_z, \gamma_z)\right], \end{aligned} \tag{28}$$

and

$$E_{\phi(\boldsymbol{z}_{Se})}\left[\log p(\boldsymbol{z}_S|\alpha_z)\right] = \frac{1}{2}\log|\alpha_z K_S|_+$$

$$-\frac{\alpha_z}{2}\boldsymbol{m}_{\boldsymbol{z}_S}^T K_S \boldsymbol{m}_{\boldsymbol{z}_S} - \frac{\alpha_z}{2}\boldsymbol{\nu}_{\boldsymbol{z}_S}^T \mathrm{diag}(K_S)\boldsymbol{\nu}_{\boldsymbol{z}_S}, \tag{29}$$

$$E_{\phi(\boldsymbol{z}_{Te})}\left[\log p(\boldsymbol{z}_T|\beta_z,\gamma_z)\right] = \frac{1}{2}\Big(\log|\beta_z\tilde{K}_{pr}+\gamma_z\tilde{K}_{tr}|_+ + \log N\Big) - \frac{1}{2}\boldsymbol{m}_{\boldsymbol{z}_T}^T\left(K_T+\boldsymbol{1}\boldsymbol{1}^T\right)\boldsymbol{m}_{\boldsymbol{z}_T} - \frac{1}{2}\boldsymbol{\nu}_{\boldsymbol{z}_T}^T\mathrm{diag}\left(K_T+\boldsymbol{1}\boldsymbol{1}^T\right)\boldsymbol{\nu}_{\boldsymbol{z}_T}, \tag{30}$$

where $\mathrm{diag}(K)$ is a diagonal matrix whose diagonal equals that of K.

Our objective is to find the smoothness parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{m}$ and $\boldsymbol{\nu}$ to maximize the lower bound L (27). To this end, we consider the gradients with respect to those parameters. For the spatial smoothness parameter α_z, the gradient is given by:

$$\frac{\partial L}{\partial \alpha_z} = \frac{P-1}{2\alpha_z} - \frac{1}{2}\boldsymbol{m}_{\boldsymbol{z}_S}^T K_S \boldsymbol{m}_{\boldsymbol{z}_S} - \frac{1}{2}\boldsymbol{\nu}_{\boldsymbol{z}_S}^T\mathrm{diag}(K_S)\boldsymbol{\nu}_{\boldsymbol{z}_S}. \tag{31}$$

For the smoothness parameters β_z and γ_z that characterize the temporal dependence, the gradient appears to be complicated because of the log-determinant term. However, recall that $\beta_z\tilde{K}_{pr}+\gamma_z\tilde{K}_{tr} = K_T = (\gamma_z K_{tr}) \oplus (\beta_z K_{pr})$ (18). Due to the properties of Kronecker sum [31, Ch. 13], the eigenvalue matrix Λ_T of K_T boils down to:

$$\Lambda_T = (\gamma_z\Lambda_{tr}) \oplus (\beta_z\Lambda_{pr}) = \beta_z\tilde{\Lambda}_{pr} + \gamma_z\tilde{\Lambda}_{tr}, \tag{32}$$

where Λ_{tr} and Λ_{pr} are the eigenvalue matrices corresponding respectively to K_{tr} and K_{pr}, $\tilde{\Lambda}_{pr} = I_{tr}\otimes\Lambda_{pr}$, and $\tilde{\Lambda}_{tr} = \Lambda_{tr}\otimes I_{pr}$. Consequently, the log-determinant term in (30) can be simplified as:

$$\log|\beta_z\tilde{K}_{pr}+\gamma_z\tilde{K}_{tr}|_+ = \log|\beta_z\tilde{\Lambda}_{pr}+\gamma_z\tilde{\Lambda}_{tr}|_+ \tag{33}$$

$$= \sum_{\{i:\tilde{\lambda}_{pri}+\tilde{\lambda}_{tri}>0\}} \log(\beta_z\tilde{\lambda}_{pri}+\gamma_z\tilde{\lambda}_{tri}), \tag{34}$$

and therefore the gradient of L with regard to β_z and γ_z equals:

$$\frac{\partial L}{\partial \beta_z} = \frac{1}{2}\sum_{\{i:\tilde{\lambda}_{pri}+\tilde{\lambda}_{tri}>0\}} \frac{\tilde{\lambda}_{pri}}{\beta_z\tilde{\lambda}_{pri}+\gamma_z\tilde{\lambda}_{tri}} - \frac{1}{2}\boldsymbol{m}_{\boldsymbol{z}_T}^T\tilde{K}_{pr}\boldsymbol{m}_{\boldsymbol{z}_T} - \frac{1}{2}\boldsymbol{\nu}_{\boldsymbol{z}_T}^T\mathrm{diag}(\tilde{K}_{pr})\boldsymbol{\nu}_{\boldsymbol{z}_T}, \tag{35}$$

$$\frac{\partial L}{\partial \gamma_z} = \frac{1}{2}\sum_{\{i:\tilde{\lambda}_{pri}+\tilde{\lambda}_{tri}>0\}} \frac{\tilde{\lambda}_{tri}}{\beta_z\tilde{\lambda}_{pri}+\gamma_z\tilde{\lambda}_{tri}} - \frac{1}{2}\boldsymbol{m}_{\boldsymbol{z}_T}^T\tilde{K}_{tr}\boldsymbol{m}_{\boldsymbol{z}_T} - \frac{1}{2}\boldsymbol{\nu}_{\boldsymbol{z}_T}^T\mathrm{diag}(\tilde{K}_{tr})\boldsymbol{\nu}_{\boldsymbol{z}_T}. \tag{36}$$

On the other hand, the gradient $w.r.t$ the variational mean and standard deviation corresponding to the spatial components $\boldsymbol{z}_S$ of the GEV parameters can be computed as:

$$\nabla_{\boldsymbol{m}_{\boldsymbol{z}_S}} L = E_{\phi(\boldsymbol{y}_e)}\Big\{\nabla_{\boldsymbol{z}_S}\Big[\sum_{\{i,j\}\in\mathcal{V}_O}\log f\big(x_{ij}|\xi_{Ti}+\xi_{Sj},\zeta_{Ti}+\zeta_{Sj},\mu_{Ti}+\mu_{Sj}\big)\Big]\Big\} - \alpha_z K_S\boldsymbol{m}_{\boldsymbol{z}_S}, \tag{37}$$

$$\nabla_{\boldsymbol{\nu}_{\boldsymbol{z}_S}} L = E_{\phi(\boldsymbol{y}_e)}\Big\{\nabla_{\boldsymbol{z}_S}\Big[\sum_{\{i,j\}\in\mathcal{V}_O}\log f\big(x_{ij}|\xi_{Ti}+\xi_{Sj},\zeta_{Ti}+\zeta_{Sj},\mu_{Ti}+\mu_{Sj}\big)\Big]\odot\boldsymbol{z}_{Se}\Big\} - \alpha_z\mathrm{diag}(K_S)\boldsymbol{\nu}_{\boldsymbol{z}_S} + 1\oslash\boldsymbol{\nu}_{\boldsymbol{z}_S}, \tag{38}$$

where $\odot$ and $\oslash$ denote componentwise product and division respectively, and $\phi(\boldsymbol{y}_e) = \prod_{z\in\{\xi,\zeta,\mu\}}\phi(\boldsymbol{z}_{Se})\phi(\boldsymbol{z}_{Te})$. The detailed derivation is presented in Appendix A. Similarly, for the temporal components,

$$\nabla_{\boldsymbol{m}_{\boldsymbol{z}_T}} L = E_{\phi(\boldsymbol{y}_e)}\Big\{\nabla_{\boldsymbol{z}_T}\Big[\sum_{\{i,j\}\in\mathcal{V}_O}\log f\big(x_{ij}|\xi_{Ti}+\xi_{Sj},\zeta_{Ti}+\zeta_{Sj},\mu_{Ti}+\mu_{Sj}\big)\Big]\Big\} - K_T\boldsymbol{m}_{\boldsymbol{z}_T} - \Big(\sum\boldsymbol{m}_{\boldsymbol{z}_T}\Big)\boldsymbol{1}, \tag{39}$$

$$\nabla_{\boldsymbol{\nu}_{\boldsymbol{z}_T}} L = E_{\phi(\boldsymbol{y}_e)}\Big\{\nabla_{\boldsymbol{z}_T}\Big[\sum_{\{i,j\}\in\mathcal{V}_O}\log f\big(x_{ij}|\xi_{Ti}+\xi_{Sj},\zeta_{Ti}+\zeta_{Sj},\mu_{Ti}+\mu_{Sj}\big)\Big]\odot\boldsymbol{z}_{Te}\Big\} - \mathrm{diag}\Big(K_T+\boldsymbol{1}\boldsymbol{1}^T\Big)\cdot\boldsymbol{\nu}_{\boldsymbol{z}_T} + 1\oslash\boldsymbol{\nu}_{\boldsymbol{z}_T}. \tag{40}$$

The gradient of the logarithm of the GEV densities $\log f(x_{ij}|\xi_{Ti}+\xi_{Sj},\zeta_{Ti}+\zeta_{Sj},\mu_{Ti}+\mu_{Sj})$ in the above expressions is listed in Appendix B.

Since the expectations in (37)-(40) are intractable, we approximate them stochastically using Monte Carlo integration. The resulting unbiased stochastic approximation of the gradients are called stochastic gradient. Replacing the exact gradients in (37)-(40) leads to a stochastic optimization algorithm for inferring the optimal variational parameters [32]. Concretely, in each iteration κ, we use one realization of the exact gradients, namely, we draw one sample $\hat{\boldsymbol{y}}_e$ from $\phi(\boldsymbol{y}_e)$ and evaluate the gradients at $\hat{\boldsymbol{y}}^{(\kappa)} = \boldsymbol{\nu}^{(\kappa)}\odot\hat{\boldsymbol{y}}_e + \boldsymbol{m}^{(\kappa)}$, where $a^{(\kappa)}$ denotes the value of parameter a in iteration κ. Therefore,

$$\tilde{\nabla}_{\boldsymbol{m}}L|_{\boldsymbol{m}=\boldsymbol{m}^{(\kappa)}} = \nabla_{\boldsymbol{y}}\log p(\boldsymbol{x}|\boldsymbol{y})\big|_{\boldsymbol{y}=\hat{\boldsymbol{y}}^{(\kappa)}} + \nabla_{\boldsymbol{m}}E_{\phi(\boldsymbol{y}_e)}\big[\log p(\boldsymbol{y}|\boldsymbol{\theta})\big]\big|_{\boldsymbol{m}=\boldsymbol{m}^{(\kappa)}}, \tag{41}$$

$$\tilde{\nabla}_{\boldsymbol{\nu}}L|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(\kappa)}} = \nabla_{\boldsymbol{y}}\log p(\boldsymbol{x}|\boldsymbol{y})\big|_{\boldsymbol{y}=\hat{\boldsymbol{y}}^{(\kappa)}}\odot\hat{\boldsymbol{y}}_e + \nabla_{\boldsymbol{\nu}}E_{\phi(\boldsymbol{y}_e)}\big[\log p(\boldsymbol{y}|\boldsymbol{\theta})\big]\big|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(\kappa)}} + 1\oslash\boldsymbol{\nu}^{(\kappa)}, \tag{42}$$

where $\tilde{\nabla}_{\boldsymbol{m}}L$ and $\tilde{\nabla}_{\boldsymbol{\nu}}L$ represent the stochastic gradients. Only the first terms on the right hand side of the above two equations are approximated stochastically, whereas the other terms can be computed in closed form. We then update all the parameters following a gradient ascent approach:

$$\boldsymbol{m}^{(\kappa+1)} = \boldsymbol{m}^{(\kappa)} + \rho^{(\kappa)}\tilde{\nabla}_{\boldsymbol{m}}L|_{\boldsymbol{m}=\boldsymbol{m}^{(\kappa)}}, \tag{43}$$

$$\boldsymbol{\nu}^{(\kappa+1)} = \boldsymbol{\nu}^{(\kappa)} + \rho^{(\kappa)}\tilde{\nabla}_{\boldsymbol{\nu}}L|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(\kappa)}}, \tag{44}$$

$$\boldsymbol{\theta}^{(\kappa+1)} = \boldsymbol{\theta}^{(\kappa)} + \rho^{(\kappa)}\nabla_{\boldsymbol{\theta}}L|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\kappa)}}, \tag{45}$$

where $\rho^{(\kappa)}$ is the learning rate (or step size) in iteration κ.

When the learning rate schedule follows the Robbins-Monro conditions [32]:

$$\sum_{\kappa=1}^{\infty}\rho^{(\kappa)} = \infty, \qquad \sum_{\kappa=1}^{\infty}\big(\rho^{(\kappa)}\big)^2 < \infty, \tag{46}$$

the stochastic optimization algorithm converges to a local maximum of L. Due to the noisy gradient used in each iteration, the SVI algorithm can easily escape from shallow local maxima of the complex objective function, and converges to at least a significant local maximum. The proposed algorithm can be viewed as a stochastic variational extension of the expectation conjugate gradient algorithm proposed in [33].

The computational complexity of the SVI algorithm is now linear in $|\mathcal{V}_O|$, where $|\mathcal{V}_O|$ denotes the number of nodes in $\mathcal{V}_O$, and it equals $\mathcal{O}(D)$ when there is no missing data. As pointed out in [21], the stochastic gradient can be obtained using a mini-batch of $M \leq |\mathcal{V}_O|$ factors of the joint distribution $p(\boldsymbol{x}, \boldsymbol{y}|\theta)$ (22). The resulting computational cost can be further reduced to be sublinear. As an example, let us focus on the stochastic gradient of $m_{z_{Sj}}$ $(j = 1, \cdots, P)$:

$$\begin{aligned}\frac{\partial L}{\partial m_{z_{Sj}}} =& \sum_{i=1}^{N} \frac{\partial \log f\big(x_{ij}|\xi_{Ti} + \xi_{Sj}, \zeta_{Ti} + \zeta_{Sj}, \mu_{Ti} + \mu_{Sj}\big)}{\partial z_{Sj}} \\ & - \alpha_z [K_S]_{j,:} \boldsymbol{m}_{\boldsymbol{z}_S},\end{aligned} \tag{47}$$

where $[K_S]_{j,:}$ is the jth row of matrix K_S. The first term on the right hand side of the above equation can be equivalently expressed as:

$$NE_{p(i)}\left[\frac{\partial \log f\big(x_{ij}|\xi_{Ti} + \xi_{Sj}, \zeta_{Ti} + \zeta_{Sj}, \mu_{Ti} + \mu_{Sj}\big)}{\partial z_{Sj}}\right], \tag{48}$$

where $p(i)$ is a discrete uniform distribution on the set $\{1, \cdots, N\}$. Thus, we randomly draw a mini-batch of i from the set, compute the corresponding partial derivatives, and approximate the expectation in (48) by the mean value of the partial derivatives. The stochastic gradients $w.r.t$ other parameters can be computed in a similar fashion. In the most extreme case, to compute all the required stochastic gradients, we only need to draw $M = \max(N, P)$ samples x_{ij} uniformly *without replacement* from $\mathcal{V}_O$ such that all the indices i $(i = 1, \cdots, N)$ and j $(j = 1, \cdots, P)$ appear at least once. Then the computational complexity is only $\mathcal{O}(\max(N, P))$. Our numerical experiments demonstrate that using a mini-batch of $\mathcal{V}_O$ can greatly accelerate the algorithm, reducing the computational time from hours to minutes.

1) Selecting the Step Size: One challenge with stochastic optimization methods is setting the learning rate. As the parameters in our problem have completely different scales, if we use a unified step size to update all the parameters, the step size has to be small enough to tackle the smallest scale. The resulting algorithm would converge slowly. To address this concern, we exploit the ADADELTA method [34], which adaptively sets individual dynamic step size for each component of the parameter vector. Specifically, ADADELTA defines the step size as:

$$\boldsymbol{\rho}^{(\kappa)} = \rho_0 \oslash \sqrt{\tilde{E}[(\nabla L)^2]^{(\kappa)} + \epsilon}, \tag{49}$$

where ϵ is a small constant that servers the purpose to better condition the denominator, and $\tilde{E}[(\nabla L)^2]$ is an exponentially decaying moving average of the squared gradients which can be updated as:

$$\tilde{E}[(\nabla L)^2]^{(\kappa)} = \eta \tilde{E}[(\nabla L)^2]^{(\kappa-1)} + (1 - \eta)(\nabla L^{(\kappa)})^2, \tag{50}$$

where $(\nabla L^{(\kappa)})^2$ denotes componentwise square of $\nabla L^{(\kappa)}$. Since the denominator employs the squared gradient information, large gradients have smaller learning rates and vice versa. The ADADELTA method has the nice property as in second-order methods (e.g., Newton's method) that the progress along each dimension evens out over time. On the other hand, as shown in [35], [36], the moving average of squared gradient $\tilde{E}[(\nabla L)^2]$ is a good approximation to $E[(\nabla L)^2]$, which can be further decomposed as:

$$E[(\nabla L)^2] = E[\nabla L]^2 + V[\nabla L], \tag{51}$$

where $V[\nabla L]$ is the variance of the gradient. As a result, the step size decreases with the growing of the variance of the gradient, thus mitigating the risk of taking a large step in a wrong direction.

In our experiments, we follow [34] to set $\eta = 0.95$ and $\epsilon = 10^{-6}$. Additionally, we initialize $\rho_0 = 10^{-3}$ and scale it every 1000 iterations by a factor of 0.99 in a similar manner as in [20], so as to guarantee the convergence of the algorithm.

2) Reducing the Variance of the Gradient: In order to increase the step size and improve the convergence rate, one has to design methods that can reduce the variance of the stochastic gradient. It has been demonstrated in [37]-[39], both empirically and theoretically, that utilizing a fixed-window moving average of stochastic gradients can effectively reduce the variance and highly speed up the stochastic gradient algorithm both empirically and theoretically. Here, we further extend the idea and employ an exponentially decaying moving average in which the decaying rate depends on the variance of the gradient. As in [39], we only compute the moving average of the Monte Carlo approximation part of the stochastic gradient (e.g., the first terms in Eq. (37)-(40)), since other terms are deterministic values that have no influence on the variance. Specifically, let $E[g]$ denote the exact value of the first terms in Eq. (37)-(40), g the Monte Carlo approximation of $E[g]$, and $\tilde{E}[g]$ the exponentially decaying moving average. $\tilde{E}[g]$ can be updated as:

$$\tilde{E}[g]^{(\kappa)} = \left(1 - \frac{1}{\tau_g^{(\kappa)}}\right)\tilde{E}[g]^{(\kappa-1)} + \frac{1}{\tau_g^{(\kappa)}} g^{(\kappa)}, \tag{52}$$

where $\tau_g^{(\kappa)}$ can be viewed as the window size of the moving average in iteration κ. We want the window size to increase when the variance is large, and to decay if the variance becomes small. Note that a good measure of the variance given the stochastic gradients in each iteration is:

$$\omega = \frac{\sum \tilde{E}[\nabla L]^2}{\sum \tilde{E}[(\nabla L)^2]} \approx \frac{\sum E[\nabla L]^2}{\sum E[\nabla L]^2 + \sum V[\nabla L]}, \tag{53}$$

where $\tilde{E}[\nabla L]$ and $\tilde{E}[(\nabla L)^2]$ are the exponentially decaying moving average of the gradient and the squared gradient respectively with the decaying rate η as defined in (50). It is evident that ω grows with the inverse of the variance. Given the current measure of the variance $\omega^{(\kappa)}$, we update the window size $\tau_g^{(\kappa+1)}$ for the next iteration as:

$$\tau_g^{(\kappa+1)} = (1 - \omega^{(\kappa)})\tau_g^{(\kappa)} + 1. \tag{54}$$

As such, $\tau_g \geq 1$ as the algorithm proceeds, and it changes with the variance of the gradient as desired. Interestingly, it can be observed that the length of the moving window will decrease if we take a big step (i.e., ∇L is large and $E[\nabla L]$ increases) in the current iteration. In this case, the gradients in the previous iterations are unreliable, and we can see that they will contribute less to $\tilde{E}[g]$ in the next iteration according to the proposed method. Therefore, such mechanism makes $\tilde{E}[g]$ approximate $E[g]$ more accurately. Note that similar methods are applied in [35], [36] for step size selection. We find that this method works as well for variance reduction. Moreover, it is straightforward to combine the method with ADADELTA to achieve better performance. The method can be further extended to yield different τ_g for different sets of parameters, cf. [35]. Here we only compute a unified τ_g for simplicity. In our experiments, we initialize $\tau_g^{(1)} = 1$. We then replace $g^{(\kappa)}$ by $\tilde{E}[g]^{(\kappa)}$ when computing the stochastic gradient in each iteration.

3) Bounding the GEV Parameters: Note that when the shape parameter $\xi_{ij} = 0$, the expression of the PDF (16) does not involve ξ_{ij}. In order to obtain the partial derivatives with respect to ξ_{ij}, we approximate it by a small value, e.g., $\xi_{ij} = 10^{-6}$, and use the PDF in (14) instead. Additionally, given an observation x_{ij}, the GEV parameters must satisfy the constraint $1 + \xi_{ij}(x_{ij} - \mu_{ij})/\exp(\zeta_{ij}) > 0$, so as to guarantee that the log-likelihood and the corresponding gradient are well defined. However, the variational distribution $q(\boldsymbol{y})$ is defined in $\mathbb{R}$. To address this issue, we borrow the idea of Lagrangian multipliers, and extend the domain of the likelihood of the GEV parameters to $\mathbb{R}$ as follows:

$$\begin{aligned}&\tilde{f}(x_{ij}|\xi_{ij}, \zeta_{ij}, \mu_{ij}) \\ =&\begin{cases} f(x_{ij}|\xi_{ij}, \zeta_{ij}, \mu_{ij}) & \text{if } 1 + \xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right) > 0, \\ \exp\left\{c_1\left[1 + \xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right] - c_2\right\} & \text{otherwise,} \end{cases}\end{aligned} \tag{55}$$

where c_1 and c_2 are sufficiently large positive constants. According to the above definition, when a sample $\hat{\boldsymbol{y}}$ from the variational distribution $q(\boldsymbol{y})$ fails to satisfy the constraint, the corresponding stochastic gradient will move the mean vector of $q(\boldsymbol{y})$ in the direction where the constraints can be satisfied, since in this case the gradient of $\tilde{f}(x_{ij})$ always points to the direction in which the value of $1 + \xi_{ij}(x_{ij} - \mu_{ij})/\exp(\zeta_{ij})$ increases. Moreover, when the constraint is not satisfied, $\tilde{f}(x_{ij})$ is close to zero due to the large positive constant c_2. Therefore, the shape of the original and the extended likelihood are almost the same.

After replacing $f(x_{ij})$ with $\tilde{f}(x_{ij})$ in (22), the value $\boldsymbol{y}^*$ that maximizes $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ is the same as before, as indicated by the following proposition:

Proposition 1. *Let $\boldsymbol{y}^*$ correspond to a local maximum of $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$. If the positive constant c_1 is sufficiently large, then $\boldsymbol{y}^*$ satisfies the constraint that $1 + \xi_{ij}(x_{ij} - \mu_{ij})/\exp(\zeta_{ij}) > 0$, $\forall\{i, j\} \in \mathcal{V}_O$, and $\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ also attains a local maximum at $\boldsymbol{y}^*$.*

Proof. See Appendix C. □

As a result, we can safely replace $\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ with $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ during the learning process. Although the exact value of c_1 cannot be predicted in advance, we find that $c_1 = 10^{10}$ is sufficiently large in our experiments. We summarize the overall algorithm in Table I.

TABLE I: Stochastic variational inference of the spatio-temporal model.

Initialize $E[\nabla_{\boldsymbol{\theta}} L^2]^{(0)} = 0$, $E[\Delta\boldsymbol{\theta}^2]^{(0)} = 0$. Iterate the following steps until $\boldsymbol{m}$ and $\boldsymbol{\nu}$ converge.

1) Draw one sample $\hat{\boldsymbol{y}}^{(\kappa)}$ from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{y}; \boldsymbol{m}^{(\kappa)}, C^{(\kappa)}(C^{(\kappa)})^T)$:
$$\hat{\boldsymbol{y}}_e \sim \phi(\boldsymbol{y}_e), \quad \hat{\boldsymbol{y}}^{(\kappa)} = \boldsymbol{\nu}^{(\kappa)} \odot \hat{\boldsymbol{y}}_e + \boldsymbol{m}^{(\kappa)}.$$
2) Compute the gradient $w.r.t$ the smoothness parameters $\nabla_{\boldsymbol{\theta}} L|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\kappa)}}$ (cf. Eq. (31), (35) and (36)) and the stochastic gradients $w.r.t$ the variational parameters(cf. Eq. (37)-(40)):
$$\begin{aligned}\tilde{\nabla}_{\boldsymbol{m}} L|_{\boldsymbol{m}=\boldsymbol{m}^{(\kappa)}} =& \nabla_{\boldsymbol{y}} \log p(\boldsymbol{x}|\boldsymbol{y})\big|_{\boldsymbol{y}=\hat{\boldsymbol{y}}^{(\kappa)}} \\ &+ \nabla_{\boldsymbol{m}} E_{\phi(\boldsymbol{y}_e)}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right]\big|_{\boldsymbol{m}=\boldsymbol{m}^{(\kappa)}}, \\ \tilde{\nabla}_{\boldsymbol{\nu}} L|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(\kappa)}} =& \nabla_{\boldsymbol{y}} \log p(\boldsymbol{x}|\boldsymbol{y})\big|_{\boldsymbol{y}=\hat{\boldsymbol{y}}^{(\kappa)}} \odot \hat{\boldsymbol{y}}_e \\ &+ \nabla_{\boldsymbol{\nu}} E_{\phi(\boldsymbol{y}_e)}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right]\big|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(\kappa)}} + 1 \oslash \boldsymbol{\nu}^{(\kappa)}.\end{aligned}$$
3) Compute the moving average the first terms of the above two equations to reduce the variance of the stochastic gradients, as described in Section IV-A2.
4) Set the componentwise step size $\rho^{(\kappa)}$ as described in Section IV-A1.
5) Update $\boldsymbol{m}$, $\boldsymbol{\nu}$, $\boldsymbol{\theta}$ as follows:
$$\begin{aligned}\boldsymbol{m}^{(\kappa+1)} &= \boldsymbol{m}^{(\kappa)} + \rho^{(\kappa)} \odot \tilde{\nabla}_{\boldsymbol{m}} L|_{\boldsymbol{m}=\boldsymbol{m}^{(\kappa)}}, \\ \boldsymbol{\nu}^{(\kappa+1)} &= \boldsymbol{\nu}^{(\kappa)} + \rho^{(\kappa)} \odot \tilde{\nabla}_{\boldsymbol{\nu}} L|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(\kappa)}}, \\ \boldsymbol{\theta}^{(\kappa+1)} &= \boldsymbol{\theta}^{(\kappa)} + \rho^{(\kappa)} \odot \nabla_{\boldsymbol{\theta}} L|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\kappa)}}.\end{aligned}$$

B. Prediction of Future GEV Parameters

A primary goal of the spatio-temporal model is to predict future GEV distributions. Since the spatial component of each GEV parameter is time invariant, we only need to extrapolate the temporal component to time points $N+1, N+2, \cdots$. Recall that the thin-plate model with the zero curvature boundary condition serves as a natural tool to predict the future by means of the current trend (cf. Section II). We thus incorporate the future temporal components of GEV parameters at the right end of the neighborhood structure of the temporal model (see Fig. 2c). As a result, the future parameters $\boldsymbol{z}_{Tf}$ and the observed parameters $\boldsymbol{z}_{To}$ together form a Gaussian graphical model with precision matrix $\tilde{K}_T$. According to Schur complement, we can obtain the MAP estimates:

$$\boldsymbol{z}_{Tf} = -[\tilde{K}_T]_{ff}^{-1}[\tilde{K}_T]_{fo}\boldsymbol{z}_{To}. \tag{56}$$

Note that $\boldsymbol{z}_{To}$ denotes the posterior estimates of the GEV parameters inferred by the SVI methods presented in the last subsection. Due to the special structure of the temporal thin-plate model, the expression (56) boils down to:

$$\begin{aligned}&\boldsymbol{z}_{T\,N+(i-1)\tau+1:N+i\tau} = \gamma_z\left(\beta_z K_{pr} + \gamma_z I_{pr}\right)^{-1} \\ &\cdot\left(2\boldsymbol{z}_{T\,N+(i-2)\tau+1:N+(i-1)\tau} - \boldsymbol{z}_{T\,N+(i-3)\tau+1:N+(i-2)\tau}\right),\end{aligned} \tag{57}$$

for $i = 1, 2, \cdots$. Since $(\beta_z K_{pr} + \gamma_z I_{pr})$ is a sparse matrix, the computational complexity of solving the linear system is linear in τ, when applying algorithms such as belief propagation [40] and embedded subgraphs algorithm [41]. Therefore, the proposed model provides an efficient tool for forecasting future GEV distributions. The final estimation of GEV parameters at time point $N+i$ and site j is $z_{N+i,j} = z_{T\,N+i} + z_{Sj}$.

V. Numerical Results

In this section, we apply our model to synthetic and real data. We first show the sublinear computational complexity of the proposed SVI algorithm. We then compare the SVI algorithm with Gibbs sampling when learning the spatio-temporal model. In addition, we also benchmark the proposed spatio-temporal model (STM) against a spatial model (SM; without considering the temporal variation) [5], [6], a temporal model (TM; without considering the spatial variation) [7], and a model with the same shape and scale parameter for all locations and time points (SSSM) [9]. The proposed SVI algorithm is employed to learn the parameters of all four models. We compare the four models using the deviance information criterion (DIC) [17]:

$$\text{DIC} = \bar{D} + p. \tag{58}$$

The first term is defined as the posterior expectation of the deviance:

$$\bar{D} = E_{q(\boldsymbol{y})}[-2\log p(\boldsymbol{x}|\boldsymbol{y})]. \tag{59}$$

It can be regarded as a Bayesian measure of model fit, which attains smaller values for better models. The second term measures the model complexity by the effective number of parameters:

$$p = E_{q(\boldsymbol{y})}[-2\log p(\boldsymbol{x}|\boldsymbol{y})] + 2\log p(\boldsymbol{x}|E_{q(\boldsymbol{y})}[\boldsymbol{y}]). \tag{60}$$

The DIC is a hierarchical model generalization of the Akaike information criterion and the Bayesian information criterion, and it is particularly useful in Bayesian model selection problems [12], [13], [16]. In addition, for synthetic data, we compute mean squared error (MSE) between the estimated GEV parameter and the ground truth for both observed time series and future GEV distributions in the next year. For real data, we assess the predictive performance by evaluating the averaged absolute fractional prediction errors (AAFPE) [15]. More concretely, let $\hat{x}_{N+i,j}$ be the median of the estimated future GEV distribution at time instant $N+i$ and location j. Then, the AAFPE is given by [15]:

$$\text{AAFPE} = \frac{1}{N_f P}\sum_{i=1}^{N_f}\sum_{j=1}^{P}\left\|\frac{\hat{x}_{N+i,j} - x_{N+i,j}}{x_{N+i,j}}\right\|, \tag{61}$$

where $x_{N+i,j}$ is the observed block maximum at time point $N+i$ and location j, and $\|\cdot\|$ denotes the absolute value.

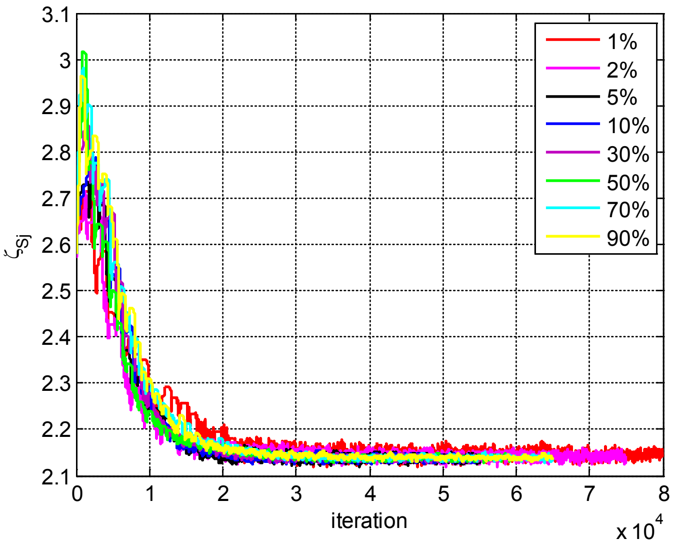


Fig. 3: Estimates of the spatial component of a scale parameter *w.r.t.* number of iterations using a proportion of all the observations.

TABLE II: Performance of the SVI algorithm when the size of the mini-batch changes.

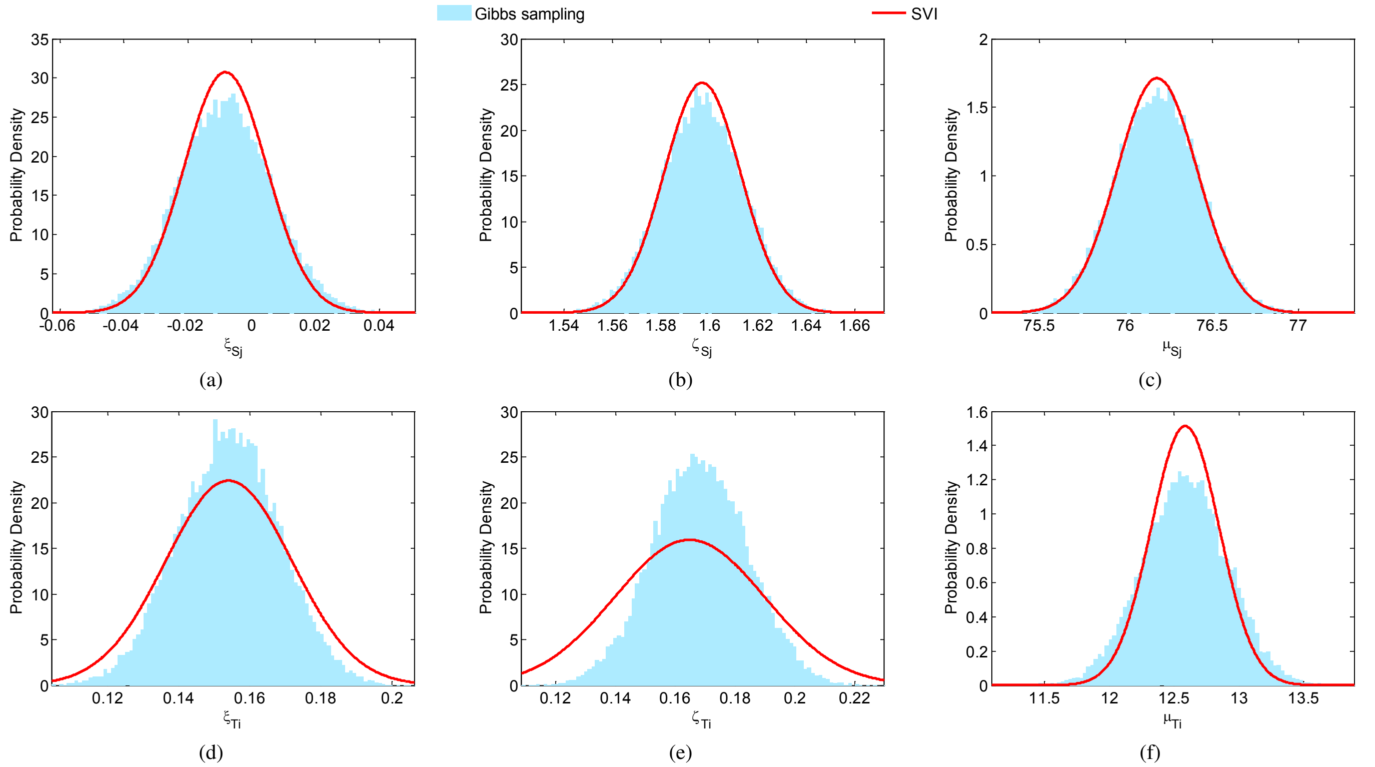
Size of	MSE			Computational	No. of
mini-batch	ξ	ζ	μ	time (s)	iterations
1%	6.08×10^{-4}	1.98×10^{-3}	1.24×10^{-1}	4.93×10^{2}	80000
2%	6.71×10^{-4}	2.00×10^{-3}	1.25×10^{-1}	7.09×10^{2}	75000
5%	6.22×10^{-4}	1.95×10^{-3}	1.25×10^{-1}	1.23×10^{3}	55000
10%	6.01×10^{-4}	1.93×10^{-3}	1.25×10^{-1}	2.45×10^{3}	55000
30%	5.97×10^{-4}	1.92×10^{-3}	1.25×10^{-1}	6.11×10^{3}	55000
50%	6.07×10^{-4}	1.92×10^{-3}	1.25×10^{-1}	1.09×10^{4}	60000
70%	6.03×10^{-4}	1.91×10^{-3}	1.25×10^{-1}	1.51×10^{4}	65000
90%	6.21×10^{-4}	1.92×10^{-3}	1.25×10^{-1}	1.94×10^{4}	65000

A. Synthetic Data

We generate synthetic data by first specifying the GEV parameters in the spatio-temporal domain and then drawing one single sample at each location and time instant. The GEV parameters are defined as quadratic Legendre polynomials of the latitude and longitude of the measuring stations. We then specify the temporal variation by means of trigonometric functions with period $\tau = 12$. Finally, we add an overall polynomial trend to the GEV parameters across time. Concretely, we consider 256 sites arranged on a 16 × 16 regular lattice with 360 monthly maximum observations for each site. We use the data of the first 348 months to learn the model, and retain the rest 12-month data to test the prediction algorithm. Therefore, $D = 89,088$ in this case.

We first explore how the performance of the SVI algorithm changes when using a smaller mini-batch of samples from $\mathcal{V}_O$ to compute the stochastic gradient. Concretely, we use 1%, 2%, 5%, 10%, 30%, 50%, 70%, and 90% of all the samples sequentially. We show how the spatial component of a randomly selected scale parameter ζ_{Sj} varies as the algorithm proceeds in Fig 3. Other related information, such as the accuracy of estimation, the computational time, and the total number of iterations, is listed in Table II. It can be seen that the SVI algorithm converges to the same optimal point regardless of the size of the mini-batch. More importantly, although gradients resulting from a very small minibatch (i.e., 1% and 2%) are very noisy and therefore it takes more iterations before the algorithm can converge, the small computational complexity in each iteration successfully shortens the overall computational time, from hours to minutes. Therefore, unless otherwise stated, we only use 1% of observations to compute the stochastic gradient in the following simulations.

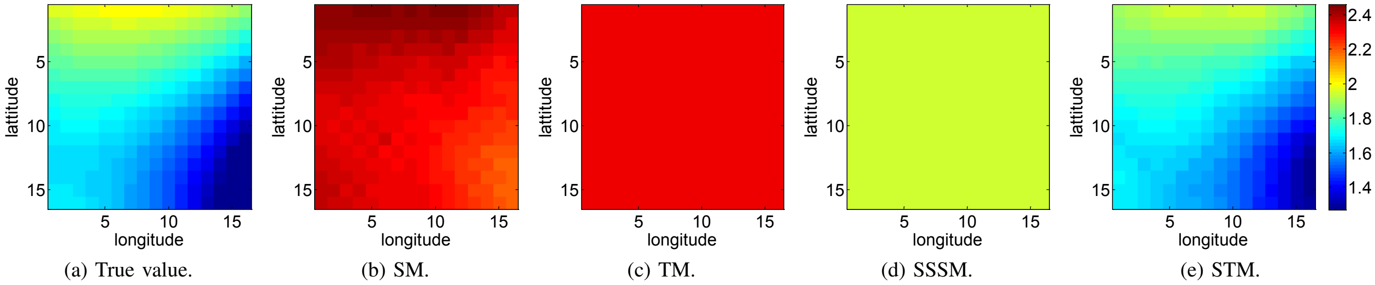
Next, we compare the proposed SVI algorithm with Gibbs sampling to investigate how well the variational distribution can approximate the simulated true posterior distribution. The



(a) (b) (c)

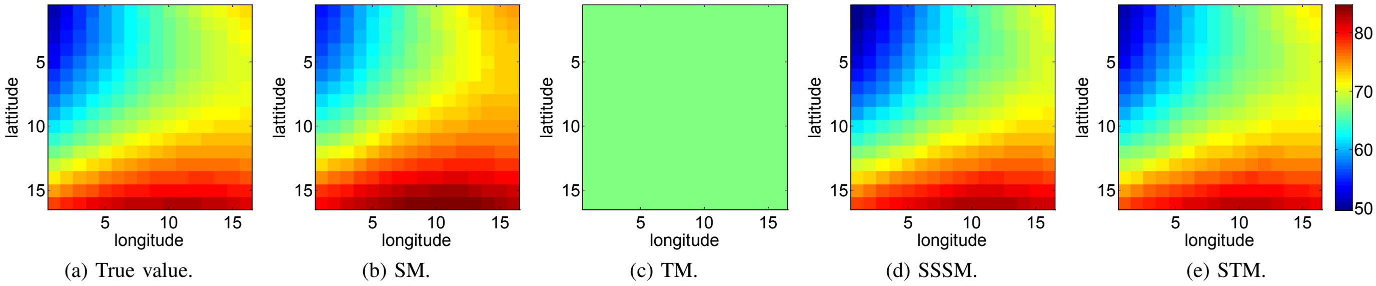
(d) (e) (f)

Fig. 4: Comparison the results from the MCMC method and the SVI method: the distribution of the spatial component of a shape (a), a scale (b), and a location parameter (c) and the distribution of the temporal component of a shape (d), a scale (e), and a location parameter (f).



(a) True value. (b) SM. (c) TM. (d) SSSM. (e) STM.

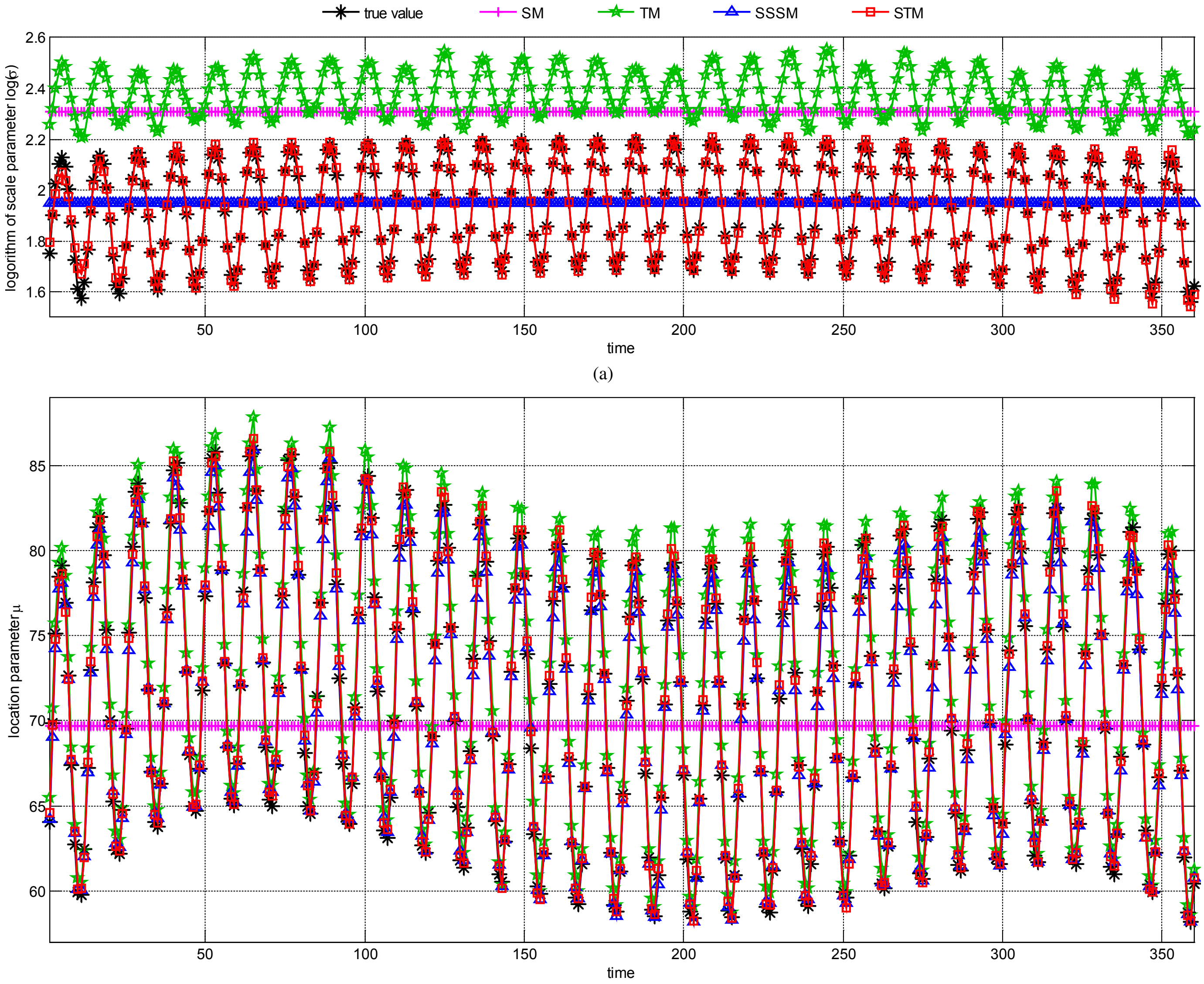
Fig. 5: Estimates of scale parameter σ across all sites at one time point.



(a) True value. (b) SM. (c) TM. (d) SSSM. (e) STM.

Fig. 6: Estimates of location parameter μ across all sites at one time point.

Gibbs sampling procedure is outlined in Appendix D. Similar methods are employed in [8]. Here, we draw $500,000$ samples. We discard the first 5000 samples as burn-in iterations, and further thin the rest samples by a factor 20. We depict in Fig. 4 the estimated distributions of randomly selected spatial and temporal components of GEV parameters resulting from the two methods. As shown in the figure, although the variances of the variational distributions are less consistent with those of the simulated posterior distributions, the mean values are almost identical. Indeed, the MSE between the mean value of the Gibbs samples and the ground truth for the three GEV parameters $(\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\mu})$ is 2.87×10^{-4}, 1.45×10^{-3} and 1.10×10^{-1} respectively, while the corresponding MSE for the SVI algorithm is 6.08×10^{-4}, 1.98×10^{-3}, and 1.24×10^{-1}



(a)

(b)

Fig. 7: Estimates of GEV parameters across time: (a) estimates of scale parameter σ across time; (b) estiamtes of location parameter μ across time.

TABLE III: Quantitative comparison of different models for spatio-temporal extreme-value data

Models	MSE (observed)			MSE (predicted)			DIC
	ξ	ζ	μ	ξ	ζ	μ	
SM	1.68×10^{-2}	2.30×10^{-1}	6.44×10^{1}	3.05×10^{-2}	2.98×10^{-1}	6.24×10^{1}	6.93×10^{5}
TM	2.76×10^{-2}	2.70×10^{-1}	5.23×10^{1}	2.29×10^{-2}	3.31×10^{-1}	5.22×10^{1}	7.11×10^{5}
SSSM	3.18×10^{-2}	6.63×10^{-2}	1.01	5.43×10^{-2}	8.67×10^{-2}	1.38	6.29×10^{5}
STM	6.08×10^{-4}	1.98×10^{-3}	1.24×10^{-1}	2.51×10^{-3}	3.22×10^{-3}	6.93×10^{-1}	6.15×10^{5}

respectively (see Table II), indicating that the SVI algorithm performs comparably with the Gibbs sampling algorithm in terms of MSE. However, it takes 4.09×10^5 seconds to generate all the Gibbs samples, whereas the SVI algorithm only runs for 4.93×10^2 seconds. The computational time of the proposed SVI algorithm is three orders of magnitude shorter than that of the Gibbs sampling.

Now we compare the proposed model with three other models, including a SM, a TM, and a SSSM. The results are summarized in Fig. 5 to Fig. 7 and in Table III. Specifically, Fig. 5 and Fig. 6 shows the estimated scale and location parameters respectively across space resulting from the four models, while Fig. 7 shows the estimated scale and location parameters across time. The results of the shape parameters are qualitatively similar to that of the scale parameters, so we omit them. Table III lists the DIC scores, and the MSE for the estimated GEV parameters with respect to the observed monthly maxima as well as for the predicted GEV parameters in the next year.

As shown in Table III, the proposed STM outshines the competing models in terms of the MSE and the DIC score, and also provides a reliable tool to forecast the GEV distributions in the future. Moreover, we can observe from Fig. 5e, Fig. 6e, and Fig. 7 that the STM yields estimates that closely follow the true temporal and spatial pattern. By contrast, the SM mistakenly ignores the temporal variation (see Fig. 7), and yields biased estimates of the scale and location parameters

TABLE IV: Quantitative comparison of different models for data simulated from the SM, the TM, and the SSSM.

Models	Data simulated from the SM				Data simulated from the TM				Data simulated from the SSSM			
	MSE			DIC	MSE			DIC	MSE			DIC
	ξ	ζ	μ		ξ	ζ	μ		ξ	ζ	μ	
SM	1.68×10^{-4}	2.98×10^{-4}	5.01×10^{-2}	6.07×10^{5}	1.28×10^{-2}	2.22×10^{-1}	6.42×10^{1}	6.88×10^{5}	2.64×10^{-2}	2.13×10^{-1}	6.16×10^{1}	7.00×10^{5}
TM	2.96×10^{-2}	2.89×10^{-1}	5.23×10^{1}	7.82×10^{5}	3.51×10^{-4}	5.19×10^{-4}	9.85×10^{-2}	6.08×10^{5}	3.29×10^{-2}	2.03×10^{-1}	5.22×10^{1}	6.96×10^{5}
SSSM	1.51×10^{-2}	3.64×10^{-2}	6.03×10^{-1}	6.15×10^{5}	1.71×10^{-2}	2.32×10^{-2}	2.49×10^{-1}	6.15×10^{5}	4.21×10^{-5}	6.90×10^{-6}	1.02×10^{-1}	6.36×10^{5}
STM	2.45×10^{-4}	3.73×10^{-4}	5.21×10^{-2}	6.08×10^{5}	6.20×10^{-4}	9.20×10^{-4}	9.04×10^{-2}	6.09×10^{5}	5.03×10^{-4}	5.61×10^{-4}	9.81×10^{-2}	6.36×10^{5}

across space at the randomly selected time instant (see Fig. 5b and Fig. 6b). Similarly, the TM fails to explain the spatial variation of the GEV parameters (see Fig. 5c and Fig. 6c), while wrongly estimating the location and scale parameters in time domain (see Fig. 7). In addition, since the location parameters are assumed to be the same across time and space respectively in the SM and the TM, the observations are more different from the corresponding location parameters in these two models than in other models. In order to capture such large deviance from the location parameters, the scale parameters are overestimated, as demonstrated in Fig. 5 and Fig. 7a. The SSSM, on the other hand, performs better than the above mentioned two models when describing the spatio-temporal dependence among the location parameters. Unfortunately, the assumption that the shape and scale parameters are constant seriously limits the modeling power, and therefore, the resulting DIC score is larger than that of the STM. In addition, the assumption influences the estimation of the location parameters as well. As a consequence, the corresponding estimates are less accurate than that of the STM, which can be seen from Fig. 6d, Fig. 7, and Table III. In summary, we can conclude that it is essential to consider the spatio-temporal variation for all the three GEV parameters when modeling the synthetic data at hand.

In order to further compare the four models, we generate another three synthetic data sets, respectively simulated from the three benchmark models. The GEV parameters are predefined in the same way as before. There are still 256 sites allocated on a 16×16 lattice with 348 monthly maxima observed at each site. Our objective is to show that the proposed model performs as well as the other models, even for data generated by these three models. We summarize the results in Table IV. It is evident that the proposed STM can flexibly handle different types of data, and achieves comparable performance to the underlying true model. The SSSM performs the second best, probably because the varying location parameters are able to capture most of the variation in the data, but not as good as the STM in terms of the MSE and the DIC score. The SM and the TM, however, only yield good results when the data are simulated from these models. From all these results, it becomes clear that the proposed STM is a flexible model with an efficient learning procedure.

Finally, we investigate whether the proposed SVI algorithm can yield accurate estimates of GEV parameters when observations are missing at random. Here, we use the first data set whose GEV parameters vary across both space and time. In this case, $|\mathcal{V}_O| < D$. Fig. 8 shows the MSE for each GEV parameter and for the observed variables ($\{i, j\} \in \mathcal{V}_O$) and the unobserved ones ($\{i, j\} \notin \mathcal{V}_O$) respectively as a function of the percentage of missing variables across 100 trials. We can see that the MSE increases with the number of unmonitored sites and time points, in agreement with our expectation. However, the MSE is still small even when only 10% variables are observed. In conclusion, the proposed model is applicable to cases with missing data.

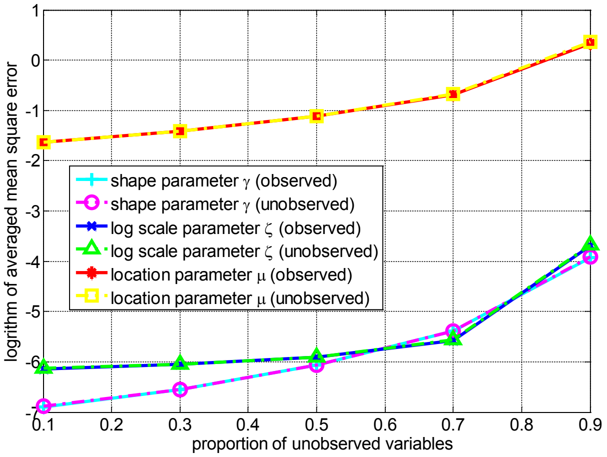
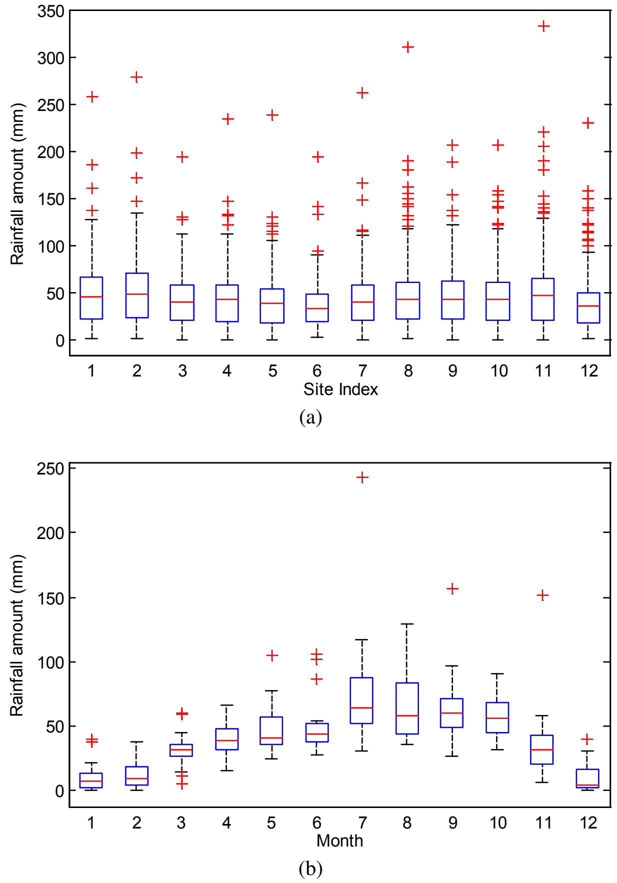


Fig. 8: MSE for varying proportion of missing data (averaged over 100 trials). The MSE increases with growing proportion of missing data, as expected.

B. Real Data

1) Nigeria Precipitation Data: We now consider the extreme precipitation in South Nigeria. The daily rainfall data available at measuring stations from 1979-2005 is interpolated onto a grid with resolution $0.1°$ in [43]. We choose 256 sites arranged on a 16×16 lattice, and extract the monthly maxima for each site. We fit the four models (i.e., the SM, the TM, the SSSM, and the STM) to the first 26-year data, and retain the monthly maxima in 2005 to check the predictive performance.

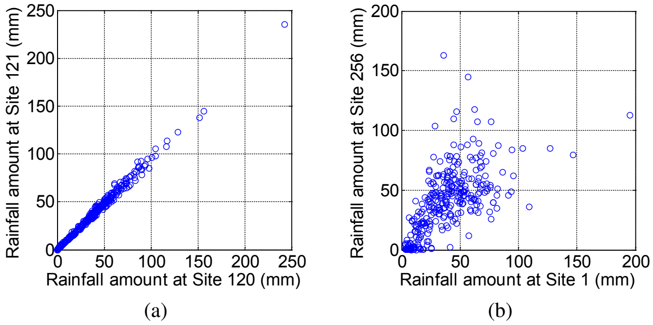
We first conduct an exploratory study on the data. Fig. 9 illustrates the non-stationarity of the data across space and time. We can see that the distribution of monthly maximum rainfall amount varies significantly more across time than across space. Additionally, Fig. 10 shows that there exists strong spatial association in monthly maximum rainfall amount for pairs of nearby sites, but less strong dependence for pairs of sites situated at opposite points of the lattice. This indicates that the GEV distributions, or equivalently the GEV parameters, are similar at nearby sites.



(a)

(b)

Fig. 9: Non-stationarity in extreme rainfall data: (a) Distribution of monthly maximum rainfall of all months in 26 years at 12 randomly selected sites; (b) Distribution of monthly maximum rainfall from January to December of all 26 years at a random site. The distribution of extreme rainfall clearly depends on the location and the month.



(a)

(b)

Fig. 10: Scatter plots of monthly maximum rainfall at pairs of sites: (a) two contiguous sites; (b) two distant sites.

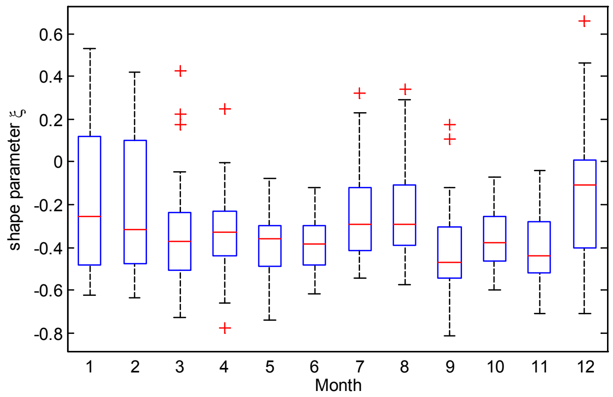


Fig. 11: Estimated shape parameters in different months.

We next estimate parameters of the four models using the SVI algorithm described in Section IV. The DIC scores of the four models are respectively 7.67×10^5, 5.61×10^5, 6.24×10^5, and 5.52×10^5. It is obvious that the proposed STM fits the data the best. We further combine the estimated shape parameters of all 26 years given by the STM and depict in Fig. 11 the distribution of shape parameters for different months at the same site as in Fig. 9b. It can be seen that shape parameters tend to be larger than zero (i.e., the extreme values follow Fréchet distributions) only in January, February, July, August, and December, which happen to be the months with heavy-tailed distributions in Fig. 9b. The remaining months are more likely to have distributions with bounded upper tails. As different months exhibit different tail behaviors, it is improper to assume that the shape parameter is constant across time. This results in the relatively large DIC score of the SSSM in comparison with the STM. Moreover, since there is less heterogeneity across space than across time as demonstrated in Fig. 9, the TM that ignores the spatial variation performs the second best. The spatial model, however, cannot capture the temporal non-stationarity well, leading to the worst fitting. On the other hand, the AAFPEs for the four models are 1.35 (SM), 0.83 (TM), 0.80 (SSSM), and 0.77 (STM). The proposed model attains the smallest prediction error, suggesting that it can forecast extreme-value distributions in the future more reliably.

2) Japan Precipitation Data: We next analyze a real data set of the monthly maximum rainfall amount in Japan. The daily rainfall data from 1900-2011 is compiled and interpolated onto a grid with resolution $0.05°$ [42]. We select one 32×32 regular grid in central Japan, where heavy rainfall is often the cause of floods. Once again, we extract the monthly maxima from 1900 to 2009 to learn the model and hold out the data in 2010 to 2011 for validation purpose. Note that the number of dimensions in this case is $D = 1024$ sites $\times 110$ years $\times 12$ months $= 1,351,680$. The computational time of the proposed SVI algorithm for such large-scale data is only 2.22×10^4 seconds. Similar to the results of the Nigeria data, the DIC scores of the four models are 1.22×10^7 (SM), 1.18×10^7 (TM), 1.07×10^7 (SSSM), and 1.01×10^7 (STM) respectively, while the AAFPEs are 2.07, 2.02, 1.72, 1.68. Hence, the proposed STM achieves the best performance.

VI. Conclusion

In this paper, a novel statistical model is proposed to describe spatio-temporal extreme-value data. Such data are modeled by GEV distributions. The proposed model allows all the three GEV parameters (shape, location, and scale) to change in the spatio-temporal domain, thus characterizing the spatial and temporal dependence in a flexible manner. More explicitly, we assume each GEV parameter can be decomposed as the sum of a spatial and a temporal component, as in a generalized additive model. Graphical models, particularly, thin-plate models, are then imposed on the spatial and the temporal components to capture the spatial and temporal dependence. A stochastic variational inference algorithm is developed to learn the model parameters. Due to the stochastic nature of the algorithm, the computational complexity is sublinear in the number of variables. Thus, as demonstrated in the numerical experiments, the model can handle thousands or even millions of variables in the spatio-temporal domain. Results of both synthetic and real data show that the proposed model can recover the underlying spatio-temporal pattern in an automated manner, given one single observation at each site and time point. Furthermore, it can reliably predict distributions of extreme events in the future.

It is noteworthy that the proposed model can also be easily extended to analyze extreme events with multiple covariates [19], [28]. In future work, we would like to accommodate the covariates of extreme events, and compare the proposed model with existing models. Another interesting direction is to employ graphical models to further capture the dependence between extreme values rather than the GEV parameters [45], [46], which may assist in prediction of the extreme events in the future. Additionally, it is possible to replace the marginal GEV distributions by other distribution families and then apply the model to other types of data. Finally, we will explore multiscale graphical models [47] to capture temporal dependence, since such models are able to model long-range dependence. Therefore, they may yield more reliable long-term predictions.

Appendix A
Derivation of the Gradients

The gradient of the lower bound L with respect to the variational parameters $\boldsymbol{m}_{\boldsymbol{z}_S}$ and $\boldsymbol{\nu}_{\boldsymbol{z}_S}$ can be derived as follows:

$$
\begin{aligned}
&\nabla_{\boldsymbol{m}_{\boldsymbol{z}_S}} L = \nabla_{\boldsymbol{m}_{\boldsymbol{z}_S}} \Big\{ E_{\phi(\boldsymbol{y}_e)} \left[\log p(\boldsymbol{x}|C\boldsymbol{y}_e + \boldsymbol{m})\right] + \\
&E_{\phi(\boldsymbol{z}_{Se})} \left[\log p(\boldsymbol{z}_S|\alpha_z)\right] \Big\} \\
&\quad = E_{\phi(\boldsymbol{y}_e)} \Big\{ \nabla_{\boldsymbol{z}_S} \Big[\sum_{\{i,j\}\in\mathcal{V}_O} \log f(x_{ij}|\xi_{Ti}+\xi_{Sj}, \zeta_{Ti}+\zeta_{Sj}, \mu_{Ti} \\
&\quad + \mu_{Sj}) \Big] \nabla_{\boldsymbol{m}_{\boldsymbol{z}_S}} \boldsymbol{z}_S \Big\} - \alpha_z K_S \boldsymbol{m}_{\boldsymbol{z}_S} \\
&\quad = E_{\phi(\boldsymbol{y}_e)} \Big\{ \nabla_{\boldsymbol{z}_S} \Big[\sum_{\{i,j\}\in\mathcal{V}_O} \log f(x_{ij}|\xi_{Ti}+\xi_{Sj}, \zeta_{Ti}+\zeta_{Sj}, \mu_{Ti} \\
&\quad + \mu_{Sj}) \Big] \Big\} - \alpha_z K_S \boldsymbol{m}_{\boldsymbol{z}_S}
\end{aligned} \tag{62}
$$

$$
\begin{aligned}
&\nabla_{\boldsymbol{\nu}_{\boldsymbol{z}_S}} L = \nabla_{\boldsymbol{\nu}_{\boldsymbol{z}_S}} \Big\{ E_{\phi(\boldsymbol{y}_e)} \left[\log p(\boldsymbol{x}|C\boldsymbol{y}_e + \boldsymbol{m})\right] + \\
&E_{\phi(\boldsymbol{z}_{Se})} \left[\log p(\boldsymbol{z}_S|\alpha_z)\right] \Big\} + 1 \oslash \boldsymbol{\nu}_{\boldsymbol{z}_S} \\
&\quad = E_{\phi(\boldsymbol{y}_e)} \Big\{ \nabla_{\boldsymbol{z}_S} \Big[\sum_{\{i,j\}\in\mathcal{V}_O} \log f(x_{ij}|\xi_{Ti}+\xi_{Sj}, \zeta_{Ti}+\zeta_{Sj}, \mu_{Ti} \\
&\quad + \mu_{Sj}) \Big] \nabla_{\boldsymbol{\nu}_{\boldsymbol{z}_S}} \boldsymbol{z}_S \Big\} - \alpha_z \mathrm{diag}(K_S) \boldsymbol{\nu}_{\boldsymbol{z}_S} + 1 \oslash \boldsymbol{\nu}_{\boldsymbol{z}_S} \\
&\quad = E_{\phi(\boldsymbol{y}_e)} \Big\{ \nabla_{\boldsymbol{z}_S} \Big[\sum_{\{i,j\}\in\mathcal{V}_O} \log f(x_{ij}|\xi_{Ti}+\xi_{Sj}, \zeta_{Ti}+\zeta_{Sj}, \mu_{Ti} \\
&\quad + \mu_{Sj}) \Big] \odot \boldsymbol{z}_{Se} \Big\} - \alpha_z \mathrm{diag}(K_S) \boldsymbol{\nu}_{\boldsymbol{z}_S} + 1 \oslash \boldsymbol{\nu}_{\boldsymbol{z}_S}.
\end{aligned} \tag{63}
$$

Note that $\phi(\boldsymbol{y}_e) = \prod_{z\in\{\xi,\zeta,\mu\}} \phi(\boldsymbol{z}_{Se})\phi(\boldsymbol{z}_{Te})$.

Appendix B
Partial Derivatives of the Logarithm of GEV densities

The logarithm of the PDF of a GEV distribution can be written as:

$$
\begin{aligned}
&\log f(x_{ij}|\xi_{Ti}+\xi_{Sj}, \zeta_{Ti}+\zeta_{Sj}, \mu_{Ti}+\mu_{Sj}) \\
&= -\zeta_{ij} - \left(\frac{1}{\xi_{ij}}+1\right) \log\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right] \\
&\quad - \left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}}},
\end{aligned} \tag{64}
$$

where $\xi_{ij} = \xi_{Ti}+\xi_{Sj}$, $\zeta_{ij} = \zeta_{Ti}+\zeta_{Sj}$, and $\mu_{ij} = \mu_{Ti}+\mu_{Sj}$. As a result, the partial derivatives can be computed as:

$$
\begin{aligned}
&\frac{\partial \log f(x_{ij})}{\partial \xi_{Ti}} = \frac{\partial \log f(x_{ij})}{\partial \xi_{Sj}} \\
&= \frac{1}{\xi_{ij}^2} \log\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right] \\
&\quad - \left(\frac{1}{\xi_{ij}}+1\right)\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-1}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right) - \\
&\quad \frac{1}{\xi_{ij}^2}\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}}} \log\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right] \\
&\quad + \frac{1}{\xi_{ij}}\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}}-1}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right),
\end{aligned} \tag{65}
$$

$$
\begin{aligned}
&\frac{\partial \log f(x_{ij})}{\partial \zeta_{Ti}} = \frac{\partial \log f(x_{ij})}{\partial \zeta_{Sj}} \\
&= -1 + (1+\xi_{ij})\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-1}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right) \\
&\quad - \left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}}-1}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right),
\end{aligned} \tag{66}
$$

$$
\begin{aligned}
&\frac{\partial \log f(x_{ij})}{\partial \mu_{Ti}} = \frac{\partial \log f(x_{ij})}{\partial \mu_{Sj}} \\
&= (1+\xi_{ij})\left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-1}\left(\frac{1}{\exp(\zeta_{ij})}\right) \\
&\quad - \left[1+\xi_{ij}\left(\frac{x_{ij}-\mu_{ij}}{\exp(\zeta_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}}-1}\left(\frac{1}{\exp(\zeta_{ij})}\right).
\end{aligned} \tag{67}
$$

Appendix C
Proof of Proposition 1

We can prove Proposition 1 via contradiction. Let $a_{ij} = 1+\xi_{ij}(x_{ij}-\mu_{ij})/\exp(\zeta_{ij})$.

Suppose that there is a local maximizer $\hat{\boldsymbol{y}}$ of $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ that breaks the constraints that $a_{ij} > 0$ for all $\{i, j\}$ in a set $\mathcal{V}_{-} \subseteq \mathcal{V}_O$. According to the definition in (55), $\forall \{i, j\} \in \mathcal{V}_{-}$,

$$\tilde{f}(x_{ij}|\xi_{ij}, \zeta_{ij}, \mu_{ij}) = \exp\left\{ c_1 \left[1 + \xi_{ij} \left(\frac{x_{ij} - \mu_{ij}}{\exp(\zeta_{ij})} \right) \right] - c_2 \right\}. \tag{68}$$

Thus, we can obtain:

$$\frac{\partial \log \tilde{f}(x_{ij})}{\partial \hat{z}_{Ti}} = \frac{\partial \log \tilde{f}(x_{ij})}{\partial \hat{z}_{Sj}} = \frac{\partial \log \tilde{f}(x_{ij})}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial \hat{z}_{ij}} = c_1 \frac{\partial a_{ij}}{\partial \hat{z}_{ij}}, \tag{69}$$

where $z \in \{\xi, \zeta, \mu\}$. For scale parameters $\hat{\zeta}_{ij}$, it follows from the inequality $1 + \hat{\xi}_{ij}(x_{ij} - \hat{\mu}_{ij}) / \exp(\hat{\zeta}_{ij}) \leq 0$ that

$$\frac{\partial a_{ij}}{\partial \hat{\zeta}_{ij}} = -\hat{\xi}_{ij} \left(\frac{x_{ij} - \hat{\mu}_{ij}}{\exp(\hat{\zeta}_{ij})} \right) \geq 1. \tag{70}$$

Therefore,

$$\frac{\partial \log \tilde{f}(x_{ij})}{\partial \hat{\zeta}_{Ti}} = \frac{\partial \log \tilde{f}(x_{ij})}{\partial \hat{\zeta}_{Sj}} = c_1 \frac{\partial a_{ij}}{\partial \hat{\zeta}_{ij}} \geq c_1. \tag{71}$$

As a result, if c_1 is sufficiently large, for example,

$$c_1 > \left[\left[\left(K_T + \mathbf{1}\mathbf{1}^T \right) \hat{\boldsymbol{\zeta}}_T \right]_i - \sum_{\{j: \{i,j\} \in \mathcal{V}_O | \mathcal{V}_{-}\}} \frac{\partial \log f_{ij}}{\partial \hat{\zeta}_{ij}} \right]_+, \tag{72}$$

for one $i \in \{i : \{i, j\} \in \mathcal{V}_{-}\}$, where $t_+ = \max(t, 0)$, then

$$\begin{aligned} \frac{\partial \log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{\partial \hat{\zeta}_{Ti}} =& c_1 \sum_{\{j: \{i,j\} \in \mathcal{V}_{-}\}} \frac{\partial a_{ij}}{\partial \hat{\zeta}_{ij}} + \sum_{\{j: \{i,j\} \in \mathcal{V}_O | \mathcal{V}_{-}\}} \frac{\partial \log f_{ij}}{\partial \hat{\zeta}_{ij}} \\ & - \left[\left(K_T + \mathbf{1}\mathbf{1}^T \right) \hat{\boldsymbol{\zeta}}_T \right]_i \\ \geq & c_1 + \sum_{\{j: \{i,j\} \in \mathcal{V}_O | \mathcal{V}_{-}\}} \frac{\partial \log f_{ij}}{\partial \hat{\zeta}_{ij}} \\ & - \left[\left(K_T + \mathbf{1}\mathbf{1}^T \right) \hat{\boldsymbol{\zeta}}_T \right]_i \\ > & 0. \end{aligned} \tag{73}$$

The above inequality contradicts the assumption that $\hat{\boldsymbol{y}}$ is a local maximizer of $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$. Therefore, $\hat{\boldsymbol{y}}$ can be a maximum of $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ only if it satisfies the constraint that $a_{ij} > 0$ for all $\{i, j\} \in \mathcal{V}_O$. Furthermore, if the constraint is satisfied, then $\tilde{f}(x_{ij}) = f(x_{ij})$ for all $\{i, j\} \in \mathcal{V}_O$, and therefore, $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) = \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$. As a result, a local maximizer of $\log \tilde{p}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ is also a local maximizer of $\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$.

Appendix D
Gibbs Sampling Method to Learn the Spatio-Temporal Model

In order to employ Gibbs sampling, we construct a full Bayesian model here by imposing Gamma priors on the smoothness parameters. More specifically, we set the shape and rate parameters (a, b) of the Gamma priors to be very small (i.e., 10^{-6}) such that the priors are non-informative. The detailed steps of the Gibbs sampling algorithm are as follows:

1) Updating the spatial components of GEV parameters at each site
   Each component of $\boldsymbol{z}_S = [z_{Sj}]^T$ is updated individually via the Metropolis-Hastings (MH) algorithm. Let us take a location parameter μ_{Sj} as an example. In iteration κ, we first generate a proposal $\mu_{Sj}^{(p)}$ from a Gaussian distribution with mean value $\mu_{Sj}^{(\kappa-1)}$, and then compute the acceptance probability:
   $$r = \min(1, \alpha), \tag{74}$$
   where α is a ratio between GEV likelihoods times the thin-plate model likelihood when $\mu_{Sj}^{(\kappa)} = \mu_{Sj}^{(p)}$ and when $\mu_{Sj}^{(\kappa)} = \mu_{Sj}^{(\kappa-1)}$. Note that other parameters are set to their most recent values. With probability r, $\mu_{Sj}^{(\kappa)}$ is set to $\mu_{Sj}^{(p)}$; otherwise it remains at $\mu_{Sj}^{(\kappa-1)}$. The spatial components of the scale and shape parameters are updated similarly.
2) Updating the temporal components of GEV parameters at each site
   Each component of $\boldsymbol{z}_T = [z_{Ti}]^T$ is updated singly in a similar vein as in the previous step.
3) Updating the smoothness parameters α_z
   The conditional distribution of α_z conditioned on other parameters has a closed form, that is, a Gamma distribution $\text{Gamma}(\alpha_z; a + (P-1)/2, b + \boldsymbol{z}_S^T K_S \boldsymbol{z}_S / 2)$. We therefore draw one sample from this distribution and set $\alpha_z^{(\kappa)}$ to the value of this sample.
4) Updating the smoothness parameters β_z and γ_z
   We update β_z and γ_z using the MH approach since the Gamma priors are not conjugate to the likelihood of β_z and γ_z. Here, we specify the proposal distributions as Gamma distributions. Due to its asymmetry, we need the Hastings correction when computing the ratio.

References

[1] H. Yu, L. Zhang, and J. Dauwels, "Spatio-temporal Graphical Models for Extreme Events," in *Proc. 2014 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 2032-2036, 2014.

[2] M. L. Parry, O. F. Canziani, J. P. Palutikof, van der Linden, and C. E. Hanson, Eds., IPCC, 2007: summary for policymakers, In: *Climate Change 2007: Impacts, Adaptation, and Vulnerability*. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, 2007.

[3] G. Kuhn, S. Khan, A. R. Ganguly, and M. L. Branstetter, " Geosptial-temporal dependence among weekly precipitation extremes with applications to observations and climate model simulations in South America," *Advances in Water Resources*, vol. 30, pp. 2401-2423.

[4] S. G. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer, London, 2001.

[5] P. J. Northrop and P. Jonathan, "Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights," *Environmetrics*, vol. 22, pp. 799–809, 2011.

[6] H. Yu, Z. Choo, J. Dauwels, P. Jonathan, and Q. Zhou, "Modeling Spatial Extreme Events using Markov Random Field Priors," *Proc. ISIT*, pp. 1453–1457, 2012.

[7] V. Chavez-Demoulin and A. C. Davison, "Modelling Times Series Extremes," *Statistical Journal*, vol. 10, pp. 109-133, 2012.

[8] F. Laurini, and F. Pauli, "Smoothing sample extremes: The mixed model approach," *Computational Statistics & Data Analysis*, vol. 53, no. 11, pp. 3842-3864, 2009.

[9] S. E. Neville, M. J. Palmer, and M. P. Wand, "Generalized Extreme Value Additive Model Analysis Via Mean Filed Variational Bayes," *Australian & New Zealand Journal of Statistics*, vol. 53, no. 3, pp. 305-330, 2011.

[10] M. P. Wand, J. T. Ormerod, S. A. Padoan, and R. Frühwirth, "Mean Field Variational Bayes for Elaborate Distributions," *Bayesian Analysis*, vol. 7, no. 2, pp. 847-900, 2012.

[11] G. Huerta and B. Sansó, "Time-Varying Models for Extreme Values," *Environmental and Ecological Statistics*, vol. 14, pp. 285-299, 2007.

[12] H. Sang, and A. E. Gelfand, "Hierarchical modeling for extreme values observed over space and time," *Environmental and Ecological Statistics* vol. 16, pp. 407–426, 2009.

[13] S. Ghosh and B. K. Mallick, "A hierarchical Bayesian spatio-temporal model for extreme precipitation events,", *Environmetrics*, vol. 22, pp. 192-204, 2011.

[14] Y. Liu, M. T. Bahadori, and H. Li, "Sparse-GEV: Sparse Latent Space Model for Multivariate Extreme Value Time Series Modeling," *Proc. ICML*, 2012.

[15] B. Mahmoudian, and M. Mohammadzadeh, "A spatio-temporal dynamic regression model for extreme wind speeds," *Extremes*, vol. 17, no. 2, pp. 221-245, 2014.

[16] D. Cooley, D. Nychka, P. Naveau, "Bayesian spatial modeling of extreme precipitation return levels," *J Am Stat Assoc* vol. 102, pp. 824-840, 2007.

[17] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. van der Linde "Bayesian measures of model complexity and fit," *J Roy Stat Soc*, Ser B, vol. 64, no. 4, pp. 583-639, 2002.

[18] A. T. Ihler, S. Kirshner, M. Ghil, A. W. Robertson, and P. Smyth, "Graphical Models for Statistical Inference and Data Assimilation," *Physica D* vol. 230, pp. 72-87, 2007.

[19] V. Chavez-Demoulin and A. C. Davison, "Generalized additive modelling of sample extremes," *J. Roy. Stat. Soc. C-App.*, vol. 54, pp. 207-222, 2005.

[20] M. Titsias, M. Lázaro-Gredilla, "Doubly Stochastic Variational Bayes for non-Conjugate Inference," *Journal of Machine Learning Research*, W&CP, vol. 32, no. 1, 1971-1979, 2014.

[21] M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303-1347, 2013.

[22] S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, 1996.

[23] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall, 2005.

[24] D. M. Malioutov, J. K. Johnson, M. J. Choi, and A. S. Willsky, "Low-rank variance approximation in GMRF models: Single and multiscale approaches," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4621-4634, 2008.

[25] P. L. Speckman and D. C. Sun, "Fully Bayesian spline smoothing and intrinsic autoregressive priors," *Biometrika*, vol. 90, pp. 289-302, 2003.

[26] Y. Yue and P. L. Speckman, "Nonstationary Spatial Gaussian Markov Random Fields," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 96-116, 2010.

[27] C. J. Paciorek, "Spatial models for point and areal data using Markov random fields on a fine grid," *Electronic Journal of Statistics*, vol. 7, pp. 946-972, 2013.

[28] H. Yu, J. Dauwels, P. Jonathan, "Extreme-Value Graphical Models with Multiple Covariates," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5734-5747, 2014.

[29] G. Casella, "An Introduction to Empirical Bayes Data Analysis," *American Statistician*, vol. 39, no. 2, pp. 83-87, 1985.

[30] M. J. Wainwright, and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, nos. 1-2, pp. 1-305, 2008.

[31] A. J. Laub, *Matrix Analysis for Scientists and Engineers*, SIAM: Society for Industrial and Applied Mathematics, 2004.

[32] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400-407, 1951.

[33] R. Salakhutdinov, S. Roweis, and Z. Ghahramani, "Optimization with EM and Expectation-Conjugate-Gradient," *Proc. ICML*, 2003.

[34] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *arXiv:1212.5701*, 2012.

[35] T. Schaul, S. Zhang, Y. LeCun, "No more Pesky Learning Rates," *Proc. ICML*, 2013.

[36] R. Ranganath, C. Wang, D. M. Blei, and E. P. Xing, "An Adaptive Learning Rate for Stochastic Variational Inference," *J. Mach. Learning Research*, W&CP vol. 28, 2013.

[37] M. Schmidt, N. L. Roux, and F. Bach, "Minimizing Finite Sums with the Stochastic Average Gradient," *arXiv:1309.2388*, 2013.

[38] R. Johnson, and T. Zhang, "Accelerating Stochastic Graident Descent using Predictive Variance Reduction," *Proc. NIPS*, 2013.

[39] S. Mandt, and D. Blei, "Smoothed Gradients for Stochastic Varianational Inference," *Proc. NIPS*, 2014.

[40] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Neural Comput.*, vol. 13, pp. 2173-2200, 2001.

[41] V. Chandrasekaran, J. K. Johnson, and A. S. Willsky, "Estimation in Gaussian graphical models using tractable subgraphs: A walk-sum analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1916-1930, 2008.

[42] K. Kamiguchi, O. Arakawa, A. Kitoh, A. Yatagai, A. Hamada, and N. Yasutomi, "Development of APHRO_JP, the first Japanese high-resolution daily precipitation product for more than 100 years," *Hydrological Research Letters* vol. 4, pp. 60-64, 2010.

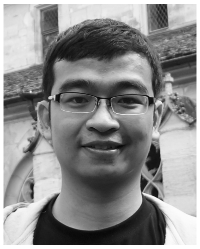
[43] J. Sheffield, G. Goteti, and E. F. Wood, "Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling," *J. Climate*, vol. 19, no. 13, pp. 3088-3111, 2006.

[44] F. Fujibe, N. Yamazaki, M. Katsurayama and K. Kobayashi, "The increasing trend of intense precipitation in Japan based on four-hourly data for a hundred years. *SOLA*, vol. 1, pp. 41-44, 2005.

[45] H. Yu, Z. Choo, W. I. T. Uy, J. Dauwels, and P. Jonathan, "Modeling Extreme Events in Spatial Domain by Copula Graphical Models," *Proc. Fusion 2012*, pp. 1761- 1768, 2012.

[46] H. Yu, W. I. T. Uy, and J. Dauwels, "Modeling Spatial Extremes via Ensemble-of-Trees of Pairwise Copulas," *Proc. ICASSP 2014*, pp. 2415-2419, 2014.

[47] M. J. Choi, V. Chandrasekaran, D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Multiscale stochastic modeling for tractable inference and data assimilation", *Comput. Methods Appl. Mech. Engrg.*, vol. 197, pp. 3492-3515, 2008.



Hang Yu (S'12) received the B.E. degree in electronic and information engineering from University of Science and Technology Beijing (USTB), China in 2010. He is currently working towards the Ph.D. degree in electrical and electronic engineering at Nanyang Technological University (NTU), Singapore.

His research interests include statistical signal processing, machine learning, graphical models, copulas, and extreme-events modeling.

Justin Dauwels (S'02-M'05-SM'12) is an Assistant Professor with School of Electrical and Electronic Engineering (EEE) at the Nanyang Technological University (NTU) in Singapore. He is also the Deputy Director of the ST Engineering-NTU Corporate Lab and the Director of the Neuroengineering Program at the School of EEE. His research interests are in Bayesian statistics, iterative signal processing, and computational neuroscience. He obtained the PhD degree in electrical engineering at the Swiss Polytechnical Institute of Technology (ETH) in Zurich in December 2005. He was a postdoctoral fellow at the RIKEN Brain Science Institute (2006-2007) and a research scientist at the Massachusetts Institute of Technology (2008-2010). He has been a JSPS postdoctoral fellow (2007), a BAEF fellow (2008), a Henri-Benedictus Fellow of the King Baudouin Foundation (2008), and a JSPS invited fellow (2010, 2011). His research on intelligent transportation systems has been featured by the BBC, national TV, Straits Times, and various other media outlets. His research on Alzheimer's disease is featured at a 5-year exposition at the Science Center in Singapore. His research team has won several best paper awards at international conferences. He has filed 5 US patents related to data analytics.