Modeling Spatial Extremes via Ensemble-of-Trees of Pairwise Copulas

Hang Yu, *Member, IEEE*, Wayne Isaac T. Uy, and Justin Dauwels, *Senior Member, IEEE*

***Abstract*—Assessing the risk of extreme events in a spatial domain, such as hurricanes, floods, and droughts, presents a unique significance in practice. Unfortunately, the existing extreme-value statistical models are typically not feasible for practical large-scale problems. Graphical models, on the other hand, are capable of handling sizable number of variables, but have yet to be explored in the realm of extreme-value analysis. To bridge the gap, an extreme-value graphical model is introduced in this paper, i.e., an ensemble-of-trees of pairwise copulas (ETPC). In the proposed graphical model, extreme-value marginal distributions are stitched together by means of a pairwise copulas, which in turn are the building blocks of the ensemble of trees. Novel linear-complexity stochastic gradient-based algorithms are then developed for learning the ETPC model and inferring missing data. As a result, the ETPC model is applicable to extreme-value problems with thousands of variables. It can be proven that, under mild conditions, the ETPC model exhibits the favorable property of tail-dependence between an arbitrary pair of sites (variables); consequently, the model is able to reliably capture statistical dependence between extreme values at different sites. Experimental results for both synthetic and real data demonstrate the advantages of the ETPC model in modeling fitting, imputation, and computational efficiency.**

***Index Terms*—Extreme events, pairwise copulas, graphical models, ensemble of trees, stochastic gradient, log-determinant, linear computational complexity, tail dependence.**

I. Introduction

EXTREME events, such as hurricanes, floods, and droughts, often have a major impact on our society. For example, Hurricane Katrina in 2005 was the costliest and deadliest Atlantic hurricane, with a death toll of at least 1,833 and a total loss of 108 billion USD [2]. To assess the likelihood of such events, extreme-value theory has been developed [3], yielding statistical models that can reliably capture extreme events occurring in spatial domain. Unfortunately, the existing extreme-value models are often limited to tens of variables. Yet many practical problems, for instance in Earth Sciences, involve thousands or millions of sites (variables). Graphical models can easily handle such large number of variables [4]–[6], nevertheless, they have rarely been applied in the realm of extreme-event analysis. In this paper, we intend to address this gap by introducing an extreme-value graphical model, i.e., ensemble-of-trees of pairwise copulas (ETPC).

Spatial extremes are often modeled in two stages [7]. First the marginal distributions of the extreme values at each site are learned. Since extreme values are by definition rare, there are only few samples available to infer the parameters of the marginal distributions. To improve the accuracy of the estimated parameter values, the marginal parameters are typically coupled in space. For instance, the parameters are assumed to be Legendre polynomials of the location in [8], however, such parametric models are prone to model misspecification. An alternative is to use nonparametric models to smooth the marginal parameters across space, such as a Gaussian process [9] or a conditional autoregressive model [10]. Unfortunately, these hierarchical models are fitted via Markov Chain Monte Carlo (MCMC) methods and thus it requires a significant amount of computational effort.

In the second stage, the dependence of the extreme values across space is incorporated via copulas or max-stable processes. For example, a Gaussian copula connects the marginal distributions to form a joint distribution in [11]. However, due to the dense covariance matrix in the Gaussian latent layer, the model is computationally intractable for high-dimensional data. In [12], the Gaussian copula is replaced by a copula Gaussian graphical model with a sparse inverse covariance matrix, and consequently, that model is applicable to large-scale problems. The model has a severe limitation though: Gaussian copulas are asymptotically tail independent, and therefore, they are not able to capture the dependence between extremes (e.g., extreme wind gusts at different locations). This disadvantage has sparked interest in the statistics community in models that are asymptotically tail dependent, including max-stable processes and extreme-value (max-stable) copulas [13], [14]. Unfortunately, such models are defined by distribution functions with complicated functional forms (see supplementary material). The corresponding likelihood function involves differentiation with respect to (w.r.t.) all the variables, resulting in a combinatorial explosion. As an example, the likelihood of a 10-dimensional extreme value copula would include a sum over more than 100,000 terms [14]. As a result, when fitting max-stable copulas or processes to extremal data, composite likelihood is often employed [15]. This method replaces the original likelihood by a pseudo-likelihood constructed from pairwise likelihoods. The main drawback, however, is that there is no principled approach to impose interpretable sparse structure on composite likelihood. Consequently, for the sake of model selection, different configurations have to be tested before the best one can be determined. On the other hand, vine copula models [16],

Manuscript received February 20, 2016; revised August 5, 2016; accepted September 2, 2016. Date of publication September 29, 2016; date of current version November 23, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ami Wiesel. This work was supported by MOE (Singapore) Tier 2 project MOE2013-T2-2-015. Part of the material in this paper was presented at International Conference on Acoustics, Speech, and Signal Processing, Florence, Italy, May 2014 [1] and at Fusion, Singapore, July 2012 [12].

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

W. I. T. Uy is with the Center for Applied Mathematics, Cornell University, Ithaca, NY 14850 USA (e-mail: wtu4@cornell.edu).

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

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org provided by the authors. This includes additional material not included in the paper itself. The material is 227 KB in size.

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.2016.2614485

[17] can also be built from pairwise copulas. Specifically, they decompose the joint density into a product of conditional densities and further approximate the latter by products of pairwise copulas [16]. The configuration of such models can be described by a hierarchy of trees whose edges are pairwise copulas. In particular, the structure of regular vine copulas can be determined from the data by selecting all trees to be maximum spanning trees [17]. Moreover, it has been proven in [18] that vine copulas can capture tail dependence under mild conditions. However, vine copulas suffer from the issue of the quadratically increasing number of parameters w.r.t. the dimension. Although different variants of vine copulas are proposed to attenuate this problem [19]–[22], they are still limited to applications with fewer than 100 variables.

In this paper, we propose a class of graphical models for both stages of extreme-value analysis in a spatial domain. As opposed to the aforementioned models, the structure of graphical models are characterized by the conditional independence between variables [4], [5]. For spatial extremes, it is reasonable and straightforward to assume that variables (i.e., extreme events occurring at certain measuring sites in the spatial domain) only have conditional dependence with their neighbors (i.e., extreme events at neighboring sites) [12], [23]. By imposing such a sparse structure, graphical models allow a compact representation of large-scale data as well as efficient learning and inference methods [4], [5].

The proposed model is derived as follows. In the first stage of spatial extremes modeling, we consider the spatial dependence among the parameters of marginal extreme-value distributions by smoothing them using a Gaussian graphical model as prior, specifically a thin-membrane model [12], [23], [24]. The special properties of thin-membrane models enable us to derive efficient algorithms scalable to high-dimensional data [24]. We next proceed to the main focus of this paper, that is, capturing the spatial dependence among extreme values. Motivated by the above discussion, we attempt to construct graphical models with pairwise copulas (especially extreme value copulas) as building blocks. More specifically, we propose to use ensemble-of-trees of pairwise copulas (ETPC) [25], [26]. As a starting point, the sites in spatial domain are arranged on a grid (cf. Fig. 1a). The probability density function (PDF) of the ETPC model is a weighted sum over the PDF of all possible spanning trees on that grid. The PDF of these trees are in turn constructed from pairwise copulas.

Unfortunately, the biggest hurdle in applying the ensemble-of-trees framework presented in [25], [26] to large-scale data is the $\mathcal{O}(NP^3)$ computational complexity of its learning algorithm, where N is the sample size and P is the number of variables. In this paper, we overcome the challenge by designing a novel stochastic gradient based algorithm for learning the model. The computational complexity of the proposed algorithms is only *linear* in the number of variables. Furthermore, we also develop a set of algorithms with *linear* complexity for imputation in the ETPC model. To the best of our knowledge, we are the first to propose scalable algorithms for learning the ETPC model and inferring missing data. Note that the computational cost of the previous works for spatial extremes analysis [11]–[22] is quadratic or higher. As a consequence, the ETPC model can better handle spatial extremal data with thousands of variables.

In addition, it is worthwhile to emphasize that, under the setting of the ETPC model, the extremes can be modeled as either asymptotically tail dependent or independent by choosing the pairwise copulas appropriately. Furthermore, we prove that tail dependence in the ETPC model is preserved if all pairwise copulas in the model are tail dependent. Numerical results for both synthetic data simulated from max-stable processes and extreme rainfall data in Japan suggest that the proposed ETPC model is suitable for spatial extreme-event analysis, in terms of model fitting and extreme value imputation.

This paper is structured as follows. In Section II, we give a brief introduction to undirected graphical models, which is the main tool used in this paper. In Section III, we describe the extreme-value marginal distributions and the method to capture the spatial dependence between marginal parameters. Next, we proceed to the second stage of extreme events modeling in Section IV. Specifically, we explain the ETPC model, which couples the extreme-value marginal distributions through pairwise copulas, arranged in an ensemble of trees. Learning and imputation methods for the ETPC model are then developed respectively in Section V and Section VI. A theoretical guarantee on tail dependence of the ETPC model is presented in Section VII, and numerical results for synthetic and real data are provided in Section VIII. Finally, we offer concluding remarks in Section IX.

II. Undirected Graphical Models

In this section, we first give a brief introduction to undirected graphical models. We then focus on two special types of graphical models, which will be used in the subsequent sections.

An undirected graphical model (i.e., a Markov random field (MRF)) can be defined as a multivariate probability distribution $p(\boldsymbol{x})$ that factorizes according to a graph $\mathcal{G}$ which consists of nodes $\mathcal{V}$ and edges $\mathcal{E}$. More concretely, each node $i \in \mathcal{V}$ is associated with a random variable x_i. An edge $(i,j) \in \mathcal{E}$ is absent if and only if the corresponding two variables x_i and x_j are conditionally independent: $p(x_i, x_j | x_{\mathcal{V}|i,j}) = p(x_i|x_{\mathcal{V}|i,j})p(x_j|x_{\mathcal{V}|i,j})$, where $\mathcal{V}|i,j$ denotes all the nodes except i and j. The Hammersley-Clifford theorem [27] then relates such properties to a factorization of the probability distribution $p(\boldsymbol{x})$ over *cliques* (i.e., fully connected subgraphs), that is,

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C), \tag{1}$$

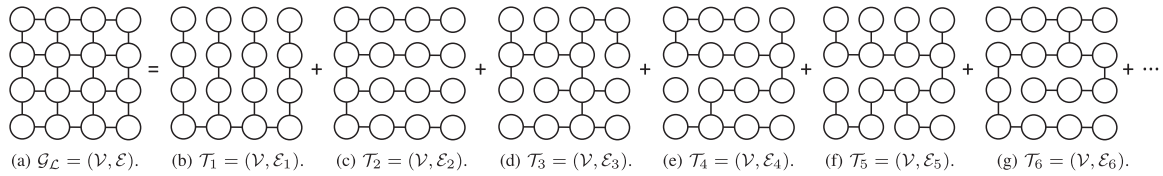
where $\psi_C(\boldsymbol{x}_C)$ is a compatibility function defined on a clique C, $\mathcal{C}$ is the set of all cliques in $\mathcal{G}$, and Z is a normalization term called the partition function. In this paper, we restrict our attention to pairwise Markov random fields in which the cliques are chosen as the nodes and edges in the graph. The resulting PDF can be written as:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{i \in \mathcal{V}} \psi_i(x_i) \prod_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j), \tag{2}$$

where $\psi_i(x_i)$ is a node potential and $\psi_{ij}(x_i, x_j)$ is an edge potential.

If an undirected graph does not include any loops, it is called a tree (i.e., an acyclic graph), cf. Fig. 1b-g. It is known from the junction tree theory [28] that the corresponding joint distribution factorizes as:

$$p(\boldsymbol{x}) = \prod_{i \in \mathcal{V}} p_i(x_i) \prod_{(i,j) \in \mathcal{E}} \frac{p_{ij}(x_i, x_j)}{p_i(x_i) p_j(x_j)}. \tag{3}$$



(a) $\mathcal{G}_{\mathcal{L}} = (\mathcal{V}, \mathcal{E})$. (b) $\mathcal{T}_1 = (\mathcal{V}, \mathcal{E}_1)$. (c) $\mathcal{T}_2 = (\mathcal{V}, \mathcal{E}_2)$. (d) $\mathcal{T}_3 = (\mathcal{V}, \mathcal{E}_3)$. (e) $\mathcal{T}_4 = (\mathcal{V}, \mathcal{E}_4)$. (f) $\mathcal{T}_5 = (\mathcal{V}, \mathcal{E}_5)$. (g) $\mathcal{T}_6 = (\mathcal{V}, \mathcal{E}_6)$.

Fig. 1. An ETPC model: the lattice and several decomposed spanning trees (b)–(g).

In other words, the node marginals $p_i(x_i)$ and the pairwise joint distributions on edges $p_{ij}(x_i, x_j)$ fully describe a tree graphical model. In this case, the node potential is $\psi_i(x_i) = p(x_i)$, the edge potential is $\psi_{ij}(x_i, x_j) = p(x_i, x_j)/[p(x_i)p(x_j)]$, and the partition function is $Z = 1$.

For cyclic graphs (cf. Fig. 1a), it is usually intractable to compute the partition function Z. One exception is the Gaussian graphical model (GGM) in which all the variables are Gaussian distributed, that is, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ and Σ are the mean and covariance matrix respectively. The GGM can be written equivalently as $\mathcal{N}(K^{-1}\boldsymbol{h}, K^{-1})$ with a precision matrix $K = \Sigma^{-1}$ and a potential vector $\boldsymbol{h} = K\boldsymbol{\mu}$. The resulting PDF can be expressed as:

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{P}{2}} |K|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\boldsymbol{h}^T K^{-1}\boldsymbol{h} \Big\} \cdot \exp\Big\{ -\frac{1}{2}\boldsymbol{x}^T K\boldsymbol{x} + \boldsymbol{h}^T\boldsymbol{x} \Big\}. \tag{4}$$

where $(2\pi)^{-\frac{P}{2}} |K|^{\frac{1}{2}} \exp\{-\frac{1}{2}\boldsymbol{h}^T K^{-1}\boldsymbol{h}\}$ is the closed-form partition function, and P is the number of variables. The corresponding node and edge potentials are:

$$\psi_i(x_i) = \exp\Big\{ -\frac{1}{2}K_{ii}x_i^2 + h_i x_i \Big\}, \tag{5}$$

$$\psi_{ij}(x_i, x_j) = \exp\Big\{ -x_i K_{ij} x_j \Big\}. \tag{6}$$

For GGMs, $K_{ij} = 0$ implies that x_i and x_j are conditionally independent.

III. Extreme-Value Marginal Distributions

In this section, we describe how we infer the spatially dependent marginal distribution at each site (i.e., the first stage of spatial extremes modeling). Suppose that we have N observations $x_i^{(n)}$ of block maxima (e.g., annual maxima) at each of P sites, where $i = 1, \cdots, P$ and $n = 1, \cdots, N$. Without loss of generality, we assume that the sites are arranged in a rectangular lattice as shown in Fig. 1a. The proposed model can be easily extended to the case of irregular allocation of measuring sites by specifying the adjacency structure via Delaunay triangulation [12]. According to the Fisher-Tippett-Gnedenko theorem, the block maxima x_i at each location can be well described by the Generalized Extreme Value (GEV) distribution with cumulative distribution function (CDF) [3]:

$$F(x_i) = \begin{cases} \exp\Big\{ -\Big[1 + \dfrac{\xi_i}{\sigma_i}(x_i - \mu_i)\Big]^{-\frac{1}{\xi_i}} \Big\}, & \xi_i \neq 0 \\ \exp\Big\{ -\exp\Big[-\dfrac{1}{\sigma_i}(x_i - \mu_i) \Big] \Big\}, & \xi_i = 0, \end{cases} \tag{7}$$

for $1 + \xi_i/\sigma_i(x_i - \mu_i) \geq 0$ if $\xi_i \neq 0$ and $x_i \in \mathbb{R}$ if $\xi_i = 0$, where $\mu_i \in \mathbb{R}$ is the location parameter, $\sigma_i > 0$ is the scale parameter, and $\xi_i \in \mathbb{R}$ is the shape parameter. It should be noted that the shape parameter governs the tail behavior of the distribution, and therefore the extreme value family can be further divided into three classes: Gumbel, Fréchet and Weibull distributions, corresponding to the case where the shape parameter ξ_i is zero, positive or negative respectively.

A. Local Estimates of GEV Parameters

We first estimate the parameters μ_i, σ_i, and ξ_i locally at each site i by the Probability-Weighted Moment (PWM) method [29], which aims to match the probability-weighted moments $E[x_i(F(x_i))^r]$ with the empirical ones b_r, where r is a real number. For the GEV distribution, $E[x_i(F(x_i))^r]$ can be written as:

$$\frac{1}{r+1}\Big\{ \mu_i - \frac{\sigma_i}{\xi_i}[1 + (r-1)^{\xi_i}\Gamma(1 - \xi_i)] \Big\}, \tag{8}$$

where $\xi_i < 1$ and $\xi_i \neq 0$, and $\Gamma(\cdot)$ is the gamma function. The resulting PWM estimates $\hat{\boldsymbol{\mu}} = (\hat{\mu}_i)$, $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_i)$ and $\hat{\boldsymbol{\xi}} = (\hat{\xi}_i)$ are the solution of the following system of equations:

$$\begin{cases} b_0 = \mu_i - \dfrac{\sigma_i}{\xi_i}(1 - \Gamma(1 - \xi_i)) \\ 2b_1 - b_0 = \dfrac{\sigma_i}{\xi_i}\Gamma(1 - \xi_i)(2^{\xi_i} - 1) \\ \dfrac{3b_2 - b_0}{2b_1 - b_0} = \dfrac{3^{\xi_i} - 1}{2^{\xi_i} - 1}. \end{cases} \tag{9}$$

The PWM method are known to give reliable estimates when the sample size of extreme observations is small [29], [30].

B. Spatial-Dependent Estimates of GEV Parameters

To improve the accuracy of the estimated GEV parameters, we couple those local estimates in space by means of thin-membrane models as in [12], [23], [24]:

$$p(\boldsymbol{z}) \propto \exp\Big\{ -\frac{1}{2}\alpha \sum_{i \in \mathcal{V}} \sum_{j \in N(i)} (z_i - z_j)^2 \Big\} \tag{10}$$

$$\propto \exp\Big\{ -\frac{1}{2}\alpha \boldsymbol{z}^T K_{\text{tm}} \boldsymbol{z} \Big\}, \tag{11}$$

where z_i stands for a GEV parameter (i.e., either ξ_i, σ_i, or μ_i) at site i, $N(i)$ denotes the neighboring nodes of node i, K_{tm} is a graph Laplacian matrix such that $[K_{\text{tm}}]_{i,i}$ is the number of neighbors of site i, while the off-diagonal elements $[K_{\text{tm}}]_{i,j}$ are equal to -1 if nodes i and j are adjacent and 0 otherwise, and α controls the smoothness across space for each GEV parameter. The thin-membrane model is a GGM with precision matrix αK_{tm}.

Next, let $\boldsymbol{y} = (y_1, y_2, \ldots, y_P)$ denote the local estimates of $\boldsymbol{z}$, that is, $\boldsymbol{y}$ is either $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\sigma}}$, or $\hat{\boldsymbol{\xi}}$. Since the PWM estimates follow Gaussian distributions asymptotically w.r.t. the sample size N, we assume that local estimates can be modeled as as $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{b}$, where $\boldsymbol{b} \sim N(\boldsymbol{0}, R_z)$ is a zero-mean Gaussian random vector

with diagonal covariance matrix R_z, that is,

$$p(\boldsymbol{y}|\boldsymbol{z}) \propto \exp\Big\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{z})^T R_z^{-1}(\boldsymbol{y}-\boldsymbol{z})\Big\}. \quad (12)$$

The resulting posterior distribution is given by:

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto \exp\Big\{-\frac{1}{2}\boldsymbol{z}^T(\alpha_z K_{\text{tm}} + R_z^{-1})\boldsymbol{z} + \boldsymbol{z}^T R_z^{-1}\boldsymbol{y}\Big\}. \quad (13)$$

The maximum a posteriori (MAP) estimate of z is given by:

$$\hat{\boldsymbol{z}} = \operatorname{argmax} p(\boldsymbol{z}|\boldsymbol{y}) = (\alpha_z K_{\text{tm}} + R_z^{-1})^{-1} R_z^{-1}\boldsymbol{y}. \quad (14)$$

The noise covariance R_z can be estimated by the bootstrap approach described in [31]. Note that $\hat{\boldsymbol{z}}$ in (14) can be evaluated efficiently with complexity $\mathcal{O}(P \log P)$ as shown in [24].

We adopt an empirical Bayesian approach to estimate both the smooth GEV parameters and the smoothness parameter α_z via expectation maximization (EM) [12]. In the E-step, we compute:

$$\begin{aligned} Q(\alpha_z, \hat{\alpha}_z^{(\kappa)}) &= E_{p(\boldsymbol{z}|\boldsymbol{y},\hat{\alpha}_z^{(\kappa)})}[\log p(\boldsymbol{y},\boldsymbol{z}|\alpha_z)] \\ &= -\frac{1}{2}\alpha_z\Big\{\text{tr}[K_{\text{tm}}(\hat{\alpha}_z^{(\kappa)}K_{\text{tm}} + R_z^{-1})^{-1}] \\ &\quad + \Big(\hat{\boldsymbol{z}}^{(\kappa)}\Big)^T K_p\,\hat{\boldsymbol{z}}^{(\kappa)}\Big\} + \frac{1}{2}\log\det(\alpha_z K_{\text{tm}}), \quad (15) \end{aligned}$$

where $\hat{\boldsymbol{z}}^{(\kappa)}$ is computed as in (14). Since the posterior distribution of $\boldsymbol{z}$ is Gaussian, the MAP estimate is also the mean of the posterior. In the M-step, we select the value $\hat{\alpha}_z^{(\kappa+1)}$ of α_z that maximizes $Q(\alpha_z, \hat{\alpha}_z^{(\kappa)})$. A closed-form expression of $\hat{\alpha}_z^{(\kappa+1)}$ exists:

$$\hat{\alpha}_z^{(\kappa+1)} = \frac{P-1}{\text{tr}[K_{\text{tm}}(\hat{\alpha}_z^{(\kappa)}K_{\text{tm}} + R_z^{-1})^{-1}] + (\hat{\boldsymbol{z}}^{(\kappa)})^T K_p \hat{\boldsymbol{z}}^{(\kappa)}}, \quad (16)$$

where P is the number of sites.

IV. Ensemble-of-Trees of Pairwise Copulas

So far, we have considered marginal distributions, yet our objective is to design multivariate models of extreme values. To this end, we use copula theory. Specifically, we tie the GEV marginal distributions together by means of copulas, i.e., ensemble-of-trees of pairwise copulas (ETPC).

A. Copulas

According to Sklar's Theorem [32], any joint distribution can be expressed as:

$$\begin{aligned} F(x_1,\ldots,x_P) &= C\left(F_1(x_1),\ldots,F_P(x_P)\right) \quad (17)\\ &= C(u_1,\ldots,u_P), \quad (18) \end{aligned}$$

where the function C is defined to be the copula, F_i is the marginal CDF of x_i, and the u_i follows unit uniform distributions. The copula $C(u_1,\cdots,u_P)$ can be uniquely determined for distributions $F(x_1,\ldots,x_P)$ with continuous margins.

Assuming that the partial derivatives exist, the probability density function can be written as:

$$f(x_1,\ldots,x_P) = c(F_1(x_1),\ldots,F_P(x_P))\prod_{i=1}^{P} f_i(x_i), \quad (19)$$

where c is the copula density function [33].

We are primarily concerned with upper tail dependence in this study which is relevant in the analysis of extremes. This can be mathematically expressed as:

$$\begin{aligned} \lambda_U &= \lim_{u\to 1^-} P(U_1 > u | U_2 > u) \quad (20)\\ &= \lim_{u\to 1^-} \frac{1-2u+C(u,u)}{1-u}, \quad (21) \end{aligned}$$

where $\lambda_U \in [0,1]$ is the upper tail dependence coefficient, U_1 and U_2 are uniformly distributed and C is the copula defined on U_1 and U_2. If $\lambda_U > 0$, this indicates that there is dependence in the upper tail, and for $\lambda_U = 0$ there is independence. We consider here eight commonly used copulas: Gaussian, student t, Clayton, Frank, Gumbel, Galambos, t-EV, and Hüsler-Reiss copula. They can be categorized into three major types, namely elliptical (Gaussian and t), Archimedean (Clayton, Frank and Gumbel), and extreme value copulas (Gumbel, Galambos, t-EV, and Hüsler-Reiss). Among them, only Gaussian, Clayton and Frank copulas are not upper-tail dependent. We give a brief introduction of their distribution functions and tail properties in supplementary material, and we refer readers to [33], [34], [35] and references therein for more details. Due to the flexility of copulas, they have found applications in a wide variety of domains, such as sensor fusion [36], GPS navigation [37], image classification [38], texture representation [39], and source localization [40]. Here, we aim to utilize copulas to analyze spatial extremes.

B. From Copulas to Trees and Ensemble-of-Trees

It follows from (3) and (19) that a tree graphical model $\mathcal{T}_i = (\mathcal{V}, \mathcal{E}_i)$ can be written as [26]:

$$f(\boldsymbol{x}|\mathcal{T}_i) = \prod_{j\in\mathcal{V}} f_j(x_j) \prod_{(j,k)\in\mathcal{E}_i} c_{jk}(F_j(x_j), F_k(x_k)), \quad (22)$$

that is, the edge potentials of a tree are the corresponding pairwise copulas. Since trees have limited flexibility in modeling dependence, we construct the Ensemble-of-Trees of Pairwise Copulas (ETPC) model by computing the weighted sum over all possible spanning trees of Fig. 1a as follows:

$$f(\boldsymbol{x}) = \sum_{\mathcal{T}_i} p(\mathcal{T}_i) f(\boldsymbol{x}|\mathcal{T}_i), \quad (23)$$

where $P(\mathcal{T}_i)$ is a decomposable prior proposed in [41] which assigns the weight of each spanning tree $\mathcal{T}_i$ as the product of the weights β_{jk} of all the edges $(j,k) \in \mathcal{E}_i$ in the tree, i.e.,

$$\begin{aligned} p(\mathcal{T}_i) &= \frac{1}{\sum_{\mathcal{T}_i=(\mathcal{V},\mathcal{E}_i)} \prod_{(j,k)\in\mathcal{E}_i} \beta_{jk}} \prod_{(j,k)\in\mathcal{E}_i} \beta_{jk} \quad (24)\\ &= \frac{1}{Z} \prod_{(j,k)\in\mathcal{E}_i} \beta_{jk}. \quad (25) \end{aligned}$$

In order to compute the normalization term Z, we first introduce Kirchhoff's matrix-tree theorem, cf. [42].

Theorem 1 (Matrix-tree theorem): For a graph $\mathcal{G}$, let $\boldsymbol{\beta}$ denote the edge weight matrix, which is symmetric with the diagonal entries being zero and with the off-diagonal entries (j,k) being β_{jk}, and let the Laplacian matrix $L(\boldsymbol{\beta}) = \text{diag}(\boldsymbol{\beta}\mathbf{1}) - \boldsymbol{\beta}$, where $\mathbf{1}$ is a column vector of all ones and $\text{diag}(\boldsymbol{\beta}\mathbf{1})$ is a diagonal matrix with $\boldsymbol{\beta}\mathbf{1}$ on the diagonal. For any two vertices v

and w of $\mathcal{G}$, $Q(\boldsymbol{\beta}) = L(\boldsymbol{\beta})_v^u$ is the matrix obtained by deleting column u and row v in $L(\boldsymbol{\beta})$, then the (weighted) number of spanning trees in $\mathcal{G}$ is given by

$$\omega(\mathcal{G}) = \sum_{\mathcal{T}_i} \prod_{\{j,k\}\in\mathcal{T}_i} \beta_{jk} = \det Q(\boldsymbol{\beta}). \tag{26}$$

According to the matrix-tree theorem, the normalization constant in eq. (25) $Z = \sum_{\mathcal{T}_i} \prod_{(j,k)\in\mathcal{E}_i} \beta_{jk} = \det[Q(\boldsymbol{\beta})]$. For simplicity, we set $Q(\boldsymbol{\beta})$ to be the first $P-1$ rows and columns of the $P \times P$ Laplacian matrix $L(\boldsymbol{\beta})$.

By substituting (22) and (25) into (23), the ETPC model can be succinctly formulated as [25], [26], [41]:

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{\mathcal{T}_i} \frac{1}{Z} \prod_{j\in\mathcal{V}} f_j(x_j) \prod_{(j,k)\in\mathcal{E}_i} \beta_{jk} c_{jk}(F_j(x_j), F_k(x_k)) \\ &= \frac{1}{Z}\left[\prod_{i\in\mathcal{V}} f_i(x_i)\right] \sum_{\mathcal{T}_i}\left[\prod_{(j,k)\in\mathcal{E}_i} \beta_{jk} c_{jk}(F_j(x_j), F_k(x_k))\right] \\ &= \prod_{j\in\mathcal{V}} f_j(x_j) \frac{\det[Q(\boldsymbol{\beta}\odot\boldsymbol{c})]}{\det[Q(\boldsymbol{\beta})]}, \end{aligned} \tag{27}$$

where $\odot$ denotes componentwise multiplication and $\boldsymbol{c}$ is the copula density matrix whose (j,k) entry equals $c_{jk}(F_j(x_j), F_k(x_k))$.

We wish to point out that the tree density in (22) is also a truncated vine copula [19]. However, in a vine copula model, a nested set of another $P-2$ trees is introduced, where the edges corresponds to conditional pairwise copulas, so as to model all higher-order dependencies. The resulting number of parameters grows quadratically with the dimension. Instead, the number of parameters in the ETPC model only increases linearly with the dimension (for a regular lattice), providing a compact representation of high-dimensional data.

V. Learning the ETPC Model

Under the assumption of the lattice graph structure (Fig. 1a), we only need to estimate the parameters of the pairwise copulas $c_{jk}(u_j, u_k)$ corresponding to the pairs of adjacent sites $(j,k) \in \mathcal{E}$. For the sake of efficiency, we apply the Inference for Margins (IFM) method in which the copula parameters are estimated separately from those of the marginal distributions [33]. Specifically, the marginal distribution parameters are estimated as in Section III while the parameters of the pairwise copulas are further estimated via the Maximum Likelihood method. Moreover, we commit to the Bayesian Information Criterion (BIC) when selecting the proper type of copula, as suggested in the literature of copulas [33].

After computing the pairwise copula densities corresponding to the N observations $\boldsymbol{c}^{(n)}$, we proceed to learn $\boldsymbol{\beta}$ by maximizing its likelihood:

$$\begin{aligned} &\hat{\boldsymbol{\beta}} = \operatorname*{argmax}_{\boldsymbol{\beta}} \sum_{n=1}^{N} \log\det[Q(\boldsymbol{\beta}\odot\boldsymbol{c}^{(n)})] - N\log\det[Q(\boldsymbol{\beta})], \\ &\text{s.t.} \quad \beta_{jk} = 0 \quad \forall (j,k)\notin\mathcal{E}, \qquad \beta_{jk} \geq 0 \quad \forall (j,k)\in\mathcal{E}, \\ &\text{and} \quad \|U(\boldsymbol{\beta})\|_2 = 1, \end{aligned} \tag{28}$$

where $\|U(\boldsymbol{\beta})\|_2 = 1$ is the Euclidean norm for the upper triangular part of the matrix $\boldsymbol{\beta}$. The second term in the objective function $N\log\det[Q(\boldsymbol{\beta})]$ serves as a penalty on the weighted number of trees, hence, the model will automatically choose a small number of trees and reduce the number of parameters. Additionally, we fix the scale of $\boldsymbol{\beta}$ because if $\hat{\boldsymbol{\beta}}$ is a solution to the objective function, then $a\hat{\boldsymbol{\beta}}$ with a being a constant is also a solution. Note that we have altered the projection from the unit simplex, as in [25], to the Euclidean norm, as it allows for a larger space of solutions. The projected gradient algorithm [43] is applied to solve the optimization problem in [25]. However, the computational complexity is $\mathcal{O}(NP^3)$, thus, the algorithm is not scalable to data with high dimensions or large sample size.

In order to improve the computational efficiency, we propose a doubly stochastic gradient ascent (DSGA) method to learn the ETPC model. Before describing the method, we first relax the original constrained problem (28) as:

$$\begin{aligned} \hat{\boldsymbol{\eta}} = \operatorname*{argmax}_{\boldsymbol{\eta}} &\log\det[Q(\boldsymbol{\beta}\odot\boldsymbol{c}^{(n)})] - N\log\det[Q(\boldsymbol{\beta})] \\ &- a_1\left(\|U(\boldsymbol{\beta})\|_2^2 - 1\right)^2, \end{aligned} \tag{29}$$

where a_1 is a positive constant, and β_{jk} equals $\exp(\eta_{jk})$ for all $(j,k)\in\mathcal{E}$ and is fixed to 0 otherwise. To maximize the objective function g in (29), we consider the gradients w.r.t. $\boldsymbol{\eta}$:

$$\begin{aligned} \frac{\partial g}{\partial \eta_{jk}} = \Bigg\{ &\boldsymbol{e}_{jk}^T \left[N[Q(\boldsymbol{\beta})]^{-1} - \sum_{n=1}^{N} c_{jk}^{(n)} [Q(\boldsymbol{\beta}\odot\boldsymbol{c}^{(n)})]^{-1} \right] \boldsymbol{e}_{jk} \\ &- 4a_1\left(\|U(\boldsymbol{\beta})\|_2^2 - 1\right)\beta_{jk} \Bigg\} \beta_{jk}, \end{aligned} \tag{30}$$

where e_{jk} is a $(P-1)\times 1$ vector extracting entries in $[N[Q(\boldsymbol{\beta})]^{-1} - \sum_{n=1}^{N} c_{jk}^{(n)}[Q(\boldsymbol{\beta}\odot\boldsymbol{c}^{(n)})]^{-1}]$ that are related to β_{jk}. Due to the summation and the inverse operation in (30), the computational cost of methods based on full gradients is $\mathcal{O}(NP^3)$. In contrast, the proposed DSGA algorithm adopts noisy but unbiased estimates of the true gradients in each iteration; such stochastic gradients can be computed efficiently, leading to computational complexity linear in P.

A. Computation of Stochastic Gradients

As a first step, each iteration estimates the true gradients on the basis of a single randomly selected observation:

$$\begin{aligned} \frac{\partial \tilde{g}}{\partial \eta_{jk}} = \Bigg\{ &N\boldsymbol{e}_{jk}^T \left[[Q(\boldsymbol{\beta})]^{-1} - c_{jk}^{(n)} [Q(\boldsymbol{\beta}\odot\boldsymbol{c}^{(n)})]^{-1} \right] \boldsymbol{e}_{jk} \\ &- 4a_1\left(\|U(\boldsymbol{\beta})\|_2^2 - 1\right)\beta_{jk} \Bigg\} \beta_{jk}, \end{aligned} \tag{31}$$

where the index n is sampled uniformly from the set $\{1,\cdots,N\}$. As a result, the computational complexity is reduced to be $\mathcal{O}(P^3)$.

Furthermore, we can achieve additional computational gain by replacing the matrix inversion in (31) with a low-complexity unbiased estimator. Given a $P\times P$ sparse invertible matrix A, it is difficult to directly compute its inverse $B = A^{-1}$. As an alternative, many efficient linear-complexity methods are available (cf. [44]–[46]) to obtain the ith column of B by solving a linear system $AB_i = d_i$, where d_i is the ith standard basis vector. To calculate all columns of B, we still have to solve P linear systems, i.e., $AB = [d_1,\cdots,d_P] = I$. In this paper, we

significantly reduce the number of linear systems to be solved by finding an unbiased low-rank approximation to I, that is, $E[LL^T] = I$ and L is a random P-by-M matrix where $M \ll P$. More precisely, we first solve $AR = L$ to yield R with $\mathcal{O}(MP)$ complexity and then multiply R by L^T to obtain an unbiased estimate of B [47], [48].

When approximating B by BLL^T, the (i,j)th entry of BLL^T can be decomposed as:

$$[BLL^T]_{ij} = B_{ij} L_{j,:} L_{j,:}^T + \sum_{k \neq j} B_{ik} L_{k,:} L_{j,:}^T, \tag{32}$$

where $L_{k,:}$ is the kth row of L. To approximate B well, we seek to design L such that 1) $L_{j,:} L_{j,:}^T = 1$ and 2) the error term $\sum_{k \neq j} B_{ik} L_{k,:} L_{j,:}^T$ is small. Particularly in our problem, $A = Q(\boldsymbol{\beta})$, and we only need those entries in B that corresponds to non-zeros in A in order to compute the gradient in (31). In other words, we only need to evaluate B_{ij} when i and j are neighbors in the graph corresponding to A. Moreover, without loss of generality, we can assume that B_{ij} decreases with the distance between node i and j. Under this assumption, a locally orthogonal L can successfully shrink the error term for neighboring (i,j): If nodes i and k are close in the graph, $L_{k,:} L_{j,:}^T = 0$ since j and k are also nearby; Otherwise, B_{ik} is small and so is the error term.

One approach to design such L is presented in [47], [48]. Specifically, all nodes in the graph are first partitioned into M color classes using the greedy multicoloring algorithm [48] such that the nodes with the same color have a certain minimum distance D between them. Next, we put a column $L_{:,j}$ in the low-rank matrix for each color j. For each node i of color j, L_{ij} is assigned to be random signs $+1$ or -1 with equal probability, which assists in further reduction of the error term in (32). Other entries are set to be zero. It has been shown in [47] that $L_{i,:} L_{i,:}^T = 1$, $E[L_{i,:} L_{j,:}^T] = 0$, and hence, $E[LL^T] = I$.

Consequently, the stochastic gradient becomes:

$$\frac{\partial \tilde{g}}{\partial \eta_{jk}} = \Big\{ N \boldsymbol{e}_{jk}^T \Big[[Q(\boldsymbol{\beta})]^{-1} LL^T - c_{jk}^{(n)} [Q(\boldsymbol{\beta} \odot \boldsymbol{c}^{(n)})]^{-1} \cdot LL^T \Big] \boldsymbol{e}_{jk} - 4a_1 \big(\|U(\boldsymbol{\beta})\|_2^2 - 1 \big) \beta_{jk} \Big\} \beta_{jk}. \tag{33}$$

In our experiments, we fix $D = 4$ and hence $M = 18$ for a lattice graph of any dimensions. As a result, the computational complexity is reduced to $\mathcal{O}(P)$.

B. Convergence of the DSGA Algorithm

In each iteration, we update all parameters $\boldsymbol{\eta}$ following a gradient ascent approach:

$$\boldsymbol{\eta}^{\{\kappa\}} = \boldsymbol{\eta}^{\{\kappa-1\}} + \rho^{\{\kappa\}} \tilde{\nabla}_{\boldsymbol{\eta}} g|_{\boldsymbol{\eta} = \boldsymbol{\eta}^{\{\kappa-1\}}}, \tag{34}$$

where each entry of $\tilde{\nabla}_{\boldsymbol{\eta}} g|_{\boldsymbol{\eta} = \boldsymbol{\eta}^{\{\kappa-1\}}}$ is computed as in (33). As stated by the theory of stochastic optimization [49], the algorithm is guaranteed to converge to a local maximum when step sizes $\rho^{(\kappa)}$ satisfy the following conditions:

$$\sum_{\kappa} \rho^{\{\kappa\}} = \infty \quad \text{and} \quad \sum_{\kappa} (\rho^{\{\kappa\}})^2 < \infty. \tag{35}$$

In our DSGA algorithm, we initialize $\rho^{\{0\}} = 0.01$ and multiply it by a factor of 0.95 every $3N$ iterations. Additionally, since the objective function (28) is non-convex, the result can be sensitive to the initial estimate. A reliable initial estimate can be obtained by solving a convex upper bound that ignores the first term of (28) [25].

TABLE I
THE LEARNING ALGORITHM OF THE ETPC MODEL

1) Estimate the GEV marginals with spatially dependent parameters using the method proposed in Section III:
   a) Estimate the GEV parameters locally for each site via the PWM method (cf. Eq. (9)).
   b) Smooth the GEV parameters across space using the EM algorithm (cf. Eq. (15) and (16)).
2) Learn the pairwise copulas corresponding to edges in the lattice via maximum likelihood. Choose copula family using BIC. (cf. Subsection IV-A). Estimate the copula density matrix $\boldsymbol{c}^{(n)}$ for all N observations.
3) Substitute $\boldsymbol{c}^{(n)}$ into Eq. (28) and learn the edge weight matrix $\boldsymbol{\beta}$ of the ETPC model using the proposed DSGA algorithm (cf. Subsection V).

C. Variance Reduction of the Stochastic Gradient

When the variance of the stochastic gradient is large, the algorithm might spend a long time oscillating around, resulting in slow convergence and poor performance. Here, we reduce the variance in the DSGA algorithm by applying the stochastic average gradient technique proposed in [50]. In particular, for each sample n, we keep the most recent updates of

$$h^{(n)}(\boldsymbol{\beta}) = [Q(\boldsymbol{\beta})]^{-1} LL^T - c_{jk}^{(n)} [Q(\boldsymbol{\beta} \odot \boldsymbol{c}^{(n)})]^{-1} LL^T. \tag{36}$$

In iteration κ, we randomly pick an index n_κ from $\{1, \cdots, N\}$. We then update $h^{(n_\kappa)}$ accordingly and keep the rest $h^{(n)}$ unchanged. The stochastic average gradient is then computed as:

$$\frac{\partial \bar{g}}{\partial \eta_{jk}} = \left\{ \boldsymbol{e}_{jk}^T \left[\sum_{n=1}^{N} h^{(n)}(\boldsymbol{\beta}) \right] \boldsymbol{e}_{jk} - 4a_1 \big(\|U(\boldsymbol{\beta})\|_2^2 - 1 \big) \beta_{ij} \right\} \beta_{ij}.$$

As demonstrated in [50] both empirically and theoretically, the stochastic average gradient reduces the variance of stochastic gradients, and leads to higher convergence rate. The overall learning algorithm is summarized in Table I.

VI. INFERENCE IN THE ETPC MODEL

After learning the ETPC model, we explain how the model can be used to carry out several inference tasks. Probabilistic inference in graphical models is often concerned with inferring the unobserved variables $\boldsymbol{x}_M$ given observed ones $\boldsymbol{x}_O$. More concretely, one often would like to compute MAP estimate:

$$\hat{\boldsymbol{x}}_M = \operatorname*{argmax}_{\boldsymbol{x}_M} p(\boldsymbol{x}_M | \boldsymbol{x}_O), \tag{37}$$

or obtain the posterior marginals of each unobserved variable:

$$p(x_{Mi} | \boldsymbol{x}_O) = \int_{\boldsymbol{x}_{M|i}} p(\boldsymbol{x}_M | \boldsymbol{x}_O) d\boldsymbol{x}_{M|i}. \tag{38}$$

In this section, we consider both problems under the setting of the ETPC model, and develop efficient algorithms to tackle them. Interestingly, after applying stochastic gradients, the update rules for both problems are quite similar.

A. MAP Inference

In the ETPC model, the MAP estimates of missing values $\boldsymbol{x}_M$ at a set of sites given observed data $\boldsymbol{x}_O$ at other sites can be obtained by solving the following optimization problem:

$$\hat{\boldsymbol{x}}_M = \underset{\boldsymbol{x}_M}{\operatorname{argmax}} \log f(\boldsymbol{x}_M|\boldsymbol{x}_O) \tag{39}$$

$$= \underset{\boldsymbol{x}_M}{\operatorname{argmax}} \sum_{x_j \in \boldsymbol{x}_M} \log f_j(x_j) + \log\det[Q(\boldsymbol{\beta} \odot \boldsymbol{c}(\boldsymbol{x}_M, \boldsymbol{x}_O))], \tag{40}$$

where the notation $\boldsymbol{c}(\boldsymbol{x}_M, \boldsymbol{x}_O)$ signifies that the copula density matrix $\boldsymbol{c}$ is a function of both $\boldsymbol{x}_M$ and $\boldsymbol{x}_O$. In general graphical models (e.g., trees), the problem can be simplified as estimating $\boldsymbol{x}_M$ given the values of the respective adjacent nodes [4]. However, for the ETPC model, such simplification is not possible as indicated by the following theorem.

Theorem 2: Every pair of variables (x_j, x_k) in the ensemble of trees (ET) model corresponding to a connected graph $\mathcal{G}$ are conditionally dependent given other variables in the model.

Proof: This can be proven by contradiction. See supplementary material. ∎

As a result, the only alternative is to solve (40). Unfortunately, the $\mathcal{O}(P^3)$ computational complexity of evaluating the determinant in (40) precludes the use of general optimization algorithms, which are prohibitive for high-dimensional problems. Furthermore, due to the complex functional form of copulas (cf. supplementary material), the objective function often has multiple modes. To remedy these two problems, we propose a stochastic smoothing-based optimization (SSO) algorithm. The computational complexity is linear in P due to the utilization of stochastic gradients. Moreover, the SSO algorithm can at least attain a significant local maximum. In the sequel, we first introduce the smoothing-based optimization algorithm [52], and then elaborate on the proposed SSO algorithm and its application to the ETPC model.

Let $h(\boldsymbol{x})$ be a d-dimensional non-negative function. Its corresponding scale space function can be defined as [51]:

$$H(\boldsymbol{m}, \nu) = \int h(\boldsymbol{x}) \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \nu) d\boldsymbol{x}, \tag{41}$$

where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \nu)$ is a multivariate Gaussian with mean $\boldsymbol{m}$ and covariance $\nu^2 I$. The right hand side of (41) can be viewed as smoothing $h(\boldsymbol{x})$ with a Gaussian blur kernel with zero mean and covariance $\nu^2 I$ and further evaluating the value of the smoothed function at point $\boldsymbol{m}$. Leordeanu *et al.* [52] proposed to find the mode of $h(\boldsymbol{x})$ by updating $\boldsymbol{m}$ and ν in each iteration via coordinate ascent:

$$\boldsymbol{m}^{(\kappa+1)} = \frac{\int \boldsymbol{x} h(\boldsymbol{x}) N(\boldsymbol{x}; \boldsymbol{m}^{(\kappa)}, \nu^{(\kappa)}) d\boldsymbol{x}}{H(\boldsymbol{m}^{(\kappa)}, \nu^{(\kappa)})}, \tag{42}$$

$$\nu^{(\kappa+1)} = \sqrt{\frac{\int P^{-1} \sum_{i=1}^{P} (x_i - m_i^{(\kappa)})^2 h(\boldsymbol{x}) N(\boldsymbol{x}; \boldsymbol{m}^{(\kappa)}, \nu^{(\kappa)}) d\boldsymbol{x}}{H(\boldsymbol{m}^{(\kappa)}, \nu^{(\kappa)})}}, \tag{43}$$

such that $H(\boldsymbol{m}^{(\kappa+1)}, \nu^{(\kappa)}) \geq H(\boldsymbol{m}^{(\kappa)}, \nu^{(\kappa)})$, $H(\boldsymbol{m}^{(\kappa)}, \nu^{(\kappa+1)}) \geq H(\boldsymbol{m}^{(\kappa)}, \nu^{(\kappa)})$. Smoothing-based optimization is a mean shift procedure with adaptively updated kernel variance. The scale space function H gradually approaches h as the standard deviation ν decreases, and recovers h when $\nu = 0$. It is obvious that the maximization of H would localize $\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \nu)$ as a Dirac delta function (i.e., $\nu = 0$) at the maximum of h. Moreover, smoothing h in each iteration with a Gaussian kernel facilitates the convergence to the global maximum, since scale space theory [51] implies that for most functions the local maxima disappear very fast with the increase of the variance of the Gaussian blurring. By initially smoothing h with a large-variance Gaussian kernel and slowly decreasing the variance, the global maximum or at least a significant local maximum can be found [53].

However, we notice that the update rules (42) and (43) are implemented by means of Monte Carlo approximations in [52], thus requiring numerous expensive evaluations of the objective function that is intractable in the ETPC model. Instead of evaluating the objective function, we resort to stochastic gradients and effectively address the problem.

First, let us relax the assumption that the standard deviation ν is the same for all the variables, and the resulting multivariate Gaussian distribution can be represented as $\mathcal{N}(x; \boldsymbol{m}, CC^T)$, where C is a diagonal matrix with the vector $\boldsymbol{\nu}$ on the diagonal and therefore CC^T is a diagonal covariance matrix. As a result, $\boldsymbol{x} = C\boldsymbol{z} + \boldsymbol{m}$, where $\boldsymbol{z}$ follows $\phi(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, I)$ with zero mean and unit variance for all the components. By changing variables in (41) according to $\boldsymbol{z} = C^{-1}(\boldsymbol{x} - \boldsymbol{m})$, we obtain:

$$H(\boldsymbol{m}, \boldsymbol{\nu}) = \int h(C\boldsymbol{z} + \boldsymbol{m}) \phi(\boldsymbol{z}) d\boldsymbol{z}. \tag{44}$$

To maximize $H(\boldsymbol{m}, \boldsymbol{\nu})$, we consider the gradients over $\boldsymbol{m}$ and $\boldsymbol{\nu}$, that is,

$$\nabla_{\boldsymbol{m}} H(\boldsymbol{m}, \boldsymbol{\nu}) = E_{\phi(\boldsymbol{z})}[\nabla_{\boldsymbol{x}} h(\boldsymbol{x})], \tag{45}$$

$$\nabla_{\boldsymbol{\nu}} H(\boldsymbol{m}, \boldsymbol{\nu}) = E_{\phi(\boldsymbol{z})}[\nabla_{\boldsymbol{x}} h(\boldsymbol{x}) \odot \boldsymbol{z}]. \tag{46}$$

In other words, the gradients in (45) and (46) can be expressed as the expectation of exact quantities. As in the previous section, we replace the exact gradients by stochastic gradients when updating $\boldsymbol{m}$ and $\boldsymbol{\nu}$. In particular, we approximate the expectation in (45) and (46) by one realization of $\nabla_{\boldsymbol{x}} h(\boldsymbol{x})$, namely, drawing one sample $\hat{\boldsymbol{z}}$ from $\phi(\boldsymbol{z})$ and evaluating $\nabla_{\boldsymbol{x}} h(\boldsymbol{x})$ at $\hat{\boldsymbol{x}} = \boldsymbol{\nu} \odot \hat{\boldsymbol{z}} + \boldsymbol{m}$. As a result, beginning with a large $\boldsymbol{\nu}$, the SSO algorithm proceeds in each iteration as:

$$\hat{\boldsymbol{z}} \sim \phi(\boldsymbol{z}), \tag{47}$$

$$\hat{\boldsymbol{x}}^{(\kappa)} = \boldsymbol{\nu}^{(\kappa)} \odot \hat{\boldsymbol{z}} + \boldsymbol{m}^{(\kappa)}, \tag{48}$$

$$\boldsymbol{m}^{(\kappa+1)} = \boldsymbol{m}^{(\kappa)} + \rho_\kappa \nabla_{\boldsymbol{x}} h(\hat{\boldsymbol{x}}^{(\kappa)}), \tag{49}$$

$$\boldsymbol{\nu}^{(\kappa+1)} = \boldsymbol{\nu}^{(\kappa)} + \rho_\kappa \nabla_{\boldsymbol{x}} h(\hat{\boldsymbol{x}}^{(\kappa)}) \odot \boldsymbol{z}, \tag{50}$$

where $\nabla_{\boldsymbol{x}} h(\hat{\boldsymbol{x}}^{(\kappa)})$ is the value of $\nabla_{\boldsymbol{x}} h(\boldsymbol{x})$ at $\hat{\boldsymbol{x}}^{(\kappa)}$, and ρ_κ is the step size in iteration κ.

The only challenge of applying the SSO algorithm to the ETPC model lies in efficient computation of the gradient $\nabla_{\boldsymbol{x}_M} h(\boldsymbol{x}_M)$, where $h(\boldsymbol{x}_M)$ denotes the objective function in (40). For each $x_j \in \boldsymbol{x}_M$, the corresponding partial derivative can be computed as:

$$\frac{\partial h(\boldsymbol{x}_M)}{\partial x_j} = \left\{ \sum_{k \in N(j)} \frac{\partial \log\det[Q(\boldsymbol{\beta} \odot \boldsymbol{c})]}{\partial c_{jk}} \frac{\partial c_{jk}}{\partial u_j} \right\} \frac{du_j}{dx_j} + \frac{d \log f_j(x_j)}{dx_j}, \tag{51}$$

TABLE II
SSO FOR MAP INFERENCE IN THE ETPC MODEL

Initialize $\boldsymbol{\nu} \in \mathbb{R}^+$ (e.g. $\boldsymbol{\nu} = 20$). Iterate the following steps until all the components of $\boldsymbol{\nu}$ are sufficiently small.

1) Draw one sample $\hat{\boldsymbol{x}}_M^{(\kappa)}$ from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}^{(\kappa)}, C^{(\kappa)}(C^{(\kappa)})^T)$:
$$\hat{\boldsymbol{z}} \sim \phi(\boldsymbol{z}), \qquad \hat{\boldsymbol{x}}_M^{(\kappa)} = \boldsymbol{\nu}^{(\kappa)} \odot \hat{\boldsymbol{z}} + \boldsymbol{m}^{(\kappa)}.$$
2) Compute the gradient $\nabla_{\boldsymbol{x}_M} h(\boldsymbol{x}_M)$ at $\hat{\boldsymbol{x}}_M^{(\kappa)}$ (cf. (51)–(53)) and update $\boldsymbol{m}$ and $\boldsymbol{\nu}$ as follows:
$$\boldsymbol{m}^{(\kappa+1)} = \boldsymbol{m}^{(\kappa)} + \rho_\kappa \nabla_{\boldsymbol{x}_M} h(\hat{\boldsymbol{x}}_M^{(\kappa)}),$$
$$\boldsymbol{\nu}^{(\kappa+1)} = \boldsymbol{\nu}^{(\kappa)} + \rho_\kappa \nabla_{\boldsymbol{x}_M} h(\hat{\boldsymbol{x}}_M^{(\kappa)}) \odot \boldsymbol{z}.$$

where $c_{jk} = c_{jk}(u_j, u_k)$, $u_j = F_j(x_j)$, and $N(j)$ includes all the neighbors of node j in the lattice. We next analyze each term in (51) in turn. It is easy to compute $du_j/dx_j = f_j(x_j)$ and $d \log f_j(x_j)/dx_j$ in closed-form expressions, so we will not provide further details. In addition, due to the complicated functional form copula densities c_{jk}, we instead evaluate the partial derivative numerically as:

$$\frac{\partial c_{jk}}{\partial u_j} = \frac{c_{jk}(u_j + t, u_k) - c_{jk}(u_j - t, u_k)}{2t}, \tag{52}$$

for $t = 10^{-4}$. The remaining term $\partial \log\det[Q(\boldsymbol{\beta} \odot \boldsymbol{c})]/\partial c_{jk}$ equals:

$$\frac{\partial \log\det[Q(\boldsymbol{\beta} \odot \boldsymbol{c})]}{\partial c_{jk}} = \beta_{jk} \boldsymbol{e}_{jk}^T Q(\boldsymbol{\beta} \odot \boldsymbol{c})^{-1} \boldsymbol{e}_{jk}, \tag{53}$$

where $\boldsymbol{e}_{jk}$ is the $P \times 1$ vector defined in Expression (30). Note that $Q(\boldsymbol{\beta} \odot \boldsymbol{c})^{-1}$ can also be computed stochastically via $Q(\boldsymbol{\beta} \odot \boldsymbol{c})^{-1} LL^T$ as mentioned in the previous section. The resulting SSO algorithm for MAP inference in the ETPC model has computational complexity $\mathcal{O}(P)$. Recall that the complexity of solving (40) via general optimization approaches is $\mathcal{O}(P^3)$, therefore, the proposed inference method yields significant savings in computational complexity. The overall MAP inference algorithm is summarized in Table II.

B. Posterior Marginals Inference

We now turn our attention to the inference of posterior marginals. Since the integration in (38) is intractable, we need to exploit approximate inference methods such as MCMC and variational Bayes (VB) methods. In MCMC methods such as Metropolis-Hastings algorithms, evaluation of the objective function (40) is required, thus, the computational complexity is $\mathcal{O}(P^3)$. Instead, the VB algorithm can be proceeded efficiently with the help of stochastic gradients as discussed below.

Under the VB framework, we aim to find a tractable variational distribution $q(\boldsymbol{x}_M)$ that can best approximate the true posterior $p(\boldsymbol{x}_M|\boldsymbol{x}_O)$ by minimizing the KL divergence between them. This can be equivalently formulated as the maximization of a lower bound on the log marginal likelihood $\log p(\boldsymbol{x}_O)$:

$$\mathcal{L} = \int q(\boldsymbol{x}_M) \log \frac{p(\boldsymbol{x}_M, \boldsymbol{x}_O)}{q(\boldsymbol{x}_M)} d\boldsymbol{x}_M. \tag{54}$$

Since we are only concerned with the posterior marginals, we assume that $q(\boldsymbol{x}_M)$ is a fully factorized Gaussian distribution $\mathcal{N}(\boldsymbol{x}_M; \boldsymbol{m}, CC^T)$, where C is a diagonal matrix with the vector $\boldsymbol{\nu}$ of standard deviations on the diagonal. As a consequence, we can obtain [54]:

$$\begin{aligned}\mathcal{L} &= \int q(\boldsymbol{x}_M) \log p(\boldsymbol{x}_M, \boldsymbol{x}_O) d\boldsymbol{x}_M - \int q(\boldsymbol{x}_M) \log q(\boldsymbol{x}_M) d\boldsymbol{x}_M, \\ &= \int \mathcal{N}(\boldsymbol{x}_M; \boldsymbol{m}, \boldsymbol{\nu}) \log p(\boldsymbol{x}_M, \boldsymbol{x}_O) d\boldsymbol{x}_M + \mathcal{H}(\boldsymbol{m}, \boldsymbol{\nu}),\end{aligned} \tag{55}$$

where $\mathcal{H}(\boldsymbol{m}, \boldsymbol{\nu})$ is the entropy of $q(\boldsymbol{x}_M)$. Interestingly, the first term in $\mathcal{L}$ is identical to the scale space function in (41). Next, we change parameters according to $\boldsymbol{z} = C^{-1}(\boldsymbol{x}_M - \boldsymbol{m})$ as in the previous subsection. Then we can derive the stochastic gradient based update rules as follows [54]:

$$\hat{\boldsymbol{z}} \sim \phi(\boldsymbol{z}), \tag{56}$$

$$\hat{\boldsymbol{x}}_M^{(\kappa)} = \boldsymbol{\nu}^{(\kappa)} \odot \hat{\boldsymbol{z}} + \boldsymbol{m}^{(\kappa)}, \tag{57}$$

$$\boldsymbol{m}^{(\kappa+1)} = \boldsymbol{m}^{(\kappa)} + \rho_\kappa \nabla_{\boldsymbol{x}} h(\hat{\boldsymbol{x}}_M^{(\kappa)}), \tag{58}$$

$$\boldsymbol{\nu}^{(\kappa+1)} = \boldsymbol{\nu}^{(\kappa)} + \rho_\kappa \left[\nabla_{\boldsymbol{x}} h(\hat{\boldsymbol{x}}_M^{(\kappa)}) \odot \boldsymbol{z} + 1 \oslash \boldsymbol{\nu}^{\{k\}} \right]. \tag{59}$$

Compared with the MAP inference, the only difference is the update of the standard deviation $\boldsymbol{\nu}$. By adding the term $1 \oslash \boldsymbol{\nu}$, the variance will not degrade to zero, and thus we can find the approximate posterior marginal distributions.

VII. THEORETICAL RESULTS

In this section, we explore the tail properties of the ETPC model. Theoretically, tail dependence is essential to a multivariate extreme value model, as multivariate extreme value theory and methods can be regarded as the characterization, estimation and extrapolation of the joint tails of multidimensional distributions. Practically, one often considers the probability of extreme events occurring at one site given extreme extremes occurring at other sites. Such type of dependence is well encoded in the tail dependence of a model (cf. Eq. (20)). In the following, we start this section with a key proposition about the tail dependence between two variables that are conditionally independent given a third variable [55]. We then generalize the results to two arbitrary nodes in a tree and further in the ETPC model.

Proposition 1: Suppose that (X_1, X_2, X_3) form a Markov chain where X_1 and X_3 are conditionally independent given X_2 and (X_1, X_2) and (X_2, X_3) have upper tail dependence with coefficients λ_{12} and λ_{23} respectively. Moreover, assume that X_1 and X_3 are stochastically increasing w.r.t. X_2. Then (X_1, X_3) has upper tail dependence with coefficient λ_{13} bounded below by $\lambda_{12}\lambda_{23}$.

Proof: See Appendix A ■

Proposition 1 implies that two conditionally independent variables can be tail dependent, therefore, it sets up the foundation of analyzing tail dependence in an extreme-value graphical model. We next consider the lower bound on the upper tail dependence of two arbitrary variables in a tree.

Proposition 2: Given a tree graphical model $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ with the corresponding set of random variables $\mathbf{X} = (X_i)_{i \in \mathcal{V}}$ whose joint PDF can be written as in (22), the upper tail dependence between all components of $\mathbf{X}$ exists if all the pairwise copulas c_{ij} corresponding to the edges $\mathcal{E}$ in the tree are upper tail dependent and X_i and X_j (variables conditionally dependent in the graph) are stochastically increasing with each other. Specifically, for any two variables X_i and X_j in the graph, the

lower bound of their tail dependence coefficient is the product of the tail dependence coefficients of all the pairwise copulas corresponding to the edges in the path connecting X_i and X_j.

Proof: See Appendix B. ■

Proposition 2 allows us to prove the existence of tail dependence between any two variables in a tree graphical model. Note that similar results have been proven in [18] for vine copulas via an alternative approach. However, they only showed the existence of tail dependence. In addition, we derive here a lower bound of tail dependence coefficients.

Based on the definition of upper tail dependence, it is easy to prove that the tail dependence still exists in the Ensemble-of-Trees model if it exists in one tree graphical model. This is a consequence of the following lemma.

Lemma 1: A finite weighted sum of bivariate copulas, of which at least one is upper tail dependent, preserves upper tail dependence.

Proof: This follows directly from the fact that the limit of a finite sum is the sum of the limit of each addend. ■

More explicitly, for a mixture of trees model,

$$\lambda_{ij} = \lim_{u\to 1^-} P(U_i > u | U_j > u) \tag{60}$$

$$= \sum_i p(\mathcal{T}_i) \lim_{u\to 1^-} P(U_i > u | U_j > u; \mathcal{T}_i) P(\mathcal{T}_i). \tag{61}$$

We now present a closed-form lower bound of the upper tail dependence coefficient of any two nodes in the ETPC model.

Proposition 3: The upper tail dependence between any two nodes in the Ensemble-of-Trees model is bounded below by $\det[Q(\boldsymbol{\beta} \odot \lambda)]/\det[Q(\boldsymbol{\beta})]$ where $Q(X)$ takes the first $P-1$ rows and columns of the Laplacian matrix corresponding to the graph defined by the $P \times P$ edge weight matrix X, λ is the pairwise upper tail dependence coefficient matrix, and $\odot$ denotes componentwise multiplication.

Proof: See Appendix C. ■

Therefore, by properly selecting pairwise copulas, the ETPC model can flexibly describe the dependencies between extreme values, both in a tail dependent and independent manner. In practice, hurricane-induced extreme wave heights are deemed to be tail independent [56], whereas extreme temperature often exhibit tail dependence [57]. Instead of using different models for different type of extreme value data, the ETPC model can handle various extreme value data in a unified framework.

VIII. Experimental Results

In this section, we compare the performance of our ETPC model to two other modeling approaches:

1) Copula Gaussian graphical models (CGGM) with grid structures [12]: a Gaussian copula is utilized to tie together all the GEV marginals, and the sparse grid structure is encoded by the zero pattern of the precision matrix (inverse covariance) of the Gaussian copula. The model can be expressed as follows:

$$\boldsymbol{y} \sim \mathcal{N}(0, K^{-1}), \quad x_i = F_i^{-1}(\Phi(y_i)), \tag{62}$$

where K is the precision matrix, Φ is the CDF of the standard Gaussian distribution, and F_i is the spatially dependent GEV CDF of x_i. For learning K, we first transform GEV distributed x_i to Gaussian distributed y_i according to $y_i = \Phi^{-1}(F_i(x_i))$, and then learn non-zero entries of K through maximum likelihood given $\boldsymbol{y}$ [12]. The corresponding computational complexity is $\mathcal{O}(P^3)$. For conditional inference in CGGMs, we first simulate samples of $\boldsymbol{y}_M$ conditioned on $\boldsymbol{y}_O$ in the Gaussian latent layer using the linear-complexity method proposed in [60]. The Gaussian distributed samples is then transformed to the GEV observed layer.

2) Full regular vine copulas (R-vine) [17]: a joint distribution is first factorized as the product of conditionals:

$$p(x_1, \cdots, x_P) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots \cdot p(x_P|x_1, \cdots, x_{P-1}). \tag{63}$$

Each conditional distribution is further represented by pairwise copulas. For instance,

$$p(x_2|x_1) = c_{12}(F(x_1), F(x_2))p(x_2),$$
$$p(x_3|x_1, x_2) = c_{23|1}(F(x_2|x_1), F(x_3|x_1))p(x_3|x_1),$$

where $F(x_i|x_j)$ is the CDF of the distribution of x_i conditioned on x_j. Hence, the P-dimensional density function can be factorized into a product of $P(P-1)/2$ pairwise (conditional) copulas. A method is proposed in [17] to learn a regular vine copula whose computational complexity is $\mathcal{O}(P^2)$. Note that all pairwise copulas are selected by minimizing the score of Bayesian information criterion (BIC). For inference in vine copulas, since there are no principled methods in the literature, we perform Metropolis-Hasting sampling of $p(\boldsymbol{x}_M|\boldsymbol{x}_O)$, whose computational complexity is $\mathcal{O}(P^2)$.

We also compare two possible configurations under the ETPC model to show the importance of pairwise copula selection: 1) the case where pairwise copula selection is performed through the minimization of the BIC (ETPC model with a mixture of copulas, denoted as "ETPCm") and 2) the case when all pairwise copulas are Gaussian (ETPC model with Gaussian copula, denoted as "ETPCG"). We check the performance of different models in terms of averaged log-likelihood (AvgLogLLH) over all samples, number of parameters (Prm No.), the BIC score, and the computational time.

A. Synthetic Data

We simulate data from a max-stable random field. Theoretically, the behavior of block maxima at multiple sites in a spatial domain can be exactly described by max-stable processes. In particular, we consider a Schlather model with a powered exponential covariance function [58], and simulate 402 extreme-value samples from a 8×8, 16×16, 32×32 and 64×64 lattice respectively. We retain two samples as the testing data corresponding to the 99th and 60th quantiles of the overall extreme values in the spatial domain, while the remaining data points are treated as training data.

First, we show the performance of the proposed method for estimation of marginal distributions. Specifically, we list in Table III the mean squared error (MSE) between the true shape parameters and the estimated ones before and after smoothing.

TABLE III
MSE OF THE SHAPE PARAMETERS ACROSS SPACE BEFORE AND AFTER SMOOTHING

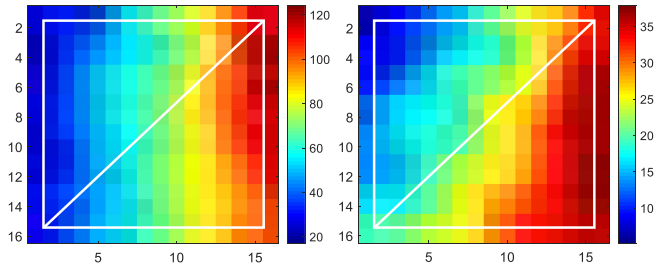
Grid Size	8×8	16×16	32×32	64×64
Local estimates	2.31×10^{-2}	8.02×10^{-4}	7.23×10^{-4}	1.10×10^{-3}
Smoothed estimates	2.28×10^{-2}	5.85×10^{-4}	2.94×10^{-4}	6.55×10^{-4}

TABLE IV
COMPARISON OF DIFFERENT MODELS WHEN FITTING THE SYNTHETIC DATA SIMULATED FROM MAX-STABLE PROCESSES

Grid Size	Models	AvgLogLLH	Prm No.	BIC	Running Time (s)
8×8	ETPCm	1.01×10^2	335	-7.89×10^4	2.87×10^2
	ETPCG	8.41×10^1	224	-6.59×10^4	1.56×10^2
	CGGM	8.67×10^1	176	-6.83×10^4	7.28×10^{-1}
	R-vine	1.13×10^2	2286	-7.71×10^4	2.32×10^3
16×16	ETPCm	3.85×10^2	1439	-3.00×10^5	8.51×10^2
	ETPCG	3.190×10^2	960	-2.495×10^5	3.07×10^2
	CGGM	3.186×10^2	736	-2.504×10^5	7.92
	R-vine	4.87×10^2	33895	-1.87×10^5	9.18×10^3
32×32	ETPCm	1.40×10^3	5950	-1.09×10^6	4.02×10^3
	ETPCG	1.141×10^3	3967	-8.89×10^5	1.22×10^3
	CGGM	1.147×10^3	3008	-8.99×10^5	5.50×10^3
	R-vine	2.75×10^3	529568	9.75×10^5	6.71×10^5
64×64	ETPCm	5.52×10^3	24190	-4.27×10^6	1.54×10^4
	ETPCG	4.54×10^3	16127	-3.53×10^6	5.13×10^3
	CGGM	4.60×10^3	12160	-3.60×10^6	2.54×10^6
	R-vine	2.11×10^4	8411613	3.35×10^7	1.53×10^7

TABLE V
COPULA TYPES SELECTED THROUGH MINIMIZATION OF THE BIC IN THE ETPCM FOR SYNTHETIC DATA

Grid size	Gaussian	t	Gumbel	Clayton	Frank	Galambos	t-EV	Hüsler-Reiss
8×8	0	110	0	0	0	0	1	1
16×16	0	0	0	0	0	0	480	0
32×32	0	0	1	0	0	0	1983	0
64×64	0	6	1	0	0	0	8057	0



(a) 99th quantile. (b) 60th quantile

Fig. 2. Testing data for imputation.

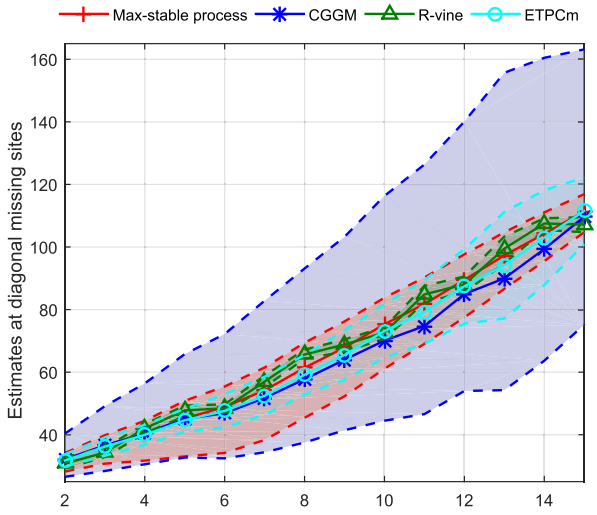
TABLE VI
MAE BETWEEN THE MEDIANS RESULTING FROM DIFFERENT MODELS AND FROM THE MAX-STABLE PROCESS

Testing Sample	ETPCm	R-vine	CGGM
99th quantile	1.29	2.23	2.00
60th quantile	0.80	1.26	0.60

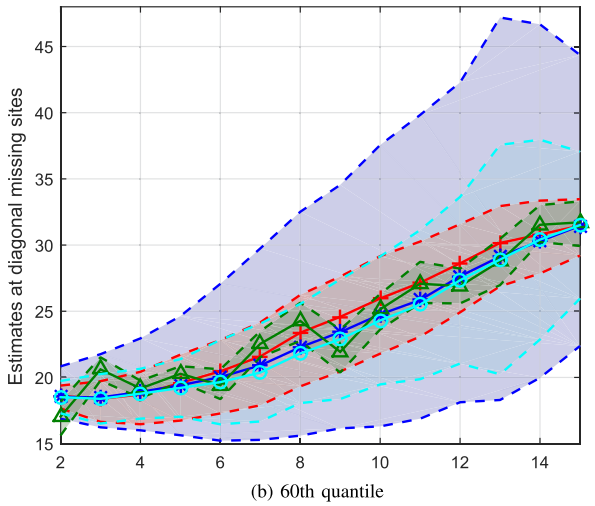
The results of other GEV parameters are similar, so we omit them for brevity. We can see from the table that the smoothed estimates are always better than the local ones, especially for high-dimensional cases. By learning the smoothness parameters from the data, the proposed approach can capture proper amount of spatial dependence between GEV parameters in an automatic manner.

We next compare the aforementioned four models in terms of fitting the data, and summarize the results in Table IV. As demonstrated in the table, the R-vine always achieves the largest average log-likelihood at the expense of employing the greatest number of parameters. In the case of low dimensions (64 variables), its BIC score is the lowest. However, the BIC score becomes the highest as the dimension increases, indicating that the R-vine overfits the data by introducing too many parameters. Moreover, the high computational cost of the R-vine also poses a formidable challenge: it becomes unacceptably slow with the increase of dimension. Therefore, we conclude that R-vine is not suitable for modeling high dimensional data. On the other hand, the CGGM requires fewest parameters. Unfortunately, although its parameter learning algorithm is fast when there are fewer than 1000 variables, the speed decreases dramatically as the number of variables increases, due to the $\mathcal{O}(P^3)$ computational complexity. Furthermore, as mentioned in Section I, Gaussian copulas are unable to accurately capture the dependence between extreme-value data, since these copulas are tail independent. Consequently, the average log-likelihood of this model is smaller in comparison with models based on tail-dependent copulas. The same phenomenon is also observed for ETPCG. Therefore, Gaussian copula based models are not recommended for dealing with extreme-value data. As opposed to the CGGM and the ETPCG, the ETPCm does not suffer from the problem of Gaussian copulas. As shown in Table V, all pairwise copulas selected using BIC have tail dependence. Indeed, the proposed ETPCm model successfully captures the interdependencies among the data with few parameters, leading to the smallest BIC score in high dimensional scenarios. By comparing ETPCm with ETPCG, we can observe that it is essential to select proper families of copulas for a specific type of data, since otherwise the result can be misleading. Additionally, we can observe that the learning algorithm of the ETPC models scales gracefully with the dimension: the computation time is an approximate linear function w.r.t. the dimension, and it is the shortest in the case of 4096 variables. Note that most of the time is spent on selection and estimation of pairwise copulas. In summary, we conclude that the ETPCm provides superior modeling capabilities over other models for large-scale spatial extremes analysis, since it is able to handle the tail dependence between extremes and the computational complexity is only $\mathcal{O}(P)$.

Finally, we present the results of the algorithm for posterior marginals inference in Section VI. Specifically, we consider the 16×16 lattice. For each testing sample, we consider the case in which the data is missing in a 14×14 area at the middle of the 16×16 grid, as shown by the white squares in Fig. 2. Here, we only present the results of the posterior marginals inference algorithm, since the MAP inference results of our model is similar to the median of the posterior marginals, and more importantly, we can better compare different models with the information of uncertainty. Furthermore, we conduct conditional simulation of $p(\boldsymbol{x}_M|\boldsymbol{x}_O)$ in the max-stable random field using the method proposed in [59], and treat the distribution as the ground truth. We depict in Fig. 3 the median, the 2.5% and 97.5% quantiles of the distributions at the locations along the diagonal of the missing region, and compute in Table VI the mean absolute error (MAE) between the estimated medians given by the original max-stable process and other models. We also show the computational time in Table VII. It can be seen from the figures that all three benchmark models can describe the increasing trend along



(a) 99th quantile.



(b) 60th quantile

Fig. 3. Posterior marginals of missing sites resulting from different models.

TABLE VII
COMPUTATIONAL TIME OF POSTERIOR MARGINALS INFERENCE IN DIFFERENT MODELS

Testing Sample	ETPCm	R-vine	CGGM
99th quantile	5.98×10^3	8.85×10^6	2.76×10^2
60th quantile	7.06×10^3	1.02×10^7	2.81×10^2

the diagonal. However, compared with the max-stable process, R-vine yields bumpy estimates with much smaller variances. Indeed, the MAE between the medians of the R-vine and the max-stable process is the largest for both low and high quantiles. This further implies that the R-vine may overfit the training data and therefore fail to express the behavior of testing data. Another shortcoming of the R-vine is that the computational complexity of the inference algorithm is $\mathcal{O}(P^2)$. Its inference algorithm is already (above) three orders of magnitude slower than that of other models in the case of a 16×16 grid, and thus, the R-vine is prohibitive for large-scale imputation. On the other hand, inference in CGGMs can be carried out efficiently. The computational time is the shortest when there are 256 variables, and the algorithm scales linearly with the dimension. Nonetheless, although the model captures the spatial dependence well when the quantile of the sample is low (i.e., 66th quantile), its MAE is larger than that of the ETPCm in the 99th-quantile case. In addition, the variance of the posterior marginals in CGGMs is much larger than that in other models, and increases more significantly with the quantile. According to the copula theory, Gaussian copulas are asymptotically independent with growth of the quantile. As a result, the imputation values have a larger uncertainty in the case of high quantiles. Different from the R-vine and the CGGM, the proposed ETPCm generates posterior marginals that are the nearest to that from the max-stable process. In particular, the MAE is the smallest in the cases of both 99th and 60th quantiles. Furthermore, the computational complexity is only $\mathcal{O}(P)$, hence it is well suited for large-scale scenarios.

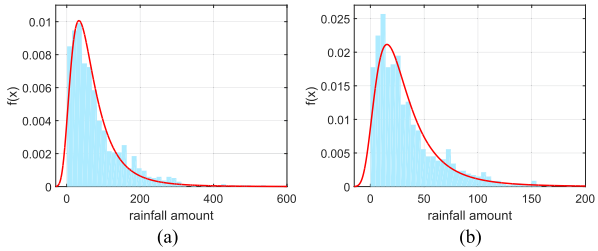


Fig. 4. Normalized histogram (light blue bars) and estimated GEV distribution (red lines) for two randomly selected sites.

B. Real Data

To assess our model in real-world scenarios, we consider the extreme precipitation in the Japanese archipelago. The daily rainfall data from 1900–2011 is compiled and interpolated onto a grid with resolution 0.05° [61]. We select a 18×38 regular grids in South Japan, where heavy rainfall is often the cause of floods. We extract the extreme precipitation values from the sites in the regions as follows: We compute the overall daily rainfall amount in that region. We next choose a high enough threshold and retain those precipitation events (possibly lasting for several days) with rainfall amount above a threshold, resulting in fewer than 10 events per year. For each rainfall event in the region, we isolate the maximum daily rainfall amount for each site. Finally, we obtain 999 samples for each site in the region. We observe that the resulting histogram of each site resembles the GEV distribution (cf. Fig. 4).

Moreover, we also verify empirically whether pairwise upper tail dependence exists in the data using the method proposed in [62]. If x_i, x_j denote extreme precipitation at two different sites, we transform them to be Gaussian distributed via $y_i = \Phi^{-1}(F_i(x_i))$ and $y_j = \Phi^{-1}(F_j(x_j))$ where Φ is the standard normal CDF and F_i, F_j are the marginal GEV distributions for x_i, x_j. The plots for y_i and y_j are shown in Fig. 5 for two arbitrarily selected pairs of sites in the 16×16 grid. The pointed contour lines at the upper right corner suggest that the pairs of sites manifest upper tail dependence. Thus, we expect that the upper tail dependent copulas are the most suitable to model such type of dependence. Additionally, we compute the value of Kendall's tau between pairs of neighboring sites. Due to the small distance between neighboring sites, the value of Kendall's tau is large (ranging from 0.92 to 0.98).

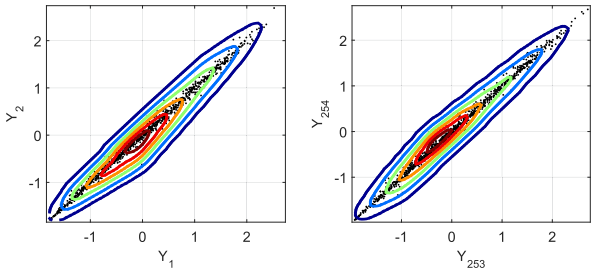


Fig. 5. Scatter plot of transformed extreme precipitation values for two randomly selected pairs of sites at the Gaussian layer.

TABLE VIII
COMPARISON OF DIFFERENT MODELS WHEN FITTING THE EXTREME PRECIPITATION DATA IN JAPAN

Models	AvgLogLLH	Prm No.	BIC	Running Time (s)
ETPCm	1.98×10^3	3936	-3.92×10^6	4.60×10^3
ETPCG	1.92×10^3	2624	-3.82×10^6	1.96×10^3
CGGM	1.96×10^3	1996	-3.89×10^6	2.89×10^2
R-vine	2.71×10^3	242643	-3.74×10^6	4.53×10^5

TABLE IX
COPULA TYPES SELECTED THROUGH MINIMIZATION OF THE BIC IN THE ETPCM FOR JAPAN RAINFALL DATA

Gaussian	t	Gumbel	Clayton	Frank	Galambos	t-EV	Hüsler-Reiss
0	1312	0	0	0	0	0	0

The results of model fitting are listed in Tables VIII and IX. In particular, for the ETPCm model, all of the chosen copulas are tail dependent, which is consistent with the empirical analysis we have performed earlier. Furthermore, the ETPCm model again achieves the best performance in terms of the BIC score. On the other hand, the R-vine overfits the data, while Gaussian copula based models cannot well capture the spatial dependence between extreme values. Interestingly, CGGMs performs better than ETPCG for this data set, since cyclic graphs offer a more natural description of the spatial dependence than the ensemble-of-trees configuration. Unfortunately, the cyclic structure can only be imposed on Gaussian copulas so far, whereas Gaussian copulas fail to capture tail dependence both empirically (as shown in our experiments) and theoretically. Moreover, the computational complexity for learning CGGMs is cubic, making them intractable to problems with a large number of sites. Therefore, the ETPCm model is currently preferable in terms of computational efficiency and tail properties.

IX. CONCLUSION

We have proposed the ETPC model for joint analysis of spatial extremes, together with highly efficient learning and imputation algorithms. The ETPC model enables us to construct the joint PDF in a flexible manner to accommodate various types of copulas for different modeling purposes, depending on the data at hand. Furthermore, it enjoys the attractive property of preserving the upper tail dependence between any two nodes under certain mild conditions, which assists in the analysis of extreme events in a spatial domain. To the best of our knowledge, this is the first work on tail dependence in undirected graphical models. Our experimental results on both synthetic and real data further show that the ETPCm model outperforms the R-vine and the CGGM in terms of model fitting, extreme data imputation and computational efficiency, indicating that the ETPC model serves as a powerful and flexible tool to model spatial extremes.

As demonstrated in Table IV and Table VIII, cyclic graphs can be a better choice to analyze spatial dependence. So far, it is unclear how to build cyclic graphical models from non-Gaussian pairwise copulas. Therefore, we would like to employ non-Gaussian copulas as the building blocks of cyclic graphical models in our future work. In addition, it would be interesting to develop efficient algorithms to draw extreme-value samples from the ETPC model, and to further estimate the return level.

APPENDIX A
PROOF OF PROPOSITION 1

Before proving the proposition, we would like to introduce the stochastic increasing condition first: for two continuously distributed variables X_1 and X_2, X_1 is said to be stochastically increasing with X_2 if the probability $F(X_1 > x_1 | X_2 = x_2)$ increases with x_2 for all x_1 in the support of X_1 [33]. Note that this condition is satisfied by all the copulas we used in this paper when modeling positive correlation.

Returning to the main thread, we can write the joint PDF of a three-node Markov chain (X_1, X_2, X_3) by applying (22):

$$p(x_1, x_2, x_3) = c_{12}(u_1, u_2) c_{23}(u_2, u_3) f_1(x_1) f_2(x_2) f_3(x_3).$$

According to the definition of upper tail dependence (20), we obtain:

$$\begin{aligned} F(U_1 > u | U_3 > u) &= \frac{F(U_1 > u, U_3 > u)}{F(U_3 > u)} \\ &= \frac{1}{1-u} \int_u^1 \int_u^1 \int_0^1 c_{12}(u_1, u_2) c_{23}(u_2, u_3) du_2 du_1 du_3 \\ &= \frac{1}{1-u} \int_0^1 \left(\int_u^1 c_{12}(u_1, u_2) du_1 \int_u^1 c_{23}(u_2, u_3) du_3 \right) du_2 \\ &= \frac{1}{1-u} \int_u^1 C_{12}(U_1 > u | u_2) C_{23}(U_3 > u | u_2) du_2 \\ &\quad + \frac{1}{1-u} \int_0^u C_{12}(U_1 > u | u_2) C_{23}(U_3 > u | u_2) du_2 \end{aligned} \quad (64)$$

Here we define $C_{12}(U_1 > u | u_2) = \int_u^1 c_{12}(u_1, u_2)\, du_1$ following the literature of copulas [16], [55]. The first term in the above expression can be regarded as computing the expected value of the product of two functions of a random variable uniformly distributed in $[u, 1]$ with probability density function $1/(1-u)$, i.e. $E\{C_{12}(U_1 > u | u_2) C_{2,3}(U_3 > u | u_2)\}$. With the stochastically increasing assumption, $C_{12}(U_1 > u | u_2)$ and $C_{2,3}(U_3 > u | u_2)$ are increasing w.r.t. $u_2 \in [u, 1)$. Since the covariance of two increasing functions of a random variable is non-negative, assuming that it exists, we can obtain the follow inequality [55]:

$$\begin{aligned} &E\{C_{12}(U_1 > u | u_2) C_{23}(U_3 > u | u_2)\} \\ &- E\{C_{12}(U_1 > u | u_2)\} E\{C_{23}(U_3 > u | u_2)\} \geq 0. \end{aligned} \quad (65)$$

Correspondingly,

$$\begin{aligned}
&\lim_{u\to 1^-}\frac{1}{1-u}\int_u^1 C_{12}(U_1>u|u_2)C_{23}(U_3>u|u_2)\,du_2\\
&\geq \lim_{u\to 1^-}\left(\frac{1}{1-u}\int_u^1 C_{12}(U_1>u|u_2)du_2\right)\\
&\quad\cdot\left(\frac{1}{1-u}\int_u^1 C_{23}(U_3>u|u_2)du_2\right)\\
&= \lim_{u\to 1^-} F(U_1>u|U_2>u)F(U_3>u|U_2>u)\\
&= \lambda_{12}\lambda_{23}. \qquad (66)
\end{aligned}$$

On the other hand, the second term in (64) can be simplified as follows:

$$\begin{aligned}
&\frac{1}{1-u}\int_0^u C(U_1>u|u_2)C(U_3>u|u_2)\,du_2\\
&= \frac{u}{1-u}\int_0^u \frac{1}{u}C(U_1>u|u_2)C(U_3>u|u_2)\,du_2\\
&= \frac{u}{1-u}E[C(U_1>u|u_2)C(U_3>u|u_2)]\\
&\geq \frac{u}{1-u}E[C(U_1>u|u_2)]E[C(U_3>u|u_2)]\\
&= \frac{u}{1-u}\left[\int_0^u \frac{1}{u}C(U_1>u|u_2)\,du_2\right]\\
&\quad\cdot\left[\int_0^u \frac{1}{u}C(U_3>u|u_2)\,du_2\right]\\
&= \frac{1}{u(1-u)}\left[\int_0^u (1-C(U_1<u|u_2))\,du_2\right]\\
&\quad\cdot\left[\int_0^u (1-C(U_3<u|u_2))\,du_2\right]\\
&= \frac{[u-C_{12}(u,u)]\,[u-C_{23}(u,u)]}{u(1-u)}. \qquad (67)
\end{aligned}$$

To compute the limit of the above, We then apply L'hôpital's rule as $[u-C_{12}(u,u)]\,[u-C_{23}(u,u)]$ and $u(1-u)$ both approach 0 when $u\to 1^-$. Noting that $C_{jk}(u,u)=\int_0^u\int_0^u c_{jk}(u_j,u_k)\,du_j du_k$, Leibniz' Integral Rule can be employed to yield

$$\begin{aligned}
\frac{d}{du}C_{jk}(u,u) &= \int_0^u c_{jk}(u,u_k)\,du_k+\int_0^u c_{jk}(u_j,u)\,du_j\\
&= C_{jk}(U_j<u|u)+C_{jk}(U_k<u|u). \qquad (68)
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\lim_{u\to 1^-}\frac{1}{1-u}\int_0^u C(U_1>u|u_2)C(U_3>u|u_2)\,du_2\\
&= \lim_{u\to 1^-}\frac{[u-C_{12}(u,u)][1-C_{23}(U_2<u|u)-C_{23}(U_3<u|u)]}{1-2u}\\
&\quad+\frac{[1-C_{12}(U_1<u|u)-C_{12}(U_2<u|u)]\,[u-C_{23}(u,u)]}{1-2u}\\
&= 0, \qquad (69)
\end{aligned}$$

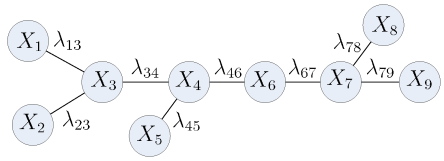


Fig. 6. An example of a tree graphical model with pairwise upper tail dependence.

because $\lim_{u\to 1^-} C_{jk}(U_j<u|u)<\infty$ by the definition of conditional cumulative distribution function.

This implies that, from (66) and (69),

$$\lambda_{13}=\lim_{u\to 1^-} F(U_1>u|U_3>u)\geq \lambda_{12}\lambda_{23}, \qquad (70)$$

where λ_{13} denotes the upper tail dependence coefficient between X_1 and X_3.

APPENDIX B
PROOF OF PROPOSITION 2

We proceed via an induction argument. Consider an undirected tree graphical model $\mathcal{T}$ with N nodes. There exists a unique path in the tree between two arbitrary distinct nodes, X_{i_1} and X_{i_n}, which we denote as $\{i_1,i_2,...,i_n\}$. We label the indices of the remaining nodes in the tree absent in the path as $i_{n+1},\ldots,i_N$.

Note that from the fact that the marginals of a copula distribution are uniformly distributed we obtain:

$$\int_0^1 c_{ij}(u_i,u_j)du_i=1. \qquad (71)$$

Therefore, the pairwise copulas which are not on the path between node i_1 and i_n will be integrated out when computing the tail dependence coefficient. As an example, the joint probability density function corresponding to the tree in Fig. 6 is

$$p(x_1,x_2,\cdots,x_{10})=c_{13}c_{23}c_{34}c_{45}c_{46}c_{67}c_{78}c_{79}\prod_{i=1}^{9}f_i(x_i),$$

where we use c_{ij} to denote $c_{ij}(u_i,u_j)$ for simplicity. Let us now analyze, for instance, the tail dependence between X_3 and X_7. According to the definition of upper tail dependence,

$$\begin{aligned}
&F(U_3>u|U_7>u)=\frac{F(U_3>u,U_7>u)}{F(U_7>u)}\\
&=\frac{1}{1-u}\int_u^1\int_u^1\left(\int_0^1\cdots\int_0^1 c_{13}c_{23}c_{34}c_{45}c_{46}c_{67}c_{78}c_{79}\right.\\
&\qquad\left.du_{1:9|3,7}\right)du_3du_7\\
&=\frac{1}{1-u}\int_u^1\int_u^1\int_0^1\int_0^1\left(\int_0^1 c_{13}du_1\int_0^1 c_{23}du_2\int_0^1 c_{45}du_5\right.\\
&\qquad\left.\cdot\int_0^1 c_{78}du_8\int_0^1 c_{79}du_9\right)\\
&\qquad\times\ c_{34}c_{46}c_{67}du_4du_6du_3du_7.
\end{aligned}$$

By utilizing (71), $F(U_3 > u|U_7 > u)$ can be further simplified as:

$$\frac{1}{1-u}\int_u^1\int_u^1\int_0^1\int_0^1 c_{34}c_{46}c_{67}du_4du_6du_3du_7, \tag{72}$$

which only involves pairwise copulas comprising the path between X_3 and X_7.

Returning to the main thread, we first deal with the case when $n=3$, that is, there are only two distinct edges which comprise the path between X_{i_1} and X_{i_3}. Let $\lambda_{i_1 i_3}$ denote the upper tail dependence coefficient between X_{i_1} and X_{i_3}. From Proposition 1, $\lambda_{i_1 i_3} = \lim_{u\to 1^-} F(U_{i_1} > u|U_{i_3} > u) \geq \lambda_{i_1 i_2}\lambda_{i_2 i_3}$, assuming that $\lambda_{i_1 i_2}$ and $\lambda_{i_2 i_3}$ both exist.

We now tackle the general case when $n > 3$. Suppose that the proposition holds for the nodes X_{i_1} and $X_{i_{n-1}}$ where the length of the path between the nodes is $n-1$. More explicitly, suppose that the upper tail dependence between X_{i_1} and $X_{i_{n-1}}$, denoted by $\lambda_{i_1 i_{n-1}}$, satisfies $\lambda_{i_1 i_{n-1}} \geq \prod_{t=1}^{n-2}\lambda_{i_t i_{t+1}}$. By the definition of upper tail dependence,

$$\begin{aligned}
&F(U_{i_1} > u|U_{i_n} > u) = \frac{F(U_{i_1} > u, U_{i_n} > u)}{F(U_{i_n} > u)}\\
&= \frac{1}{1-u}\int_u^1\int_u^1\left(\int_0^1\cdots\int_0^1 \prod_{\{i_s,i_t\}\in T} c_{i_s i_t}du_{i_2}\cdots du_{i_{n-1}}\right.\\
&\quad \left. du_{i_{n+1}}\cdots du_{i_N}\right)du_{i_1}du_{i_n}\\
&= \frac{1}{1-u}\int_u^1\int_u^1\left(\int_0^1\cdots\int_0^1 \prod_{t=1}^{n-1} c_{i_t i_{t+1}}du_{i_2}\cdots du_{i_{n-1}}\right)\\
&\quad du_{i_1}du_{i_n}\\
&= \frac{1}{1-u}\int_0^1\cdots\int_0^1\left(\int_u^1 c_{i_1 i_2}du_{i_1}\int_u^1 c_{i_{n-1}i_n}du_{i_n}\right)\\
&\quad \cdot\prod_{t=2}^{n-2} c_{i_t i_{t+1}}du_{i_2}\cdots du_{i_{n-1}}\\
&= \frac{1}{1-u}\int_0^1\cdots\int_0^1 C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\\
&\quad \cdot C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\prod_{t=2}^{n-2} c_{i_t i_{t+1}}du_{i_2}\cdots du_{i_{n-1}} \qquad (73)\\
&= \frac{1}{1-u}\int_u^1\int_0^1\underbrace{\cdots}_{n-3}\int_0^1 C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\\
&\quad \cdot C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\prod_{t=2}^{n-2} c_{i_t i_{t+1}}du_{i_2}\cdots du_{i_{n-1}}\\
&\quad + \frac{1}{1-u}\int_0^u\int_0^1\underbrace{\cdots}_{n-3}\int_0^1 C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\\
&\quad \cdot C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\prod_{t=2}^{n-2} c_{i_t i_{t+1}}du_{i_2}\cdots du_{i_{n-1}}. \qquad (74)
\end{aligned}$$

Note that expression (74) is obtained from (73) by splitting the integral w.r.t. $u_{i_{n-1}}$ into two parts. Next, we analyze the two terms in (74) separately. The first term in (74) can be regarded as the expected value of a product of two functions of $n-2$ jointly distributed random variables with joint pdf $\prod_{t=2}^{n-2} c_{i_t i_{t+1}}(u_{i_t}, u_{i_{t+1}})/(1-u)$ in the region $[0,1]^{n-3}\times[u,1]$, i.e. $E\{C_{i_1 i_2}(U_{i_1} > u|u_{i_2})C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\}$. The stochastically increasing assumption coupled with the fact that all the copulas with upper tail dependence only model positive dependence guarantee $\mathrm{Cov}\{C_{i_1 i_2}(U_{i_1} > u|u_{i_2})C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\} \geq 0$. Thus,

$$\begin{aligned}
&E\{C_{i_1 i_2}(U_{i_1} > u|u_{i_2})C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\}\\
&\geq E\{C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\}E\{C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\} \qquad (75)
\end{aligned}$$

We then find that:

$$\begin{aligned}
&E\{C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\}\\
&= \frac{1}{1-u}\int_u^1\int_0^1\underbrace{\cdots}_{n-3}\int_0^1 C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\prod_{t=2}^{n-2} c_{i_t i_{t+1}}\\
&\quad du_{i_2}\cdots du_{i_{n-1}}\\
&= \frac{1}{1-u}\int_0^1\cdots\int_0^1 C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\prod_{t=2}^{n-3} c_{i_t i_{t+1}}\\
&\quad \cdot\left(\int_u^1 c_{i_{n-2}i_{n-1}}du_{i_{n-1}}\right)du_{i_2}\cdots du_{i_{n-2}}\\
&= \frac{1}{1-u}\int_0^1\cdots\int_0^1 C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\\
&\quad \cdot C_{i_{n-2}i_{n-1}}(U_{i_{n-1}} > u|u_{i_{n-2}})\prod_{t=2}^{n-3} c_{i_t i_{t+1}}du_{i_2}\cdots du_{i_{n-2}}\\
&= F(U_{i_1} > u|U_{i_{n-1}} > u) \qquad (76)
\end{aligned}$$

which is bounded below by $\prod_{t=1}^{n-2}\lambda_{i_t i_{t+1}}$ as $u\to 1^-$ by the induction assumption. On the other hand,

$$E\{C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}}) = F(U_{i_n} > u|U_{i_{n-1}} > u) \tag{77}$$

which equals $\lambda_{i_{n-1}i_n}$ as $u\to 1^-$. Therefore, by taking (76) and (77) together, we have the following lower bound for the first term in (74):

$$\begin{aligned}
&\lim_{u\to 1^-}\frac{1}{1-u}\int_u^1\int_0^1\underbrace{\cdots}_{n-3}\int_0^1 C_{i_1 i_2}(U_{i_1} > u|u_{i_2})\\
&\quad \cdot C_{i_{n-1}i_n}(U_{i_n} > u|u_{i_{n-1}})\prod_{t=2}^{n-2} c_{i_t i_{t+1}}du_{i_2}\cdots du_{i_{n-1}}\\
&\geq \prod_{t=1}^{n-1}\lambda_{i_t i_{t+1}}. \qquad (78)
\end{aligned}$$

The second term in (74), on the other hand, can be simplified in a similar fashion as:

$$\frac{1}{1-u}\int_0^u \int_0^1 \underbrace{\cdots}_{n-3} \int_0^1 C_{i_1 i_2}(U_{i_1} > u | u_{i_2}) \cdot C_{i_{n-1} i_n}(U_{i_n} > u | u_{i_{n-1}}) \prod_{t=2}^{n-2} c_{i_t i_{t+1}} \, du_{i_2} \cdots du_{i_{n-1}} = \frac{[u - C_{i_1 i_{n-1}}(u,u)]\,[u - C_{i_{n-1} i_n}(u,u)]}{u(1-u)}, \tag{79}$$

which approaches 0 as $u \to 1^-$ using a similar argument as presented in the proof of Proposition 1.

Hence, from (78) and (79),

$$\lambda_{i_1 i_n} = \lim_{u \to 1^-} F(U_{i_1} > u | U_{i_n} > u) \geq \prod_{t=1}^{n-1} \lambda_{i_t i_{t+1}}.$$

APPENDIX C
PROOF OF PROPOSITION 3

Denote by $\mathcal{G}$ the graph under consideration and let x_1 and x_2 be distinct and arbitrary nodes. We label each spanning tree of $\mathcal{G}$ by $\mathcal{T}_i$ and label the path between x_1 and x_2 in $\mathcal{T}_i$ by $P^i_{x_1 x_2}$. By the definition of upper tail dependence,

$$\begin{aligned}
&\lim_{u \to 1^-} F(U_1 > u | U_2 > u) \\
&= \lim_{u \to 1^-} \sum_i F(U_1 > u | U_2 > u, \mathcal{T}_i) F(\mathcal{T}_i) \\
&= \frac{1}{\det[Q(\boldsymbol{\beta})]} \sum_i \left[\prod_{\{j,k\} \in \mathcal{T}_i} \beta_{jk} \right] \lim_{u \to 1^-} F(U_1 > u | U_2 > u, \mathcal{T}_i) \\
&\geq \frac{1}{\det[Q(\boldsymbol{\beta})]} \sum_i \left[\prod_{\{j,k\} \in \mathcal{T}_i} \beta_{jk} \right] \left[\prod_{\{j,k\} \in P^i_{x_1 x_2}} \lambda_{jk} \right].
\end{aligned} \tag{80}$$

Since all $\lambda_{jk} \in [0,1]$, the above expression can be further relaxed by including other λ_{jk} outside the path $P_{x_1 x_2}$, that is,

$$\begin{aligned}
&\lim_{u \to 1^-} F(U_1 > u | U_2 > u) \\
&\geq \frac{1}{\det[Q(\boldsymbol{\beta})]} \sum_i \left[\prod_{\{j,k\} \in \mathcal{T}_i} \beta_{jk} \right] \left[\prod_{\{j,k\} \in \mathcal{T}_i} \lambda_{jk} \right] \\
&= \frac{1}{\det[Q(\boldsymbol{\beta})]} \sum_i \left[\prod_{\{j,k\} \in \mathcal{T}_i} \beta_{jk} \lambda_{jk} \right] \\
&= \frac{\det[Q(\boldsymbol{\beta} \odot \lambda)]}{\det[Q(\boldsymbol{\beta})]},
\end{aligned} \tag{81}$$

where the final equality is due to the matrix-tree theorem.

We remark that because $\sum_i F(\mathcal{T}_i) = 1$ and that $\lim_{u \to 1^-} F(U_1 > u | U_2 > u, \mathcal{T}_i) \in [0,1]$, $\lim_{u \to 1^-} F(U_1 > u | U_2 > u) \in [0,1]$.

REFERENCES

[1] 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.

[2] R. D. Knabb, J. R. Rhome, and D. P. Brown, "Tropical cyclone report: Hurricane Katrina, August 23–30, 2005," *United States Nat. Ocean. Atmos. Administration's Nat. Weather Serv.*, retrieved Dec. 10, 2012.

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

[4] 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.

[5] 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, Jun. 2007.

[6] H. Yu and J. Dauwels, "Modeling spatio-temporal extreme events using graphical models," *IEEE Trans. Signal Process.*, vol. 64, no. 5, pp. 1101–1116, Mar. 1, 2015.

[7] D. Cooley *et al.*, "A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects," *REVSTAT Stat. J.*, vol. 10, pp. 135–165, 2012.

[8] P. J. Northrop and P. Jonathan, "Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights," *Environmetrics*, vol. 22, no. 7, pp. 799–809, 2011.

[9] D. Cooley and S. Sain, "Spatial hierarchical modeling of precipitation extremes from a regional climate model," *J. Agric., Biol. Environ. Stat.*, vol. 15, pp. 381–402, 2010.

[10] H. Sang and A. E. Gelfand, "Hierarchical modeling for extreme values observed over space and time," *Environ. Ecological Stat.*, vol. 16, no. 3, pp. 407–426, 2009.

[11] H. Sang and A. E. Gelfand, "Continuous spatial process models for spatial extreme values," *J. Agric., Biol. Environ. Stat.*, vol. 15, pp. 49–65, 2010.

[12] 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.

[13] A. C. Davison and M. M. Gholamrezaee, "Geostatistics of extremes," *Proc. Roy. Soc. Lond. Ser. A*, vol. 468, pp. 581–608, 2012.

[14] A. C. Davison, S. Padoan, and M. Ribatet, "Statistical modeling of spatial extremes," *Statist. Sci.*, vol. 27, no. 2, pp. 161–186, 2012.

[15] S. A. Padoan, M. Ribatet, and S. A. Sisson, "Likelihood-based inference for max-stable processes," *J. Am. Stat. Assoc.*, vol. 105, no. 489, pp. 263–277, 2010.

[16] K. Aas, C. Czado, A. Frigessi, and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance: Math. Econ.*, vol. 44, pp. 182–198, 2009.

[17] J. Dissmann, E. C. Brechmann, C. Czado, and D. Kurowicka, "Selecting and estimating regular vine copulae and application to financial returns," *Comput. Stat. Data Anal.*, vol. 59, no. 1, pp. 52–69, 2013.

[18] H. Joe, H. Li, and A. K. Nikoloulopoulos, "Tail dependence functions and vine copulas," *J. Multivariate Anal.*, vol. 101, no. 1, pp. 252–270, 2010.

[19] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with applications to financial data," *Can. J. Stat.*, vol. 40, no. 1, pp. 68–85, 2012.

[20] T. M. Erhardt, C. Czada, and U. Schepsmeier, "R-vine models for spatial time series with an application to daily mean temperature," *Biometrics*, vol. 71, pp. 323–332, 2015.

[21] T. M. Erhardt, C. Czada, and U. Schepsmeier, "Spatial composite likelihood inference using local C-vines," *J. Multivariate Anal.*, vol. 138, pp. 74–88, 2015.

[22] B. Gräler, "Modelling skewed spatial random fields through the spatial vine copula," *Spatial Stat.*, vol. 10, pp. 87–102, 2014.

[23] H. Yu, Z. Choo, J. Dauwels, P. Jonathan, and Q. Zhou, "Modeling spatial extreme events using Markov random field priors," in *Proc. ISIT*, 2012, pp. 1453–1457.

[24] H. Yu, J. Dauwels, and P. Jonathan, "Extreme-value graphical models with multiple covariates," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5734–5747, Nov. 1, 2014.

[25] Y. Lin, S. Zhu, D. D. Lee, and B. Taskar, "Learning sparse Markov network structure via ensemble-of-trees models," in *Proc. 12th Int. Conf. Artif. Intell. Stat.*, 2009, pp. 360–367.

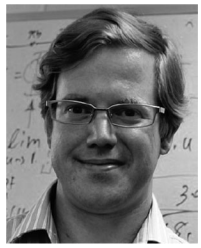
[26] S. Kirshner, "Learning with tree-averaged densities and distributions," in *Proc. 20th Int. Conf. Adv. Neural Inf. Process. Syst*, 2008, pp. 761–768.

[27] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Univ. Press, 1996.
[28] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1120–1146, May 2003.
[29] 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.
[30] P. Naveau, R. Huser, P. Ribereau, and A. Hannart, "Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection," *Water Resour. Res.*, vol. 52, pp. 2753–2769, 2016.
[31] P. Jonathan and K. Ewans, "Uncertainties in extreme wave height estimates for hurricane-dominated regions," *J. Offshore Mech. Arctic Eng.*, vol. 129, no. 4, pp. 300–305, 2007.
[32] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," (in French) *Publications de l'Institut de Statistique de L'Université de Paris*, vol. 8, pp. 229–231, 1959.
[33] P. K. Trivedi and D. M. Zimmer, *Copula Modeling: An Introduction for Practitioners*. Hanover, MA, USA: Now Publishers, 2007.
[34] S. Demarta and A. J. McNeil, "The t copula and related copulas," *Int. Stat. Rev.*, vol. 73, no. 1, pp. 111–129, 2005.
[35] G. Gudendorf and J. Segers, "Extreme-value copulas," in *Proc. Workshop Copula Theory Its Appl.*, 2010, pp. 127–146.
[36] A. Subramanian, A. Sundaresan, and P. K. Varshney, "Detection of dependent heavy-tailed signals," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2790–2803, Jun. 2015.
[37] V. Pereira, A. Giremus, and E. Grivel, "Modeling of multipath environment using copulas for particle filtering based GPS navigation," *IEEE Signal Process. Lett.*, vol. 19, no. 6, pp. 360–363, Jun. 2012.
[38] V. A. Krylov, G. Moser, S. B. Serpico, and J. Zeurbia, "Supervised high-resolution dual-polarization SAR image classification by finite mixtures and copulas," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 554–566, Jun. 2011.
[39] C. Li, J. Li, and B. Fu, "Magnitude-phase of quaternion wavelet transform for texture representation using multilevel copula," *IEEE Signal Process. Lett.*, vol. 20, no. 8, pp. 799–802, Aug. 2013.
[40] A. Sundaresan and P. K. Varshney, "Location estimation of a random signal source based on correlated sensor observations," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 787–799, Feb. 2011.
[41] M. Meilă and T. Jaakkola, "Tractable Bayesian learning of tree belief networks," *Stat. Comput.*, vol. 16, no. 1, pp. 77–92, 2006.
[42] L. W. Beineke, R. J. Wilson, and P. J. Cameron, *Topics in Algebraic Graph Theory.* New York, NY, USA: Cambridge Univ. Press, 2004.
[43] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 2003.
[44] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed. Philadelphia, PA, USA: SIAM, 2003.
[45] 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.
[46] 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, May 2008.
[47] 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, Oct. 2008.
[48] J. M. Tang and Y. Saad, "A probing method for computing the diagonal of a matrix inverse," *Numer. Linear Algebra Appl.*, vol. 19, no. 3, pp. 485–501, 2012.
[49] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.
[50] M. Schmidt, N. L. Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," doi:10.1007/s10107-016-1030-6, 2016.
[51] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Norwell, MA, USA: Kluwer, 1993.
[52] M. Leordeanu and M. Hebert, "Smoothing-based optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[53] C. Shen, M. J. Brooks, and A. van den Hengel, "Fast global kernel density mode seeking: Applications to Localization and Tracking," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1457–1469, Jun. 2007.
[54] M. Titsias and M. Lázaro-Gredilla, "Doubly stochastic variational Bayes for non-conjugate inference," *J. Mach. Learn. Res.*, vol. 32, no. 1, pp. 1971–1979, 2014.
[55] R. M. Cooke, C. Kousky, and H. Joe, "Micro Correlations and Tail Dependence," in *Dependence Modeling:Vine Copula Handbook*, D. Kurowicka and H. Joe, Eds. Singapore: World Scientific, 2011, pp. 89–113.
[56] J. L. Wadsworth and J. A. Tawn, "Dependence modeling for spatial extremes," *Biometrika*, vol. 99, pp. 253–272, 2012.
[57] A. C. Davison, R. Huser, and E. thibaud, "Geostatistics of dependent and asymptotically independent extremes," *Math. Geosci*, vol. 45, pp. 511–529, 2013.
[58] M. Schlather, "Models for stationary max-stable random fields," *Extremes*, vol. 5, no. 1, pp. 33–44, 2002.
[59] C. Dombry, F. Éry-Minko, and M. Ribatet, "Conditional simulations of max-stable processes," *Biometrika*, vol. 101, no. 1, pp. 1–15, 2014.
[60] Y. Liu, O. Kosut, and A. S. Willsky, "Sampling from Gaussian Markov random fields using stationary and non-stationary subgraph perturbations," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 576–589, Feb. 1, 2015.
[61] 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," *Hydrol. Res. Lett.*, vol. 4, pp. 60–64, 2010.
[62] B. Renard and M. Lang, "Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology," *Adv. Water Resources*, vol. 30, no. 4, pp. 897–912, 2007.

Hang Yu (S'12–M'15) received the B.E. degree in electronic and information engineering from the University of Science and Technology Beijing (USTB), Beijing, China, in 2010, and the Ph.D. degree in electrical and electronics engineering from Nanyang Technological University (NTU), Singapore, in 2015.

He is currently a Postdoctoral Research Fellow in the Centre for System Intelligence and Efficiency (EXQUISITUS), NTU, under the guidance of Prof. Justin Dauwels. His research interests include statistical signal processing, machine learning, graphical models, copulas, and extreme-events modeling.

Wayne Isaac T. Uy received the Undergraduate degree in applied math from Nanyang Technological University, Singapore, in 2013. He is currently working toward the Ph.D. degree in the Center for Applied Mathematics, Cornell University, Ithaca, NY, USA. His research interests include probabilistic surrogate models for uncertainty quantification.



Justin Dauwels (S'02–M'05–SM'12) received the Ph.D. degree in electrical engineering from the Swiss Polytechnical Institute of Technology (ETH), Zurich, Switzerland, in December 2005. He is an Associate Professor with the School of Electrical & Electronic Engineering, Nanyang Technological University (NTU), Singapore. He is the Deputy Director of the ST Engineering-NTU Corporate Laboratory on Autonomous Systems. His research interests include Bayesian statistics, iterative signal processing, and computational neuroscience. 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, Straits Times, and various other media outlets. His research on Alzheimer's disease is featured at a five-year exposition at the Science Center in Singapore. His research team has won several best paper awards at international conferences. He has filed five U.S. patents related to data analytics.