Extreme-Value Graphical Models With Multiple Covariates

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

***Abstract*—To assess the risk of extreme events such as hurricanes, earthquakes, and floods, it is crucial to develop accurate extreme-value statistical models. Extreme events often display heterogeneity (i.e., nonstationarity), varying continuously with a number of covariates. Previous studies have suggested that models considering covariate effects lead to reliable estimates of extreme events distributions. In this paper, we develop a novel statistical model to incorporate the effects of multiple covariates. Specifically, we analyze as an example the extreme sea states in the Gulf of Mexico, where the distribution of extreme wave heights changes systematically with location and storm direction. In the proposed model, the block maximum at each location and sector of wind direction are assumed to follow the Generalized Extreme Value (GEV) distribution. The GEV parameters are coupled across the spatio-directional domain through a graphical model, in particular, a three-dimensional (3D) thin-membrane model. Efficient learning and inference algorithms are developed based on the special characteristics of the thin-membrane model. We further show how to extend the model to incorporate an arbitrary number of covariates in a straightforward manner. Numerical results for both synthetic and real data indicate that the proposed model can accurately describe marginal behaviors of extreme events.**

***Index Terms*—Covariates, extreme events modeling, Gaussian graphical models, Kronecker product, Laplacian matrix.**

I. Introduction

EXTREME events, such as heat waves, cold snaps, tropical cyclones, hurricanes, heavy precipitation and floods, droughts and wild fires, have possibly tremendous impact on people's lives and properties. For instance, China experienced massive flooding of parts of the Yangtze River in the summer of 1998, resulting in about 4,000 dead, 15 million homeless and 26 billion USD in economic loss. To make matters worse, both observational data and computer climate models suggest that the occurrence and sizes of such catastrophes will increase in the future [3]. It is therefore imperative to model such events, assess the risk, and further take precaution measurements.

Extreme-value theory governs the statistical behavior of extreme values of variables, such as extreme wave heights during hurricanes. The theory provides closed-form distribution functions for the extremes of single variables (marginals), such as block maxima (monthly or annually) and peaks over a sufficiently high threshold [4]. The main challenge in fitting such distributions to measurements is the lack of data, as extreme events are by definition very rare. The problem can be alleviated by assuming that all the collected data (e.g., extreme wave heights at different measuring sites [5]) are stationary and follow the same distribution. After combining all the data, the resulting sample size is sufficiently large to yield apparently reliable estimates. However, there usually exists clear heterogeneity in the extreme-value data caused by the underlying mechanisms that drive the weather events. Extreme temperature, for example, is greatly influenced by the altitude of the measuring site. The latter can be regarded as a covariate. Accommodating heterogeneity in the model is essential since the estimated model will be unreliable otherwise [6]. In order to handle both heterogeneity as well as the problem of small sample size, the interactions among extreme events with different covariate values are often exploited. For instance, extreme temperatures at similar altitudes behave similarly, implying that the parameters of the corresponding extreme-value distributions vary smoothly with the covariate (i.e., altitude). Such prior knowledge may help to improve the fitting of extreme-value distributions.

The large body of literature on extreme-value models with covariates can be divided into two categories: models with single [7]–[11] and multiple covariates [12]–[15]. Approaches of the first group usually treat the parameters of the marginal extreme-value distributions as a function of the covariate. In [7]–[9], directional and seasonal effects are considered when describing the marginal behavior of extreme wave heights. The dependence of the parameters on the single covariate is captured by a Fourier expansion. Spatial effects are investigated in [10]: the parameters are assumed to be Legendre polynomials of the location; extreme-value threshold is determined through quantile regression. Although these parametric models offer a simple framework to capture the covariate effect, they are prone to model misspecification. A more appealing approach is to incorporate the covariate in a non-parametric manner. The work in [11] employs a Markov random field, in particular, a conditional au-

Manuscript received January 15, 2014; revised June 22, 2014 and August 11, 2014; accepted September 04, 2014. Date of publication September 18, 2014; date of current version October 08, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pascal Larzabal. This work is supported by MOE Tier 2 project ARC5/14. Part of the material in this paper was presented at the IEEE International Symposium on Information Theory (ISIT), Cambridge, MA, USA, July 2012 [1], and at the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014 [2].

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

P. Jonathan is with Shell Research Ltd., Manchester M22 0RR, U.K., and also with the Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, U.K. (e-mail: philip.jonathan@shell.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2014.2358955

toregressive model, to induce spatial dependence among the parameters of extreme-value distributions across a spatial domain. This also enables the use of the Markov Chain Monte Carlo (MCMC) algorithm to learn the model. However, such procedures are computationally complex and can be prohibitive for large-scale systems.

On the other hand, few attempts have been made to model multiple covariates. A standard method is to predefine the distribution parameters as a function of all the covariates [4]. Unfortunately, the function can be quite complicated as the number of covariates increases, whereas only linear or log-linear model are used in [4] for simplicity. The resulting estimates may be biased due to the misspecification of the functional form. As an alternative, Eastoe *et al.* [12] proposed to remove the heterogeneity of the entire data set, both extreme and non-extreme, through preprocessing and then model the extremal part of the preprocessed data using the above mentioned standard approach. They found that the preprocessing technique can indeed remove almost all the heterogeneity, and consequently, simple linear or log-linear models are capable of expressing the residual heterogeneity. Motivated by this success, Jonathan *et al.* [13] proposed to process two different covariates individually. They removed the effect of the first covariate by whitening the data using a linear location-scale model and then employed the methods for a single covariate to accommodate the second one. A setback of their method, however, is that the dependence on the first covariate may be nonlinear and therefore the whitening step cannot completely remove the effect of the first covariate. Furthermore, capturing the two covariates independently fails to consider the possible correlation between them. To accommodate all the covariates at the same time, a spline-based generalized additive model is introduced in [14], where the spline smoothers for each covariate are added to the original likelihood function and then the penalized likelihood is maximized. Similarly, Randell *et al.* [15] addressed the problem by means of penalized tensor products of B-splines so as to obtain a smooth dependence of the distribution parameters w.r.t. all the covariates. The spline-based methods have the virtue of extending the methods for single covariates to the case of multiple covariates. Unfortunately, directly maximizing the complex penalized likelihood has several problems. First, the algorithm can be time-consuming. Typically, Newton's method is first employed and the iterative back fitting is then used to solve each Newton step. Thus, the algorithm has at least two loops, and the inner loop, i.e., the iterative back fitting, usually has a slow rate of convergence. Moreover, good initial points are essential for the algorithm to find the global solution. Finally, choosing the smoothness parameters can be problematic as pointed out in [16].

The aforementioned shortcomings spark our interest in exploiting graphical models to incorporate multiple covariates in the extreme-value model. The interdependencies between extreme values with different covariates are often highly structured. This structure can in turn be leveraged by the graphical model framework to yield very efficient algorithms [17], [18]. In the following, we briefly review the literature on inference in graphical models. The most popular tool used in this area is belief propagation (BP) [19]. It provides an efficient linear-complexity algorithm for exact inference in tree graphs. However, cyclic graphs compare favorably with tree graphs in practice due to their richer modeling power. To deal with the complex estimation problem in cyclic graphs, various algorithms have been and are still being proposed. A straightforward method is to extend BP to cyclic graphical models, resulting in loopy belief propagation (LBP) [20]. Unfortunately, LBP is not guaranteed to converge or give accurate results. Alternatively, the embedded subgraphs algorithm [21], [22] employs the idea of performing inference iteratively on tractable subgraphs. Another tempting idea is to decompose the complicated inference problem into simpler subproblems, including the recursive model-reduction method [23] and the Lagrangian relaxation technique [24].

In our previous works [1], [25]–[27], we have demonstrated the utility of graphical models for spatial and spatio-temporal extremes. In this paper, we aim to model extreme events with multiple covariates using graphical models. Our model is theoretically significant because it is among the first approaches to exploit the framework of graphical models to analyze the marginal behavior of extreme events. As an example, we model the storm-wise maxima of significant wave heights in the Gulf of Mexico (see Fig. 6), where the covariates are longitude, latitude, and wind direction. Note that the significant wave height is a standard measure of sea surface roughness; it is defined as the mean of the highest one third of waves (typically in a three-hour period). Theoretically, significant wave heights are affected by dominant wave direction (i.e., storm direction). However, we use wind direction as a surrogate since wind and wave direction are generally fairly well correlated, especially for extreme events.

The proposed model is derived from the following ideas: Motivated by the extreme value theory [4], the extreme events are assumed to follow the fat-tailed Generalized Extreme Value (GEV) distributions. The parameters of those GEV distributions are further assumed to depend smoothly on the covariates. To facilitate the use of graphical models, we discretize the continuous covariates within a finite range. In the example of extreme wave heights in the Gulf of Mexico, space is discretized as a finite homogeneous two-dimensional lattice, and the wind direction is discretized in a finite number of equal-sized sectors. More generally, the GEV distributed variables (and hence also the GEV parameters) are defined on a finite number of points indexed by the (discretized) covariates. We characterize the dependence between the GEV parameters through a graphical model prior, in particular, a multidimensional thin-membrane model where edges are only present between pairs of neighboring points (see Fig. 1). We demonstrate that the multidimensional model can be constructed flexibly from one-dimensional thin-membrane models for each covariate. The proposed model can therefore easily be extended to cope with an arbitrary number of covariates. We follow the empirical Bayes approach to learn the parameters and hyper-parameters. Specifically, both the smoothed GEV parameters and the smoothness parameters are inferred via Expectation Maximization. A major challenge lies in the scalability of the algorithm since the dimension of the model is usually quite large. Instead of using the aforementioned algorithms for solving cyclic graphical models [20]–[24], we take advan-

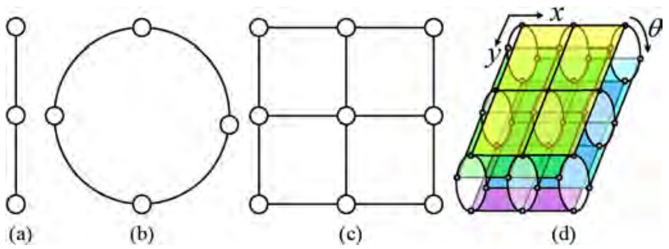


Fig. 1. Thin-membrane models: (a) chain graph; (b) circle graph; (c) lattice; (d) spatio-directional model.

tage of the special pattern of the eigenvalues and eigenvectors corresponding to the one-dimensional thin-membrane models, and derive an efficient inference algorithm specialized for the proposed model.

Our numerical results for both synthetic and real data suggest that the proposed model indeed accurately captures the effect of covariates on the statistics of extreme events. Moreover, the proposed model can flexibly accommodate a large variety of (smooth) dependencies on covariates, as the smoothness parameters of the thin-membrane model are inferred automatically from data.

The remainder of the paper is organized as follows. In Section II, we introduce thin-membrane models and their properties. In Section III, we first construct the 3D thin-membrane model based on simple 1D thin-membrane models to model extreme wave heights, and then illustrate how to generalize the model to incorporate any number of covariates. In Section IV, we discuss the efficient learning and inference algorithms at length and provide theoretical guarantees. Numerical results for both synthetic and real data are presented in Section V. Lastly, we offer concluding remarks and point out directions for future work in Section VI.

II. Thin-Membrane Models

In this section, we first give a brief introduction to graphical models, and subsequently consider the special case of thin-membrane models. We then analyze two concrete examples of thin-membrane models that will be used as building blocks in the proposed extreme-value graphical model: chain and circle models.

In an undirected graphical model (i.e., Markov random field), the probability distribution is represented by an undirected graph $\mathcal{G}$ which consists of nodes $\mathcal{V}$ and edges $\mathcal{E}$. 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: $P(Z_i, Z_j|Z_{\mathcal{V}|i,j}) = P(Z_i|Z_{\mathcal{V}|i,j})P(Z_j|Z_{\mathcal{V}|i,j})$, where $\mathcal{V}|i,j$ denotes all the variables except Z_i and Z_j.

If the random variables Z corresponding to the nodes on the graph are jointly Gaussian, then the graphical model is called a Gaussian graphical model. Let $Z \sim \mathcal{N}(\mu, \Sigma)$ with mean vector μ and positive-definite covariance matrix Σ. Since Z constitutes a Gaussian graphical model, the precision matrix (the inverse covariance) $K = \Sigma^{-1}$ is sparse with respect to the graph $\mathcal{G}$, i.e., $[K]_{i,j} \neq 0$ if and only if the edge $(i,j) \in \mathcal{E}$ [28]. The Gaussian graphical model can be written in an equivalent information form $\mathcal{N}(K^{-1}h, K^{-1})$ with a precision matrix K and a potential vector $h = \Sigma^{-1}\mu$. The corresponding probability density function (PDF) is

$$P(X) \propto \exp\left(-\frac{1}{2}X^TKX + h^TX\right). \tag{1}$$

The thin-membrane model [29] is a Gaussian graphical model that is commonly used as smoothness prior, as it minimizes the difference between values at neighboring nodes:

$$\begin{aligned} P(Z) &\propto \exp\left\{-\frac{1}{2}\alpha\sum_{i\in\mathcal{V}}\sum_{j\in N(i)}(Z_i - Z_j)^2\right\} && (2)\\ &\propto \exp\left\{-\frac{1}{2}\alpha Z^T K_{\rm tm} Z\right\}, && (3) \end{aligned}$$

where $N(i)$ denotes the neighboring nodes of node i, $K_{\rm tm}$ is a graph Laplacian matrix such that $[K_{\rm tm}]_{i,i}$ is the number of neighbors of the i^{th} node while the off-diagonal elements $[K_{\rm tm}]_{i,j}$ are equal to -1 if nodes i and j are adjacent and 0 otherwise, and α is the smoothness parameter which controls the smoothness across the domain defined by the thin-membrane model. By comparing (3) with (1), we can see that the precision matrix of the thin-membrane model is $K = \alpha K_{\rm tm}$. Since $K_{\rm tm}$ is a Laplacian matrix, K is rank deficient, i.e., $\det K = 0$. As such, the thin-membrane model is classified as a partially informative normal prior [30] or intrinsic Gaussian Markov random field [31]. To make the distribution well-defined, the improper density function is usually applied in practice:

$$\begin{aligned} P(Z) &\propto |K|_+^{0.5}\exp\left\{-\frac{1}{2}Z^TKZ\right\} && (4)\\ &= |\alpha K_{\rm tm}|_+^{0.5}\exp\left\{-\frac{1}{2}\alpha Z^T K_{\rm tm} Z\right\}, && (5) \end{aligned}$$

where $|K|_+$ denotes the product of nonzero eigenvalues of K. Note that $K_{\rm tm}1 = 0$, where 1 is a vector of all ones, indicating the eigenvalue associated with the eigenvector 1 equals 0. Thus, the thin-membrane model is invariant to the addition of $c1$, where c is an arbitrary constant, and it allows the deviation from any overall mean level without having to specify the overall mean level itself. As an illustration, we can easily find that the conditional mean of variable Z_i is [31]

$$E\left(Z_i|Z_{\mathcal{V}|i}\right) = -\frac{1}{[K_{\rm tm}]_{i,i}}\sum_{j\in N(i)}[K_{\rm tm}]_{i,j}Z_j, \tag{6}$$

i.e., the mean of its neighbors, but it does not involve an overall level. This special behavior is often desirable in applications.

We now turn our attention to the thin-membrane models of the chain and the circular graph, as shown in Fig. 1(a) and Fig. 1(b) respectively. The former can well characterize the dependence structure of nonperiodic covariates (e.g., longitude and latitude), while the latter is highly suitable for periodic covariates (e.g., directional and seasonal patterns). The corresponding Laplacian matrices are denoted as K_B and K_C respectively. It is easy to prove that the eigenvalues and eigenvectors of K_B and K_C have the following special pattern [32]:

$$\begin{aligned}\lambda_{Bk} =& 2 - 2\cos\left(\frac{k\pi}{P_K}\right),\\ v_{Bk} =& \left[\cos\!\left(\frac{k\pi}{2P_K}\right), \cos\left(\frac{3k\pi}{2P_K}\right), \cdots, \cos\left(\frac{(2P_K-1)k\pi}{2P_K}\right)\right]^T,\\ \lambda_{Ck} =& 2 - 2\cos\left(\frac{2k\pi}{P_K}\right),\\ v_{Ck} =& \left[1, \omega^K, \omega^{2k}, \cdots, \omega^{(P_K-1)k}\right]^T,\end{aligned}$$

for $k = 1, \cdots, P_K$, where P_K is the dimension of K_B and K_C, $\omega = \exp(2\pi i/P_K)$ and i is the imaginary unit. An important property of the eigenvectors is as follows: Let V_B be the eigenvector matrix of K_B, i.e., $V_B = [v_{B1}, v_{B2}, \cdots, v_{BP_K}]$, and x be a $P_K \times 1$ column vector, then $V_B x$ is identical to the discrete cosine transform of x [32, Ch. 1]. Similarly, let V_C denote the eigenvector matrix of K_C, then $V_C x$ can be computed as the discrete Fourier transform of x [32, Ch. 1]. Note that other smoothness priors, such as thin-plate models [29], [31], do not have such nicely structured eigendecomposition. This motivates us to use thin-membrane models for the sake of computational convenience.

III. Modeling Multiple Covariates

In this section, we present a novel extreme-value statistical model that incorporates multiple covariates. We assume that the block maxima (e.g., monthly or annual maxima) associated with every possible set of values for the covariates follow the Generalized Extreme Value (GEV) distribution, according to the Fisher-Tippett-Gnedenko theorem in extreme value theory [4]. The GEV parameters z are assumed to vary smoothly with the covariates. The latter are discretized within a finite range. Therefore, the GEV parameters can be indexed as $z_{i_1 i_2 \cdots i_m}$, where m is the number of covariates, and i_1 corresponds to a discretized value of the first covariate and likewise for the other indices $i_2, \cdots, i_m$. In summary, the GEV parameters are defined on a finite number of points $(i_1, i_2, \ldots, i_m)$ and those parameters are supposed to vary smoothly from one point to a nearby point. The resulting dependence structure can be well represented by a multidimensional thin-membrane model. Furthermore, we show that the multidimensional model can be constructed from one-dimensional thin-membrane models for each covariate, thereby making the proposed model generalizable to accommodate as many covariates as required. We employ the multidimensional thin-membrane model as the prior and estimate the GEV parameters through an empirical Bayes approach.

As an illustration, we present in detail the spatio-directional model to quantify the extreme wave heights. Suppose that we have N samples $x_{ijk}^{(n)}$ (block maxima) at each location, indexed by its longitude and latitude (i, j), and directional sector k, where $n = 1, \cdots, N$, $i = 1, \cdots, P$, $j = 1, \cdots, Q$, $k = 1, \cdots, D$ and P, Q, D are the number of longitude indices, latitude indices and directional sectors respectively. Consequently, the dimension M of the proposed model is given by $M = PQD$. Hence, our objective is to accurately infer the GEV parameters with the consideration of both spatial and directional dependence. Specifically, we first locally fit the GEV distribution to block maxima at each location and in each directional sector. The local estimates are further smoothed across the spatio-directional domain by means of 3D thin-membrane models.

A. Local Estimates of GEV Parameters

We assume that the block maxima x_{ijk} follow a Generalized Extreme Value (GEV) distribution [4], whose cumulative probability distribution (CDF) equals:

$$F(x_{ijk}) = \begin{cases} \exp\left\{-\left[1 + \frac{\gamma_{ijk}}{\sigma_{ijk}}(x_{ijk} - \mu_{ijk})\right]^{-\frac{1}{\gamma_{ijk}}}\right\}, & \gamma_{ijk} \neq 0 \\ \exp\left\{-\exp\left[-\frac{1}{\sigma_{ijk}}(x_{ijk} - \mu_{ijk})\right]\right\}, & \gamma_{ijk} = 0, \end{cases}$$

for x_{ijk} satisfying $1 + \gamma_{ijk}/\sigma_{ijk}(x_{ijk} - \mu_{ijk}) > 0$ if $\gamma_{ijk} \neq 0$ and $x_{ijk} \in \mathbb{R}$ if $\gamma_{ijk} = 0$, where $\mu_{ijk} \in \mathbb{R}$ is the location parameter, $\sigma_{ijk} > 0$ is the scale parameter and $\gamma_{ijk} \in \mathbb{R}$ is the shape parameter. The GEV distribution is a generalization of three classes of distributions: the Weibull, Gumbel, and Fréchet distributions, corresponding to $\gamma_{ijk} < 0$, $\gamma_{ijk} = 0$, and $\gamma_{ijk} > 0$. The three classes of distributions have distinct tail behaviors. Concretely, the upper-end point of the Weibull density function is finite, whereas that of the Gumbel and Fréchet distributions is infinite. Moreover, the density function decays exponentially for the Gumbel distribution and polynomially for the Fréchet distribution. In other words, the Weibull, Gumbel and Fréchet distributions have bounded, light and heavy tails respectively.

The Probability-Weighted Moment (PWM) method [33] is employed here to yield the estimates of GEV parameters $\hat{\mu}_{ijk}$, $\hat{\sigma}_{ijk}$ and $\hat{\gamma}_{ijk}$ locally at each location (i, j) and each direction k. The goal of the PWM method is to match the PWMs $E[x_{ijk}(F(x_{ijk}))^r]$ with the empirical ones b_r, where r is a real number. For the GEV distribution, $E[x_{ijk}(F(x_{ijk}))^r]$ can be written as:

$$\frac{1}{r+1}\left\{\mu_{ijk} - \frac{\sigma_{ijk}}{\gamma_{ijk}}\left[1 + (r-1)^{\gamma_{ijk}}\Gamma(1 - \gamma_{ijk})\right]\right\}, \tag{7}$$

where $\gamma_{ijk} < 1$ and $\gamma_{ijk} \neq 0$, and $\Gamma(\cdot)$ is the gamma function. Here we choose to compute the PWMs for $r = 0$, 1, 2, and the resulting PWM estimates $\hat{\mu}_{ijk}$, $\hat{\sigma}_{ijk}$ and $\hat{\gamma}_{ijk}$ are the solution of the following system of equations:

$$\begin{cases} b_0 & = \mu_{ijk} - \dfrac{\sigma_{ijk}}{\gamma_{ijk}}\left(1 - \Gamma(1 - \gamma_{ijk})\right), \\ 2b_1 - b_0 & = \dfrac{\sigma_{ijk}}{\gamma_{ijk}}\Gamma(1 - \gamma_{ijk})(2^{\gamma_{ijk}} - 1), \\ \dfrac{3b_2 - b_0}{2b_1 - b_0} & = \dfrac{3^{\gamma_{ijk}} - 1}{2^{\gamma_{ijk}} - 1}. \end{cases} \tag{8}$$

In practice, since solving the last equation in (8) is time-consuming, it can be approximated by [33]:

$$\hat{\gamma_i}^{\mathrm{PWM}} = -(7.859c + 2.9554c^2), \tag{9}$$

where $c = (2b_1 - b_0)/(2b_2 - b_0) - \log 2/\log 3$. The PWM method generates good estimates even when the sample size is small, as demonstrated in [33]. Therefore, the method is suitable for extreme-events modeling, especially in our case where the number of samples with the same values of covariates is limited.

However, the linear approximation (9) is accurate only when $|\gamma_{ijk}| < 0.5$. Moreover, the moments in (7) do not exist when $\gamma \geq 1$.

To address these concerns, we utilize the two-stage procedure proposed by Castillo *et al.* [34], which is referred to as the median (MED) method, when the PWM estimates $|\hat{\gamma}_{ijk}| \geq 0.5$ or is not a number (NAN). In the first stage of the MED method, the extreme-value samples can be ordered $x_{ijk}^{(1)} \leq x_{ijk}^{(2)} \leq \cdots \leq x_{ijk}^{(N)}$. A set of GEV estimates is obtained by equating the GEV CDF $F(x_{ijk}^{(n)})$ to its corresponding empirical CDF $P(x_{ijk}^{(n)}) = (n - c_0)/N$, where $c_0 \in (0, 1)$ and we choose $c_0 = 0.35$ as in [34]. More explicitly, for each n such that $2 \leq n \leq N - 1$ in turn, we have:

$$\begin{cases} x_{ijk}^{(1)} = \dfrac{\sigma_{ijk}^{(n)}}{\gamma_{ijk}^{(n)}} \left\{ \left[-\log P\left(x_{ijk}^{(1)}\right) \right]^{-\gamma_{ijk}^{(n)}} - 1 \right\} + \mu_{ijk}^{(n)}, \\ x_{ijk}^{(n)} = \dfrac{\sigma_{ijk}^{(n)}}{\gamma_{ijk}^{(n)}} \left\{ \left[-\log P\left(x_{ijk}^{(n)}\right) \right]^{-\gamma_{ijk}^{(n)}} - 1 \right\} + \mu_{ijk}^{(n)}, \\ x_{ijk}^{(N)} = \dfrac{\sigma_{ijk}^{(n)}}{\gamma_{ijk}^{(n)}} \left\{ \left[-\log P\left(x_{ijk}^{(N)}\right) \right]^{-\gamma_{ijk}^{(n)}} - 1 \right\} + \mu_{ijk}^{(n)}. \end{cases} \tag{10}$$

Note that the equations associated with the first and last sample, i.e., $x_{ijk}^{(1)}$ and $x_{ijk}^{(N)}$, will be used multiple times. By solving the three equations in (10) w.r.t. the GEV parameters independently for each n $(2 \leq n \leq N - 1)$ in turn, we can obtain a set of $(\gamma_{ijk}^{(n)}, \sigma_{ijk}^{(n)}, \mu_{ijk}^{(n)})$ for each of the $N - 2$ occurrences $x_{ijk}^{(n)}$. In the second stage, we determine the median of the sets of GEV parameters as the final MED estimates $(\hat{\mu}_{ijk}, \hat{\sigma}_{ijk}, \hat{\gamma}_{ijk})$. As shown in [34], although the MED method is more computationally demanding than the PWM method, it yields reliable estimates when the shape parameter $|\gamma_{ijk}| \geq 0.5$.

B. Prior Distribution

We assume that each of the three parameter vectors $\mu = [\mu_{ijk}]$, $\gamma = [\gamma_{ijk}]$ and $\sigma = [\sigma_{ijk}]$ has a 3D thin-membrane model (see Fig. 1(d)) as prior. Since the thin-membrane models of μ, γ, and σ share the same structure and inference methods, we present the three models in a unified form. Let z denote the true GEV parameters, that is, z is either μ, σ, or γ.

We next illustrate how to construct the 3D thin-membrane model priors from the constituent chain and circular graphs. As a first step, we build a regular lattice (see Fig. 1(c)) from Markov chains. Since both the longitude and the latitude of the measurements are nonperiodic, either of them can be characterized by the chain graph shown in Fig. 1(a). Let K_{B_x} and K_{B_y} denote the Laplacian matrices corresponding to the graph of one row (i.e., the longitude) and one column (i.e., the latitude) of sites respectively. We further assume that the smoothness across the longitude and the latitude are the same, thus, they can share one common smoothness parameter α_z. The resulting precision matrix of the regular lattice is given by:

$$K_p = (\alpha_z K_{B_x}) \oplus \left(\alpha_z K_{B_y}\right) = \alpha_z \left(K_{B_x} \oplus K_{B_y}\right) = \alpha_z K_L,$$

where $\oplus$ represents the Kronecker sum. It is easy to show that $K_L = K_{B_x} \oplus K_{B_y}$ is the graph Laplacian matrix of the lattice. According to the property of the Kronecker sum [35, Ch. 13], the eigenvalue matrix $\Lambda_L = \Lambda_{B_x} \oplus \Lambda_{B_y}$ and eigenvector matrix $V_L = V_{B_y} \otimes V_{B_x}$, where $\otimes$ denotes the Kronecker product.

In the second step, we accommodate the effect of wind direction. We discretize the wind direction into D sectors and assume that the true GEV parameters are constant in each sector. As GEV parameters of neighboring sectors are similar, the directional dependence can be encoded in a circular graph (see Fig. 1(b)), whose Laplacian matrix is K_C. Another smoothness parameter β_z is introduced to dictate the directional dependence since the smoothness across wind direction and space can be different. We can then seamlessly combine the lattice and the circle, leading to the 3D thin-membrane model (see Fig. 1(d)) with precision matrix:

$$K_{\text{prior}} = (\alpha_z K_L) \oplus (\beta_z K_C) = \alpha_z K_s + \beta_z K_d. \tag{11}$$

Interestingly, $K_s = K_L \otimes I_C$ and $K_d = I_L \otimes K_C$ corresponds to the lattices and circular graphs in the graph respectively, indicating that the former only characterizes the spatial dependence while the latter the directional dependence. Note that I_* is an identity matrix with the same dimension as K_*. Based on the property of the Kronecker sum [35, Ch. 13], the eigenvalue matrix of K_{prior} equals:

$$\Lambda_{\text{prior}} = (\alpha_z \Lambda_L) \oplus (\beta_z \Lambda_C) = \alpha_z \Lambda_s + \beta_z \Lambda_d, \tag{12}$$

where $\Lambda_s = \Lambda_L \otimes I_C$ and $\Lambda_d = I_L \otimes \Lambda_C$. The eigenvector matrix equals:

$$V_{\text{prior}} = V_L \otimes V_C = V_{B_x} \otimes V_{B_y} \otimes V_C. \tag{13}$$

By substituting (11) into (4), the density function of the 3D thin-membrane model becomes:

$$P(z) \propto |K_{\text{prior}}|_+^{0.5} \exp\left\{-\frac{1}{2} z^T K_{\text{prior}} z\right\} \tag{14}$$

$$= |\alpha_z K_s + \beta_z K_d|_+^{0.5} \exp\left\{-\frac{1}{2} Z^T (\alpha_z K_s + \beta_z K_d) Z\right\}. \tag{15}$$

The specific structure of the model is extendable to any number of covariates. Specifically, the precision matrix can be generalized as:

$$\begin{aligned} K_{\text{prior}} &= (\alpha K_a) \oplus (\beta K_b) \oplus (\gamma K_c) \oplus \cdots & (16) \\ &= \alpha K_a \otimes I_b \otimes I_c \otimes \cdots + \beta I_a \otimes K_b \otimes I_c \otimes \cdots \\ &\quad + \gamma I_a \otimes I_b \otimes K_c \otimes \cdots + \cdots, & (17) \end{aligned}$$

where $K_i \in \{K_B, K_C\}$ for $i \in \{a, b, c, \cdots\}$ is the graph Laplacian matrix associated with the dependence structure of covariate i, and where α, β and γ are corresponding smoothness parameters. The eigenvalue and eigenvector matrix of K_{prior} can be computed as:

$$\begin{aligned} \Lambda_{\text{prior}} &= (\alpha \Lambda_a) \oplus (\beta \Lambda_b) \oplus (\gamma \Lambda_c) \oplus \cdots & (18) \\ &= \alpha \Lambda_a \otimes I_b \otimes I_c \otimes \cdots + \beta I_a \otimes \Lambda_b \otimes I_c \otimes \cdots \\ &\quad + \gamma I_a \otimes I_b \otimes \Lambda_c \otimes \cdots + \cdots, & (19) \\ &= \alpha \tilde{\Lambda}_a + \beta \tilde{\Lambda}_b + \gamma \tilde{\Lambda}_c + \cdots, & (20) \\ V_{\text{prior}} &= V_a \otimes V_b \otimes V_c \otimes \cdots, & (21) \end{aligned}$$

where $\Lambda_i \in \{\Lambda_B, \Lambda_C\}$ and $V_i \in \{V_B, V_C\}$ are the eigenvalue and eigenvector matrix of K_i. The dependence structure of nonperiodic and periodic covariates can usually be described by chain and circle graphs respectively, thus it is easy to calculate Λ_i and V_i as discussed in Section II.

C. Posterior Distribution

Let y denote the local estimates of z, where y is either $\hat{\mu}$, $\hat{\sigma}$, or $\hat{\gamma}$, and z denotes the true GEV parameters. We further assume that local estimates for some locations (i, j) and directions k are missing, probably due to the two reasons listed below:

1) There are insufficient observations of extremes in some direction sectors or at some sites. Note that the PWM method needs at least three samples to solve (8). In practice, for data collected by satellites, there are always missing parts in satellite images due to the limited satellite path or the presence of clouds. For data collected by sensors, the sensors may fail during the extreme events, resulting in missing measurements of extreme-value samples. On the other hand, the rate of occurrence of events with respect to direction in particular is non-uniform. For example, for locations sheltered by land, the number of extreme events emanating from the direction of the land will be smaller than from other directions in general.
2) There exist unmonitored sites where no observations are available. For instance, the measuring stations are often irregularly distributed across space, whereas the proposed method is more applicable to the case of regular lattice (see Fig. 1(c)). As a result, we can introduce unmonitored sites such that all the sites, including both observed and unobserved ones, are located on a regular lattice. In addition, in the case of wave heights analysis, people may have particular interest in some unmonitored locations since it is easy and convenient to build offshore facilities there.

We therefore only have measurements (i.e., the local estimates y) at a subset of variables z in the 3D graphical model. Furthermore, the local estimates given by the PWM method are asymptotically Gaussian distributed with respect to the sample size [33], thus motivating us to employ a Gaussian approximation. In other words, we assume that the local estimates are corrupted by Gaussian noise due to the limited number of samples available in each site and directional sector. To summarize, the local estimates are modeled as $y = Cz + b$, where $b \sim N(0, R_z)$ is zero-mean Gaussian random vector (Gaussian white noise) with diagonal covariance matrix R_z, and C is the selection matrix that only selects z at which the noisy observations y are available. C has a single non-zero value (equal to 1) in each row. If there are adequate observations available at all locations and directions, C would simply be an identity matrix. As a consequence, the conditional distribution of the observed value y given the true value z can be written as:

$$P(y|z) \propto \exp\left\{-\frac{1}{2}(y - Cz)^T R_z^{-1}(y - Cz)\right\}. \tag{22}$$

Since we assume that the prior distribution of z is the 3D thin-membrane model (15), the posterior distribution is given by:

$$\begin{aligned} P(z|y) &\propto P(z)P(y|z) &(23)\\ &\propto |K_{\text{post}}|_+^{0.5} \exp\left\{-\frac{1}{2}z^T K_{\text{post}} z + z^T C^T R_z^{-1} y\right\}, &(24)\end{aligned}$$

where $K_{\text{post}} = K_{\text{prior}} + C^T R_z^{-1} C$ is the precision matrix of the posterior distribution.

IV. Learning and Inference

We describe here the proposed learning and inference algorithm. Concretely, we discuss how the smoothed GEV parameters, the noise covariance matrix R_z, and the smoothness parameters α_z and β_z for each of the three parameters μ, γ, and σ are computed.

A. Inferring Smoothed GEV Parameters

Given the covariance matrix R_z and the smoothness parameters α_z and β_z, the maximum *a posteriori* estimate of z is given by:

$$\hat{z} = \arg\max P(z|y) = K_{\text{post}}^{-1} C^T R_z^{-1} y. \tag{25}$$

The dimension M of K_{post} is equal to $M = PQD$. In most practical scenarios, it is intractable to compute the inverse of K_{post} due to the $\mathcal{O}(M^3)$ complexity. In the following, we explain how we can significantly reduce the computational complexity by exploiting the special configuration of the eigenvectors of K_{prior}. When the diagonal matrix $C^T R_z C$ can be well approximated by a scaled identity matrix cI_z, we have:

$$\begin{aligned} K_{\text{post}} &\approx K_{\text{prior}} + cI_z &(26)\\ &= V_{\text{prior}}^T(\alpha_z\Lambda_s + \beta_z\Lambda_d + cI_z)V_{\text{prior}} &(27)\end{aligned}$$

As a result, $\hat{z}$ can be computed as:

$$\hat{z} = V_{\text{prior}}^T(\alpha_z\Lambda_s + \beta_z\Lambda_d + cI_z)^{-1} V_{\text{prior}} C^T R_z^{-1} y. \tag{28}$$

We propose a fast thin-membrane model (FTM) solver to evaluate (28) in three steps.

1) Let $z_0 = C^T R_z^{-1} y$, and the first step computes $z_1 = V_{\text{prior}} z_0$. Note that $V_{\text{prior}} = V_{B_x} \otimes V_{B_y} \otimes V_C$ (21), thus, $z_1 = V_{\text{prior}} z_0$ is a three-dimensional integration transform as defined in [38]. More explicitly, we can first reshape the vector z_0 into a $P \times Q \times D$ array Z_0 such that $z_0 = \text{vec}(Z_0)$, where $\text{vec}(Z_0)$ denotes the vectorization operation. In the spatio-directional model, the three dimensions represent longitude, latitude and direction respectively. Next, we perform the fast cosine transform (FCT) in the first and second dimension (corresponding to V_{B_x} and V_{B_y}) and the fast Fourier transform (FFT) in the third one (corresponding to V_C). The detailed derivation is shown in Appendix A. The resulting computational complexity is $\mathcal{O}(M\log(M))$. Moreover, the computation is amenable to parallelization.
2) In the second step, $(\alpha_z\Lambda_s + \beta_z\Lambda_d + cI_z)$ is a diagonal matrix, so the complexity of the operation $z_2 = (\alpha_z\Lambda_s + \beta_z\Lambda_d + cI_z)^{-1} z_1$ is linear in M.
3) The operation $V_{\text{prior}}^T z_2$ in the final step amounts to performing the inverse the FCT and the FFT in the proper dimensions of the $P \times Q \times D$ array reshaped from z_2, and the computational effort is the same with the first step.

In summary, the computational complexity of evaluating z is $\mathcal{O}(M\log(M))$. A similar algorithm can easily be designed for the general case of multiple covariates. Since the generalized $V_{\mathrm{prior}} = V_a \otimes V_b \otimes V_c \otimes \cdots$ for $V_i \in \{V_B, V_C\}$ $(i \in \{a,b,c,\cdots\})$ (21), we can perform the FFT in the i-th dimension if V_i belongs to the V_B family, and perform the FCT otherwise.

When cI_z is not a good approximation to C^TR_zC, we decompose K_{post} into two parts K_1 and K_2 as follows:

$$K_1 = K_{\mathrm{prior}} + cI_z, \quad K_2 = K_{\mathrm{post}} - K_1,$$

where c is chosen as the largest entry in C^TR_zC. The Richardson iteration [37] is then used to solve $z^{(\kappa+1)} = K_1^{-1}(C^TR_z^{-1}y - K_2z^{(\kappa)})$ until convergence. In each iteration, computing the inverse of K_1 can be circumvented by using the FTM solver mentioned above. The following two theorems guarantee the convergence of the proposed method.

Theorem 1: Given $K_{\mathrm{post}} = K_{\mathrm{prior}} + C^TR_z^{-1}C$, $K_1 = K_{\mathrm{prior}} + cI_z$ and $K_2 = K_{\mathrm{post}} - K_1$, where c equals the largest diagonal element in $C^TR_z^{-1}C$,

1) K_{post} is strictly positive definite,
2) the spectral radius $\rho(K_1^{-1}K_2) < 1$ and the resulting Richardson iterations $z^{(\kappa+1)} = K_1^{-1}(C^TR_z^{-1}y - K_2z^{(\kappa)})$ are guaranteed to converge to $K_{\mathrm{post}}^{-1}C^TR_z^{-1}y$.

Proof: See Appendix B. □

Theorem 2: The convergence rate of the proposed algorithm $\rho(K_1^{-1}K_2)$ is bounded below and above by:

$$\rho\left(K_1^{-1}K_2\right) \geq \frac{\max\left(C^TR_z^{-1}C\right) - \min\left(C^TR_z^{-1}C\right)}{\max(\Lambda_{prior}) + \max\left(C^TR_z^{-1}C\right)}, \quad (29)$$

$$\rho\left(K_1^{-1}K_2\right) \leq \frac{\max\left(C^TR_z^{-1}C\right) - \min\left(C^TR_z^{-1}C\right)}{\min(\Lambda_{prior}) + \max\left(C^TR_z^{-1}C\right)}, \quad (30)$$

where $\max(K)$ and $\min(K)$ denote the maximum and minimum diagonal element of the matrix K respectively.

Proof: See Appendix C. □

According to Theorem 2, decreasing the difference between the largest and smallest diagonal entries of matrix $C^TR_z^{-1}C$ will reduce the upper bound on $\rho(K_1^{-1}K_2)$, resulting in faster convergence. In the limit case where $\min(C^TR_z^{-1}C) = \max(C^TR_z^{-1}C)$, that is, $C^TR_z^{-1}C = cI_z$, the Richardson iteration converges in one step as in (28).

Note that the MAP estimate of z is dependent on the covariance matrix R_z and the smoothness parameters, the calculation of which is described in the sequel.

B. Estimating Covariance Matrices R_z

We use the parametric bootstrap approach to infer the (diagonal) noise covariance matrices R_μ, R_γ and R_σ; this method is suitable when the number of available samples is small [36]. Concretely, we proceed as follows:

1) We generate the local GEV estimates $\hat{\mu}_{ijk}$, $\hat{\sigma}_{ijk}$ and $\hat{\gamma}_{ijk}$ using the method discussed in Section III-A.
2) We draw M sample sets $S_1, \cdots, S_M$, each with N GEV distributed samples based on the local estimates of GEV parameters $(\hat{\mu}_{ijk}, \hat{\sigma}_{ijk}, \hat{\gamma}_{ijk})$. We set $M = 3000$ in our experiments.
3) For each S_m, where $m = 1,\cdots,M$, we estimate the GEV parameters locally again using the method in Section III-A. The resulting GEV estimates are denoted as $(\hat{\mu}_{ijk}^{[m]}, \hat{\sigma}_{ijk}^{[m]}, \hat{\gamma}_{ijk}^{[m]})$.
4) The variance of $\hat{\mu}_{ijk}^{[m]}$ $(m = 1,\cdots,M)$ at site (i,j) and wind direction k is our estimate of the corresponding diagonal element in R_μ. Similarly, we can obtain estimates of the diagonal covariance matrices R_γ and R_σ.

C. Learning Smoothness Parameters

Since α_z and β_z are usually unknown, we need to infer them based on the local estimates y. However, since z is unknown, directly inferring α_z and β_z is impossible, and instead we solve (24) by Expectation Maximization (EM):

$$\left(\hat{\alpha}_z^{(\kappa)}, \hat{\beta}_z^{(\kappa)}\right) = \arg\max Q\left(\alpha_z, \beta_z; \hat{\alpha}_z^{(\kappa-1)}, \hat{\beta}_z^{(\kappa-1)}\right). \quad (31)$$

In Appendix D, we derive the Q-function:

$$\begin{aligned} &Q\left(\alpha_z, \beta_z; \hat{\alpha}_z^{(\kappa-1)}, \hat{\beta}_z^{(\kappa-1)}\right) \\ &\propto -\mathrm{tr}\left(K_{\mathrm{prior}}\left(K_{\mathrm{post}}^{(\kappa-1)}\right)^{-1}\right) \\ &\quad - \left(z^{(\kappa-1)}\right)^T K_{\mathrm{prior}} z^{(\kappa-1)} + \log|K_{\mathrm{prior}}|_+ + c, &(32) \\ &= -c_1\alpha_z - c_2\beta_z + \log|K_{\mathrm{prior}}|_+ + c, &(33) \end{aligned}$$

where

$$c_1 = \mathrm{tr}\left(K_s\left(K_{\mathrm{post}}^{(\kappa-1)}\right)^{-1}\right) + \left(z^{(\kappa-1)}\right)^T K_s z^{(\kappa-1)}, \quad (34)$$

$$c_2 = \mathrm{tr}\left(K_d\left(K_{\mathrm{post}}^{(\kappa-1)}\right)^{-1}\right) + \left(z^{(\kappa-1)}\right)^T K_d z^{(\kappa-1)}, \quad (35)$$

$z^{(\kappa-1)}$ is computed as in (25) with $\hat{\alpha}_z^{(\kappa-1)}$ and $\hat{\beta}_z^{(\kappa-1)}$ obtained from the previous iteration, and c stands for all the unrelated terms. In (32), we need to evaluate $\log|K_{\mathrm{prior}}|_+$. Since K_{piror} can be regarded as a generalized Laplacian matrix corresponding to the connected graph of the 3D thin-membrane model, according to the properties of Laplacian matrices, $|K_{\mathrm{prior}}|_+ = M\det S(K_{\mathrm{prior}})$, where $S(K_{\mathrm{prior}})$ denotes the first $M-1$ rows and columns of the $M \times M$ matrix K_{prior} and $S(K_{\mathrm{prior}})$ is positive definite [39]. As a result,

$$\begin{aligned} \log|K_{\mathrm{prior}}|_+ &= \log\det S(K_{\mathrm{prior}}) + \log M, &(36) \\ &= \log\det\left(\alpha_z S(K_s) + \beta_z S(K_d)\right) + c. &(37) \end{aligned}$$

Taking the partial derivatives of Q function with regard to α_z and β_z, we can obtain:

$$\frac{\partial Q}{\partial \alpha_z} = -c_1 + \mathrm{tr}\left(S(K_{\mathrm{prior}})^{-1}S(K_s)\right), \quad (38)$$

$$\frac{\partial Q}{\partial \beta_z} = -c_2 + \mathrm{tr}\left(S(K_{\mathrm{prior}})^{-1}S(K_d)\right), \quad (39)$$

where the Jacobi's formula is applied, i.e.,

$$\frac{\partial \log\det K}{\partial x} = \mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial x}\right). \quad (40)$$

By equating the partial derivatives (38) and (39) to zero, we have the following two identities:

$$c_1 = \mathrm{tr}\left(S(K_{\mathrm{prior}})^{-1}S(K_s)\right), \tag{41}$$
$$c_2 = \mathrm{tr}\left(S(K_{\mathrm{prior}})^{-1}S(K_d)\right). \tag{42}$$

As a consequence, we obtain:

$$\begin{aligned}\alpha c_1 + \beta c_2 &= \alpha \mathrm{tr}\left(S(K_{\mathrm{prior}})^{-1}S(K_s)\right) \\ &\quad + \beta \mathrm{tr}\left(S(K_{\mathrm{prior}})^{-1}S(K_d)\right) \\ &= \mathrm{tr}\left\{S(K_{\mathrm{prior}})^{-1}(\alpha_z S(K_s) + \beta_z S(K_d))\right\}\end{aligned} \tag{43}$$

Recall that $\alpha_z S(K_s) + \beta_z S(K_d) = S(K_{\mathrm{prior}})$, and therefore,

$$\alpha c_1 + \beta c_2 = M - 1. \tag{44}$$

By substituting (44) into the Q-function (32), it follows:

$$Q\left(\alpha_z, \beta_z; \hat{\alpha}_z^{(\kappa-1)}, \hat{\beta}_z^{(\kappa-1)}\right) = \log |K_{\mathrm{prior}}|_+ - (M-1) + c. \tag{45}$$

At this point, the expression (31) can be succinctly formulated as:

$$\begin{aligned}&\left(\alpha_z^{(\kappa)}, \beta_z^{(\kappa)}\right) = \arg\max \log |K_{\mathrm{prior}}|_+, \\ &\text{s.t.} \quad c_1\alpha_z + c_2\beta_z = M-1, \quad \alpha_z \geq 0, \quad \beta_z \geq 0,\end{aligned} \tag{46}$$

where the constants c_1 and c_2 are computed using (34) and (35) respectively. Since the eigenvalue matrix Λ_{prior} of K_{prior} equals $\Lambda_{\mathrm{prior}} = \alpha_z \Lambda_s + \beta_z \Lambda_d$ as in (20), we can further simplify the objective function (46) as:

$$\begin{aligned}\log |K_{\mathrm{prior}}|_+ &= \log |\Lambda_{\mathrm{prior}}|_+ \\ &= \sum_{k \in \{k : \lambda_{sk} + \lambda_{dk} > 0\}} \log(\alpha_z \lambda_{sk} + \beta_z \lambda_{dk}),\end{aligned} \tag{47}$$

Consequently, the constrained convex optimization problem (46) can be solved efficiently via the bisection method [40].

In the scenario of multiple covariates, (46) can be extended as:

$$\begin{aligned}&\left(\alpha_z^{(\kappa)}, \beta_z^{(\kappa)}, \gamma_z^{(\kappa)}, \cdots\right) \\ &= \arg\max \sum_{\{k : \tilde{\lambda}_{ak} + \tilde{\lambda}_{bk} + \cdots > 0\}} \log(\alpha_z \tilde{\lambda}_{ak} + \beta_z \tilde{\lambda}_{bk} + \gamma_z \tilde{\lambda}_{ck} + \cdots), \\ &\text{s.t.} \quad c_1\alpha_z + c_2\beta_z + c_3\gamma_z + \cdots = M - 1, \\ &\qquad \alpha_z \geq 0, \quad \beta_z \geq 0, \gamma_z \geq 0, \cdots\end{aligned} \tag{48}$$

where $\tilde{\lambda}_{ik} (i \in \{a, b, c, \cdots\})$ is the k-th diagonal element of $\tilde{\lambda}_i$ in (20). The overall learning and inference algorithm for the generalized model is summarized in Table I.

D. Bootstrapping the Uncertainty of GEV Estimates

In addition to the point estimates of GEV parameters, we also have particular interest in the uncertainly of the estimates. As demonstrated in [5], nonparametric bootstrapping provides a reliable tool to quantify the uncertainty associated with extreme-value models. The bootstrap procedure consists of the following steps:

1) Generate a large number M of sample sets $S_1, \cdots, S_m$, each with N occurrences, by resampling at random with replacement from the original N observations. $M = 1000$ in our experiments.
2) For each S_m, where $m = 1, \cdots, M$, apply the algorithm in Table I to yield smoothed GEV parameters $z^{(m)}$, which denotes either γ, σ, or μ.
3) The 95% confidence interval for GEV estimates is computed as the values corresponding to the 2.5% and 97.5% quantiles of $z^{(m)}$ with $m = 1, \cdots, M$.

TABLE I
THE LEARNING AND INFERENCE ALGORITHM FOR MODELING MULTIPLE COVARIATES

1) Estimate the GEV parameters locally using the combination of PWM and MED estimator.
2) Approximate the relation between locally fitting and true GEV parameters by a Gaussian model, that is, $y = Cz + b$. Estimate the diagonal noise covariance R_z using the parametric bootstrap approach.
3) Initialize the smoothness parameters $\hat{\alpha}_z^{(0)}$ and $\hat{\beta}_z^{(0)}$. Iterate the following steps till convergence:
   a) E-step: update the MAP estimates of GEV parameters z using the methods described in Section IV-A:
   $$\hat{z}^{(\kappa-1)} = \left\{K_{\mathrm{post}}^{(\kappa-1)}\right\}^{-1} C^T R_z^{-1} y,$$
   where $K_{\mathrm{post}}^{(\kappa-1)} = K_{\mathrm{prior}}^{(\kappa-1)} + C^T R_z^{-1} C = (\hat{\alpha}_z^{(\kappa-1)} K_a) \oplus (\hat{\beta}_z^{(\kappa-1)} K_b) \oplus (\hat{\gamma}_z^{(\kappa-1)} K_c) \oplus \cdots + C^T R_z^{-1} C$. The last expression follows from (16).
   b) M-step: update the estimate of smoothness parameters by solving (48).

V. NUMERICAL RESULTS

In this section, we test the proposed spatio-directional model (SDM) on both synthetic and real data against 5 other models, that is, a locally fit model (LFM) (i.e., assuming $\alpha_z = 0$ and $\beta_z = 0$), a directionally constant model (DCM) (assuming GEV parameters to be constant across different directions, i.e., $\beta_z = \infty$), a directionally independent model (DIM) (only considering possible spatial dependence and assuming GEV parameters corresponding to different directional sectors are independent, i.e., $\beta_z = 0$), a spatially constant model (SCM) (i.e., $\alpha_z = \infty$), and a spatially independent model (SIM) (i.e., $\alpha_z = 0$).

A. Synthetic Data

Here we draw samples from GEV distributions with parameters that depend on both location and direction. Concretely, we select 256 sites arranged in a two-dimensional 16 × 16 lattice. We then discretize the direction into 15 directional sectors. Next, we predefine GEV parameters for each site and each directional sector. More explicitly, we characterize the spatial dependence and directional dependence by a quadratic polynomial and a Fourier series expansion respectively, as suggested in [10] and [7]. Finally, we randomly generate 50 GEV distributed occurrences for each site and directional sector.

Our results are listed in Fig. 2, Fig. 3 and Table II. Fig. 2 shows the scale parameters estimated by the aforementioned 6 models across space, while Fig. 3(a) shows the estimated shape

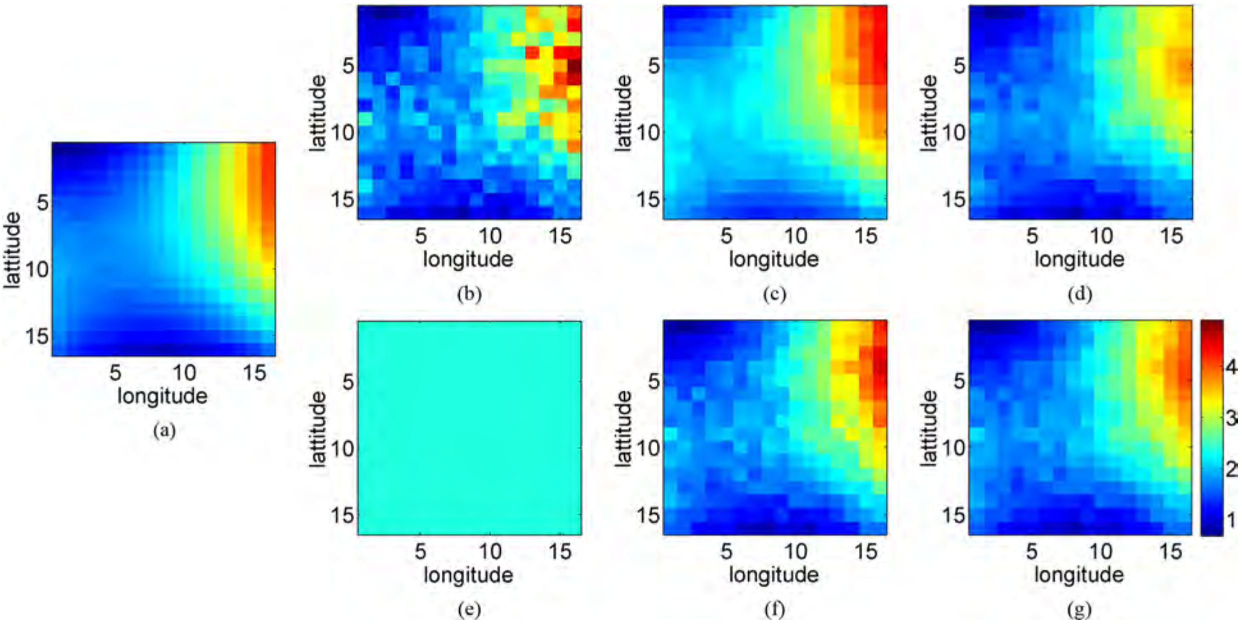


Fig. 2. Estimates of scale parameter σ across all sites in one directional sector. (a) Ground truth. (b) LFM. (c) DCM. (d) DIM. (e) SCM. (f) SIM. (g) SDM.

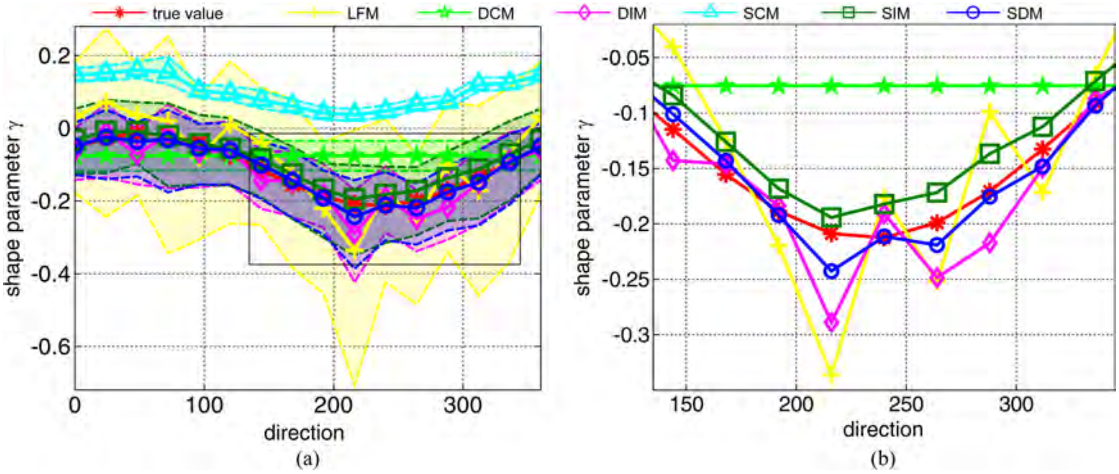


Fig. 3. (a) Estimates of the shape parameters (solid lines) with 95% confidence intervals (dashed lines) across different directions at a randomly selected site; (b) magnified image corresponding to the part in the black square.

parameters with 95% confidence interval across different wind directions. We further magnify the section in the black square in Fig. 3(a) and show it in Fig. 3(b). For other GEV parameters in the spatial or directional domain, the results are qualitatively similar, so we omit them. We can see that estimates resulting from the proposed SDM follow the ground truth closely, across both space and wind direction. On the contrary, local estimates (resulting from the LFM) fluctuate substantially and exhibit the largest uncertainty among the 6 models, since the limited number of occurrences at each location and direction is insufficient to yield reliable estimates of GEV parameters. Next, let us focus on the DCM and the SCM, which ignore one of the two covariates. The DCM is able to infer the varying trend of GEV parameters across space, but generates biased results; specifically, it overestimates the scale parameters (see Fig. 2(c)). Moreover, the DCM mistakenly ignores the directional variation of GEV parameters as shown in Fig. 3(a). Similarly, the SCM can capture the overall trend across different directions (cf. Fig. 3(a)) but fails to model the spatial variation (cf. Fig. 2(e)). Note that the confidence intervals of the DCM and the SCM are narrower than that of the SDM due to the larger sample size by combining the data from different directions (or sites) after ignoring the possible variation. On the other hand, the DIM and the SIM perform much better, yielding similar results to that of the SDM. This implies the necessity of accommodating the heterogeneity in the data. However, since these two models only capture one of the two types of dependence (i.e., directional and spatial dependence) and mistakenly assume the other independent, we can find in Fig. 3(b) that the DIM estimates fluctuate more across direction, and in Fig. 2(f) that the estimates across space given by the SIM are not as smooth as that of the SDM.

Table II summarizes the overall mean square error (MSE) for each of the three GEV parameters and the Akaike Information

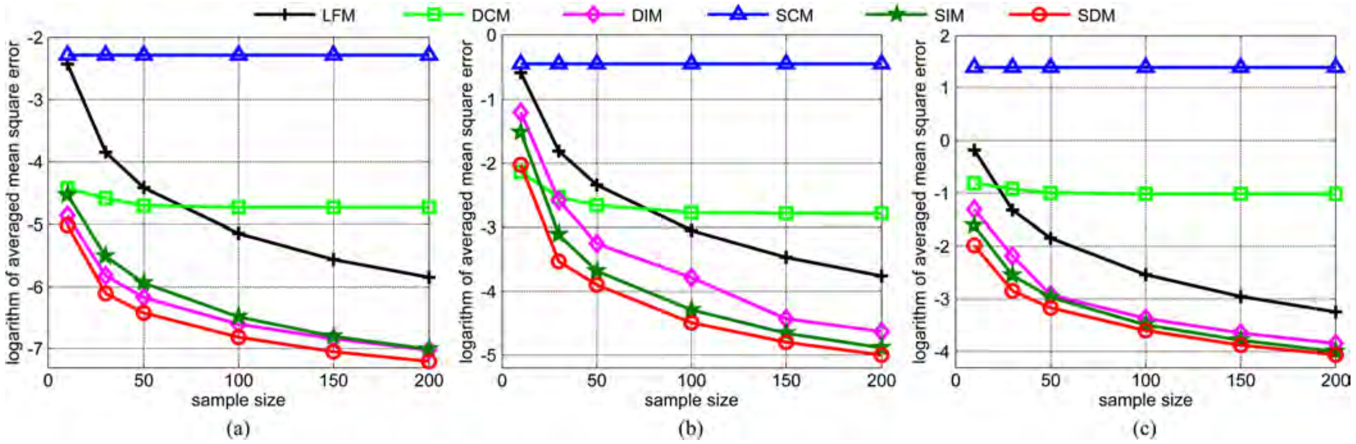


Fig. 4. Mean square error (MSE) of the GEV estimates as a function of sample size (averaged over 100 data sets). (a) Shape parameter γ. (b) Scale parameter σ. (c) Location parameter μ.

TABLE II
Quantitative Comparison of Different Models

Models	Mean Square Error (MSE)			AIC
	γ	σ	μ	
LFM	0.0119	0.1002	0.1539	1.0327×10^6
DCM	0.0090	0.0626	0.3659	9.0111×10^5
DIM	0.0021	0.0276	0.0505	8.7767×10^5
SCM	0.1033	0.6405	3.9919	1.0065×10^6
SIM	0.0026	0.0235	0.0487	8.9400×10^5
SDM	0.0016	0.0192	0.0436	8.7524×10^5

Criterion (AIC) of model fitting. The proposed model yields the smallest MSE and AIC score. The latter implies that the SDM fits the data best notwithstanding the penalty on the effective number of parameters. Note that the effective number of parameters (i.e. the degree of freedom) in the AIC score can be computed as $\mathrm{tr}\{(C^T R_z^{-1} C + \alpha_z K_s + \beta_z K_d)^{-1} C^T R_z^{-1} C\}$ in the proposed spatio-directional model, according to the definition given in [41] and [42]. By comparing the SDM with the other 5 models, we can tell that the proposed model can capture the proper amount of spatial and directional dependence in an automatic manner.

Next, we investigate the impact of sample size on the 6 models. In this set of experiments, we consider the MSE of the GEV estimates for varying sample size. Concretely, we consider sample size 10, 30, 50, 100, 150, and 200 per location per directional bin. For each sample size, we generate 100 data sets with the same parameterization as before. We then compute the MSE of GEV estimates averaged over the 100 sets, as shown in Fig. 4. The proposed SDM usually performs the best. The only exception is for the scale parameter when the number of samples is 10. In this case, the DCM slightly outperforms the SDM, whereas the SDM generates significantly better results in other cases. The proposed model yields reasonably accurate estimates even when the sample size is as small as 10, which demonstrates the utility of the proposed model in the case of small sample size.

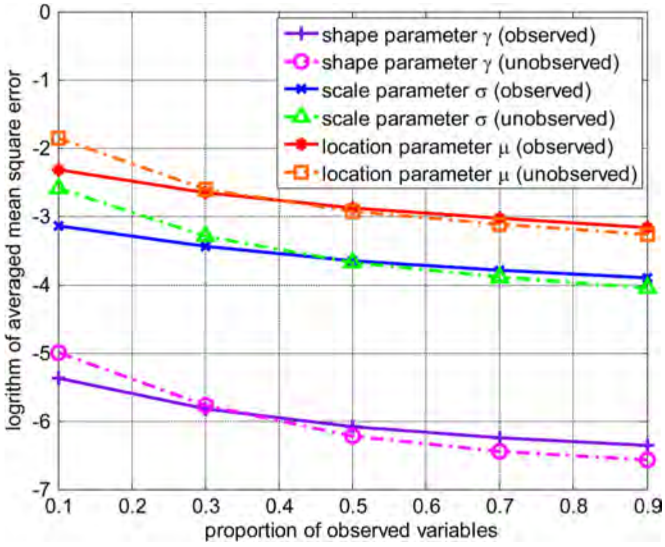


Fig. 5. MSE for varying proportion of observed sites (averaged over 100 trials). The MSE decreases with increasing number of observed sites, as expected.

Finally, we test the performance of the proposed model when dealing with randomly distributed unmonitored sites and wind directions, i.e., sites and directional sectors for which no observations are available. The selection matrix C is not an identity matrix in this case. We apply the EM algorithm to estimate all the GEV parameters, corresponding to both observed and unobserved locations and wind directions. We then depict in Fig. 5 the MSE for each GEV parameter and for the observed and unobserved variables 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 directional sectors, in agreement with our expectation. However, the MSE is still small even when only 10% variables are observed in the 3D graphical models. Therefore, the proposed model can successfully tackle missing data.

B. *Real Data*

We now investigate the extreme wave heights in the Gulf of Mexico. Such analysis could be of benefit when constructing

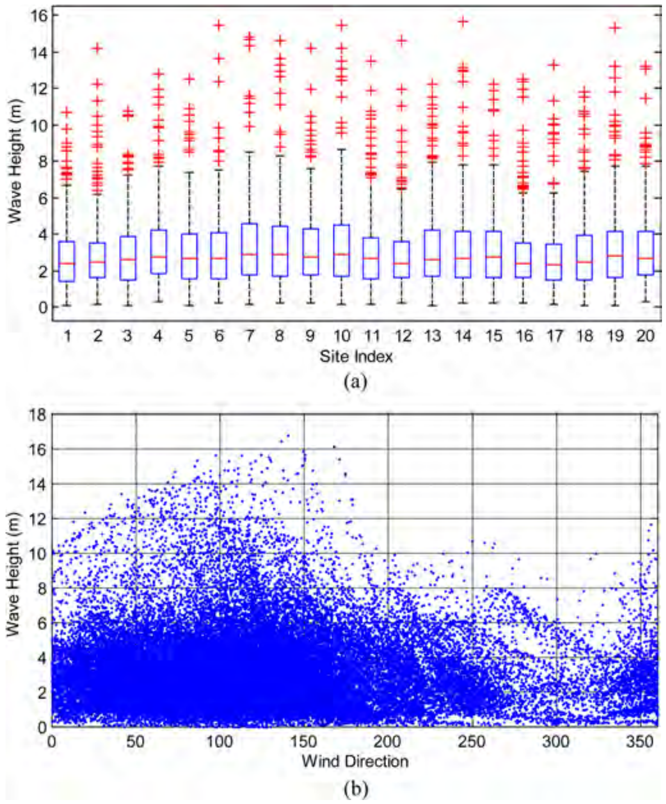


Fig. 6. Heterogeneity in GOMOS data. The wave heights clearly depend on the location in the Gulf of Mexico and the wind direction. (a) Distribution of extreme wave heights at 20 randomly selected sites. (b) Scatter plot of wave height w.r.t direction.

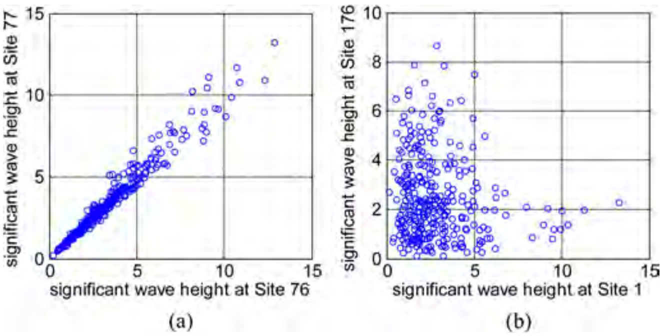


Fig. 7. Scatter plots of wave height at pairs of sites. (a) Two nearby sites. (b) Two distant sites.

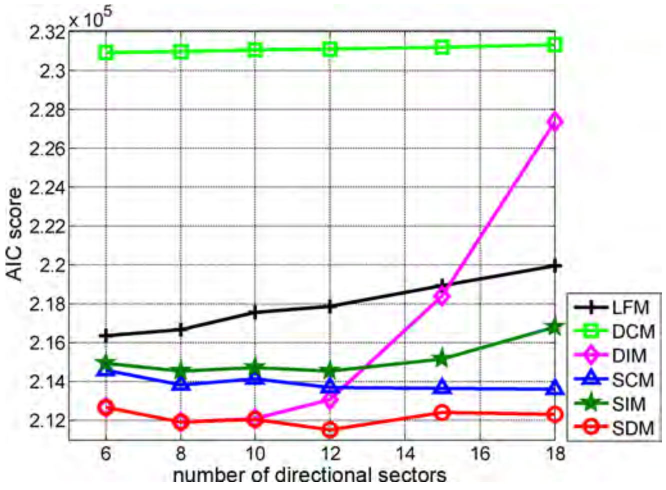


Fig. 8. AIC score as a function of the selected number of directional sectors.

oil platforms in the gulf to withstand extreme waves. The GOMOS (Gulf of Mexico Oceanographic Study) data [43] is used in this study, which cover the period from September 1900 to September 2005 inclusive, at thirty-minute intervals. It measures the significant wave heights. The hindcasts are produced by a physical model, calibrated to observed hurricane data. We isolate 315 maximum peak wave height values; each corresponds to a hurricane event in the Gulf of Mexico. We also extract the corresponding vector mean direction of the wind at the time of the peak significant wave height. We then select 176 sites arranged on a 8 × 22 lattice with spacing 0.5° (approximately 56 km), which almost covers the entire U.S. Gulf of Mexico.

An initial analysis (as shown in Fig. 6) shows the heterogeneity of the data w.r.t. location and direction. We can also see from Fig. 6 that the distribution of storm-wise maxima at each site and direction has a heavy tail, supporting the use of the GEV marginals. We further depict in Fig. 7 the scatter plot of extreme wave heights from two neighboring sites and two distant sites in the lattice. Strong spatial dependence exists between two nearby sites, however, the dependence is clearly weaker between sites that are far apart. This observation provides support for the choice of thin-membrane models, since the latter only capture the direct dependence between neighbors. In summary, the preliminary study suggests that the proposed model is well suited for this data set.

We next apply the 6 models to the data. To test the influence of the selected number of directional sectors on the results, we consider different numbers of sectors, i.e., 6, 8, 10, 12, 15 and 18 in sequence. We then compute the AIC score of the 6 models for each number of directional sectors. The results are summarized in Fig. 8. Again, the SDM always achieves the best AIC score. Moreover, the performance of this model is not sensitive to the chosen number of directional sectors. In practice, we propose to choose the number that minimizes the AIC score, which is 12 for the GOMOS data. In this case, the estimated shape parameters γ_{ijk} resulting from the SDM ranges from -0.30 to 0.34, hence, all the three classes of extreme-value distributions are employed to describe the extreme wave heights. Interestingly, the DIM performs well when the number of directional sectors is small ($\leq$10). However, the performance deteriorates significantly as the number of directional sectors increases. This indicates the need to capture the directional dependence, especially when there is a relatively large number of directional sectors and thus the directional dependence between neighboring sectors is strong. On the other hand, the SIM fails to model the spatial dependence as shown in Fig. 7(a). In contrast with the SIM, the SCM mistakenly ignores the spatial variation, which is essential to model the extreme wave heights in the Gulf of Mexico [10] (see Fig. 6(a) and 7(b)). Similarly, the DCM does not properly consider the strong directional variation visible in Fig. 6(b), and therefore leads to the highest AIC score. Finally, the LFM overfits the data by introducing too many parameters. This can be concluded from the fact that its AIC increases with the number of directional sectors. Taken together, it is evident that the proposed SDM is preferred for this GOMOS data since it models the proper amount of dependence across space and wind direction simultaneously.

VI. Conclusion and Future Work

In this paper, we proposed a novel extreme-value model to accommodate the effects of multiple covariates. More explicitly, we assume that marginal extreme values follow GEV distributions. The GEV parameters are then coupled together through a multidimensional thin-membrane model. The advantages of the proposed model can be summarized as follows: Firstly, the multidimensional thin-membrane model can be constructed flexibly from one-dimensional thin-membrane models, rendering the proposed model generalizable to incorporate any number of covariates. Secondly, component eigenvector structures provide efficient inference of smoothed GEV parameters with computational complexity $\mathcal{O}(M\log(M))$. Thirdly, the eigenvalues of the overall multidimensional model can be computed easily, and help to simplify the determinant maximization problem in the learning process of smoothness parameters. As a result, the proposed method scales gracefully with the problem size. Numerical results for both synthetic and real data support the proposed model; it achieves the most accurate estimates of GEV parameters, and models the data best. Therefore, the approach may prove to be a practical tool for modeling extreme events with covariates.

In the ongoing work, we will capture the dependence between both the GEV parameters and the extreme values. The latter is essential when considering the spatial effects on extreme-value modeling [25]. Additionally, another potential area for future application is in estimation of joint extremes, for instance, the joint distribution of significant wave height and associated spectral peak period [47]. This can be crucial for the design and assessment of offshore and coastal structures.

Appendix A
Derivation of the First Step in the FTM Solver

Due to the mixed product property of the Kronecker product [35, Ch. 13], V_{prior} in (21) can be expressed alternatively as the matrix product:

$$V_{\text{prior}} = \left(I_{B_x}\otimes I_{B_y}\otimes V_C\right)\left(I_{B_x}\otimes V_{B_y}\otimes I_C\right)\times\left(V_{B_x}\otimes I_{B_y}\otimes I_C\right). \tag{49}$$

Therefore, the calculation of $z_1 = V_{\text{prior}}z_0$ can be further divided into three substeps. Before elaborating on the substeps, we first introduce another property of the Kronecker product: for two arbitrary matrices J and K that can be multiplied together,

$$\begin{aligned}\text{vec}(JK) &= (I_K\otimes J)\text{vec}(K) && (50)\\ &= (K^T\otimes I_J)\text{vec}(J). && (51)\end{aligned}$$

Now let us focus on the first substep (as denoted by the superscript), that is, $z_1^1 = (V_{B_x}\otimes I_{B_y}\otimes I_C)z_0$. Given the above mentioned property, z_1^1 can be computed as:

$$z_1^1 = \text{vec}\left(Z_0^1 V_{B_x}^T\right) = \text{vec}\left\{\left(V_{B_x}{Z_0^1}^T\right)^T\right\}, \tag{52}$$

where Z_0^1 is a P by QD matrix such that $\text{vec}(Z_0^1) = z_0$, and P, Q and D are the dimension of V_{B_x}, V_{B_y} and V_C respectively. In addition, recall that we define Z_0 to be a 3D P by Q by D array such that $\text{vec}(Z_0) = z_0$ as in Section IV-A. As a result, it is easy to express ${Z_0^1}^T$ using the entries of Z_0:

$${Z_0^1}^T = \begin{bmatrix}[Z_0]_{111} & [Z_0]_{121} & \cdots & [Z_0]_{1Q1} & \cdots & [Z_0]_{1QD}\\ [Z_0]_{211} & [Z_0]_{221} & \cdots & [Z_0]_{2Q1} & \cdots & [Z_0]_{2QD}\\ \vdots & \vdots & & \vdots & & \vdots\\ [Z_0]_{P11} & [Z_0]_{P21} & \cdots & [Z_0]_{PQ1} & \cdots & [Z_0]_{PQD}\end{bmatrix}.$$

As mentioned in Section II, $V_{B_x}{Z_0^1}^T$ is equivalent to discrete Fourier transform on each column of ${Z_0^1}^T$, and therefore, it coincides with performing the FFT along the first dimension of the 3D array Z_0.

Similarly, in the second substep, we can first apply the property in Expression (50), and subsequently, in Expression (51). As a result, we can find that $z_1^2 = (I_{B_x}\otimes V_{B_y}\otimes I_C)z_1^1$ is identical to the FFT in the second dimension of the 3D P by Q by Q array Z_1^1 which satisfies $\text{vec}(Z_1^1) = z_1^1$. The operation in the third substep, i.e., $z_1 = (I_{B_x}\otimes I_{B_y}\otimes V_C)z_1^2$, can be proven likewise. Note that the ordering of the three integration transforms can be arbitrary, because the three matrices $(I_{B_x}\otimes I_{B_y}\otimes V_C)$, $(I_{B_x}\otimes V_{B_y}\otimes I_C)$, and $(V_{B_x}\otimes I_{B_y}\otimes I_C)$ commute with each other. This algorithm resembles the popular row-column algorithm in the literature of multidimensional Fourier transform, cf. [44].

Appendix B
Proof of Theorem 1

Since K_{prior} is a Laplacian matrix, it is positive semi-definite and singular. In addition, $C^T R_z^{-1} C$ is a diagonal matrix with positive diagonal elements corresponding to the observed variables. According to the properties of Laplacian matrices, the resulting summation K_{post} is strictly positive definite.

The second conclusion follows from the following theorem of standard Richardson iteration, proved by Adams [45].

Theorem 3: Let $K_{\text{post}} = K_1 + K_2$ be a symmetric positive definite matrix and let K_1 be symmetric and nonsingular. Then $\rho(K_1^{-1}K_2) < 1$ if and only if $K_1 - K_2$ is positive definite.

Since $K_1 = K_{\text{prior}} + cI_z$ and $c > 0$, K_1 is symmetric and nonsingular. Note that

$$\begin{aligned}K_2 &= K_{\text{post}} - K_1 && (53)\\ &= C^T R_z^{-1} C - cI_z, && (54)\end{aligned}$$

where c equals the largest diagonal element in $C^T R_z^{-1} C$. Therefore, K_2 is a diagonal matrix with diagonal elements smaller than or equal to 0. As a result, $K_1 - K_2$ is positive definite.

Appendix C
Proof of Theorem 2

Computing the eigenvalues of $K_1^{-1}K_2$ amounts to solving the generalized eigenvalue problem:

$$\lambda K_1 x = K_2 x. \tag{55}$$

It follows from Theorem 2.2 in [46] that

$$\frac{\lambda_{\max}(K_2)}{\lambda_{\max}(K_1)} \le \rho\left(K_1^{-1}K_2\right) \le \frac{\lambda_{\max}(K_2)}{\lambda_{\min}(K_1)}, \tag{56}$$

where $\lambda_{\max}(K)$ and $\lambda_{\min}(K)$ denote the largest and smallest absolute eigenvalues of K. By substituting $K_1 = K_{\rm prior} + cI_z$ and $K_2 = C^T R_z^{-1} C - cI_z$ into (56), we can obtain the bounds specified for the proposed algorithm.

Appendix D
Derivation of the Q-Function

We aim to learn the smoothness parameters α_z and β_z by maximum likelihood estimation, i.e., by maximizing

$$L(\alpha_z, \beta_z) = \log p(y|\alpha_z, \beta_z) \tag{57}$$

$$= \log \int\limits_z p(y, z|\alpha_z, \beta_z) dz. \tag{58}$$

Since maximizing $\log p(y|\alpha_z, \beta_z)$ is intractable, we apply Expectation Maximization (EM) instead.

In the E-step, we compute the Q-function, which is defined as:

$$\begin{aligned}
&Q\left(\alpha_z, \beta_z; \hat{\alpha}_z^{(\kappa-1)}, \hat{\beta}_z^{(\kappa-1)}\right)\\
&= \int\limits_z p\left(z|y, \hat{\alpha}_z^{(\kappa-1)}, \hat{\beta}_z^{(\kappa-1)} \log p(y, z|\alpha_z, \beta_z\right) dz\\
&= E_{z|y,\hat{\alpha}_z^{(\kappa-1)},\hat{\beta}_z^{(\kappa-1)}} \left\{\log p(y, z|\alpha_z, \beta_z)\right\}\\
&= E_{z|y,\hat{\alpha}_z^{(\kappa-1)},\hat{\beta}_z^{(\kappa-1)}} \left\{-\frac{1}{2}\mathrm{tr}(z^T K_{\rm prior} z) + \frac{1}{2}\log|K_{\rm prior}|_+ + c\right\}\\
&= -\frac{1}{2} E_{z|y,\hat{\alpha}_z^{(\kappa-1)},\hat{\beta}_z^{(\kappa-1)}} \left\{\mathrm{tr}\big(K_{\rm prior} z z^T\big)\right\} + \frac{1}{2}\log|K_{\rm prior}|_+\\
&\quad + c.
\end{aligned}$$

Note that the trace and the expectation operator commute. Consequently,

$$\begin{aligned}
&Q\left(\alpha_z, \beta_z; \hat{\alpha}_z^{(\kappa-1)}, \hat{\beta}_z^{(\kappa-1)}\right)\\
&\quad = -\frac{1}{2}\mathrm{tr}\left(K_{\rm prior} E_{z|y,\hat{\alpha}_z^{(\kappa-1)},\hat{\beta}_z^{(\kappa-1)}}(z z^T)\right)\\
&\quad + \frac{1}{2}\log|K_{\rm prior}|_+ + c\\
&\quad = -\frac{1}{2}\mathrm{tr}\left(K_{\rm prior}\left(K_{\rm post}^{(\kappa-1)}\right)^{-1}\right) - \frac{1}{2}\left(z^{(\kappa-1)}\right)^T K_{\rm prior} z^{(\kappa-1)}\\
&\qquad + \frac{1}{2}\log|K_{\rm prior}|_+ + c.
\end{aligned}$$

If we ignore the common coefficient 1/2, we obtain the Q-function in (32).

Acknowledgment

The authors would like to thank Jingjing Cheng, Zheng Choo, and Qiao Zhou for assistance with the Matlab coding.

References

[1] H. Yu, Z. Choo, J. Dauwels, P. Jonathan, and Q. Zhou, "Modeling spatial extreme events using markov random field priors," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2012, pp. 1453–1457.

[2] H. Yu, J. Cheng, and J. Dauwels, "Extreme-value graphical models with multiple covariates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4553–4557.

[3] "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*, M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2007.

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

[5] P. Jonathan and K. Ewans, "Uncertainties in extreme wave height estimates for hurricane-dominated regions," *J. Offshore Mech. Arctic*, vol. 129, no. 4, pp. 300–305, 2007.

[6] P. Jonathan, K. Ewans, and G. Forristall, "Statistical estimation of extreme ocean environments: The requirement for modelling directionality and other covariate effects," *Ocean Eng.*, vol. 35, pp. 1211–1225, 2008.

[7] P. Jonathan and K. C. Ewans, "The effect of directionality on extreme wave design criteria," *Ocean Eng.*, vol. 34, pp. 1977–1994, 2007.

[8] K. C. Ewans and P. Jonathan, "The effect of directionality on Northern North Sea extreme wave design criteria," *J. Offshore Mech. Arctic*, vol. 130, p. 041604, 2008.

[9] P. Jonathan and K. Ewans, "Modelling the seasonality of extreme waves in the Gulf of Mexico," *J. Offshore Mech. Arctic*, vol. 133, p. 021104, 2011.

[10] 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.

[11] H. Sang and A. E. Gelfand, "Hierarchical modeling for extreme values observed over space and time," *Environ. Ecol. Statist.*, vol. 16, pp. 407–426, 2009.

[12] E. F. Eastoe and J. A. Tawn, "Modelling non-stationary extremes with application to surface level ozone," *J. Roy. Stat. Soc. C—Appl.*, vol. 58, pp. 25–45, 2009.

[13] P. Jonathan and K. Ewans, "A spatio-directional model for extreme waves in the Gulf of Mexico," *J. Offshore Mech. Arctic*, vol. 133, p. 011601, 2011.

[14] V. Chavez-Demoulin and A. C. Davison, "Generalized additive modelling of sample extremes," *J. Roy. Statist. Soc. C—Appl.*, vol. 54, pp. 207–222, 2005.

[15] D. Randell, Y. Wu, P. Jonathan, and K. Ewans, "Modelling covariate effects in extremes of storm severity on the Australian North West shelf," in *Proc. Int. Conf. Ocean Offshore Arctic Eng.*, 2013, p. V02AT02A019.

[16] N. P. Galatsanos and A. K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 322–336, 1992.

[17] A. T. Ihler, S. Krishner, M. Ghil, A. W. Robertson, and P. Smyth, "Graphical models for statistical inference and data assimilation," *Physica D*, vol. 230, pp. 72–87, 2007.

[18] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, P. Li, and F. Kschischang, "The factor graph approach to model-based signal processing," *Proc. IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.

[19] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA, USA: Morgan Kauffman, 1988.

[20] 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.

[21] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky, "Embedded trees: Estimation of Gaussian processes on graphs with cycles," *IEEE Trans. Signal Process.*, vol. 52, no. 11, pp. 3136–3150, 2004.

[22] V. Delouille, R. Neelamani, and R. Baraniuk, "Robust distributed estimation using the embedded subgraphs algorithm," *IEEE Trans. Signal Process.*, vol. 54, pp. 2998–3010, 2006.

[23] J. K. Johnson and A. S. Willsky, "A recursive model-reduction method for approximate inference in Gaussian Markov random fields," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 70–83, 2008.

[24] J. K. Johnson, D. M. Malioutov, and A. S. Willsky, "Lagrangian relaxation for MAP estimation in graphical models," in *Proc. 45th Annu. Allerton Conf. Commun., Control, Comput,*, 2007.

[25] H. Yu, Z. Choo, W. I. T. Uy, J. Dauwels, and P. Jonathan, "Modeling extreme events in spatial domain by copula graphical models," in *Proc. 15th Int. Conf. Inf. Fusion*, 2012, pp. 1761–1768.

[26] H. Yu, W. I. T. Uy, and J. Dauwels, "Modeling spatial extremes via ensemble-of-trees of pairwise copulas," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2415–2419.

[27] H. Yu, L. Zhang, and J. Dauwels, "Spatio-temporal graphical models for extreme events," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2014, to be published.

[28] J. M. F. Moura and N. Balram, "Recursive structure of noncausal Gauss Markov random fields," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 334–354, 1992.

[29] 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. Eng.*, vol. 197, pp. 3492–3515, 2008.

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

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

[32] G. Strang, *Computational Science and Engineering*. Wellesley, U.K.: Cambridge Univ. Press, 2007.

[33] J. R. M. Hosking, J. R. Wallis, and E. F. Wood, "Estimation of the generalized extreme-value distribution by the method of probability-weighted moments," *Technometrics*, vol. 27, pp. 251–261, 1985.

[34] E. Castillo and A. S. Hadi, "Parameter and quantile estimation for the generalized extreme-value distribution," *Environmetrics*, vol. 5, pp. 417–432, 1994.

[35] A. J. Laub, *Matrix Analysis for Scientists and Engineers*. Philadelphia, PA, USA: SIAM, 2004.

[36] F. W. Scholz, "The bootstrap small sample properties," Boeing Computer Services, Research and Technology, Tech. Rep., 2007.

[37] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed. Philadelphia, PA, USA: SIAM, 2003.

[38] C. van Loan, *Computational Frameworks for the Fast Fourier Transform*. Philadelphia, PA, USA: SIAM, 1992.

[39] X. D. Zhang, "The Laplacian eigenvalues of graphs: A survey," in *Linear Algebra Research Advances*, Ling G. D., Ed. Hauppage, NY, USA: Nova Science Publishers, 2007, pp. 201–228.

[40] M. T. Heath, *Scientific Computing: An Introductory Survey*. New York, NY, USA: McGraw-Hill, 2002.

[41] T. J. Henderson and R. J. Tibshirani, *Generalized Additive Models*. London, U.K.: Chapman & Hall, 1990.

[42] J. S. Hodges and D. J. Sargent, "Counting degrees of freedom in hierarchical and other richly parameterized models," *Biometrika*, vol. 88, pp. 367–379, 2001.

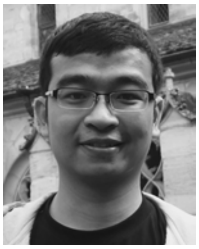
[43] Oceanweather Inc., GOMOS-USA Gulf of Mexico Oceanographic Study, Northen Gulf of Mexico Archive, 2005.

[44] R. Tolimieri, M. An, and C. Lu, *Mathematics of Multidimensional Fourier Transform Algorithms*. New York, NY, USA: Springer, 1993.

[45] L. Adams, " m-step preconditioned conjugate gradient methods," *SIAM J. Sci. Statist. Comput.*, vol. 6, pp. 452–463, 1985.

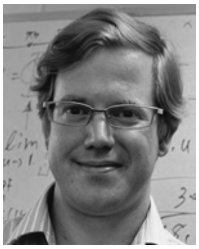
[46] O. Axelsson, "Bounds of eigenvalues of preconditioned matrices," *SIAM J. Matrix Anal. Appl.*, vol. 13, no. 3, pp. 847–862, 1992.

[47] P. Jonathan, K. Ewans, and D. Randell, "Joint modelling of extreme ocean environments incorporating covariate effects," *Coastal Eng.*, vol. 79, pp. 22–31, 2013.



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) obtained the Ph.D. 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). He is an Assistant Professor with School of Electrical and Electronic Engineering at the Nanyang Technological University (NTU) in Singapore. His research interests are in Bayesian statistics, iterative signal processing, and computational neuroscience. His research on intelligent transportation systems has been featured by the BBC, 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.

Philip Jonathan received the B.S. degree in applied mathematics from Swansea University, U.K., in 1984, where he received the Ph.D. degree in chemical physics in 1987.

He joined Shell Research Ltd., U.K., in 1988. He is currently the head of Shell's Statistics and Chemometrics group, and an Honorary Fellow in the Department of Mathematics and Statistics at Lancaster. His current research interests include extreme value analysis for ocean engineering applications, Bayesian methods for monitoring of large systems in time, and inversion methods in remote sensing. He is also interested in multivariate methods generally, particularly in application to physical systems.