Latent Tree Ensemble of Pairwise Copulas for Spatial Extremes Analysis

Hang Yu, Junwei Huang, and Justin Dauwels
School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore

***Abstract*—We consider the problem of jointly describing extreme events at a multitude of locations, which presents paramount importance in catastrophes forecast and risk management. Specifically, a novel Ensemble-of-Latent-Trees of Pairwise Copula (ELTPC) model is proposed. In this model, the spatial dependence is captured by latent trees expressed by pairwise copulas. To compensate the limited expressiveness of every single latent tree, an mixture of latent trees is employed. By harnessing the variational inference and stochastic gradient techniques, we further develop a triply stochastic variational inference (TSVI) algorithm for learning and inference. The corresponding computational complexity is only *linear* in the number of variables. Numerical results from both the synthetic and real data show that the ELTPC model provides a reliable description of the spatial extremes in a flexible but parsimonious manner.**

I. Introduction

Extreme events such as hurricanes, droughts, and floods often give rise to fatal consequences and huge economic loss. For example, Hurricane Sandy in 2012 results in a death toll of at least 233 and a total loss of 75 billion USD. Worse still, both observational data and climate models imply that both occurrences and sizes of such catastrophes will increase in the future. It is therefore imperative to analyze such events, assess the likelihood, and further take precaution measurements.

Models of spatial extremes often resort to a two-stage approach [1]. The first stage involves learning the marginal distributions at each measuring site. The marginal parameters are typically coupled in space to address the issue of inaccurate estimation due to the lack of extreme-value samples [2]. We then proceed to the second stage and couple the marginals together via copulas or max-stable processes. For instance, the marginals are glued together by a Gaussian copula in [3] and a copula Gaussian graphical model in [4]. However, Gaussian copulas are asymptotically tail independent [5], indicating that they are not able to capture the dependence between high-quantile extremes (e.g., extreme wind gusts at different locations). To address this issue, models with tail dependence are proposed, including max-stable processes and extreme-value (max-stable) copulas [5]. Unfortunately, such models are defined by complicated cumulative distribution functions (CDF). The corresponding likelihood function involves differentiation with respect to (w.r.t.) all variables, and hence is intractable in high-dimensional settings [5]. As a remedy, composite likelihood [6] replaces the original likelihood by a pseudo-likelihood composing of pairwise likelihoods. However, different configurations of composite likelihood have to be tested before the best one can be determined. Vine copula models further overcome this deficiency by expressing different orders of dependencies using pairwise-copula trees and selecting all trees to be maximum spanning trees [7]. Moreover, vine copulas can capture tail dependence by properly selecting the base pairwise copulas [8]. However, the biggest impediment to the application of vine copulas is the quadratically increasing number of parameters w.r.t. the dimension.

As a consequence, the existing extreme-value models are often limited to tens of variables. Yet many practical statistical problems, for instance in Earth Sciences, involve thousands or millions of sites (variables). On the other hand, graphical models can handle a large number of variables [9], nevertheless, they have rarely been applied to the realm of extreme-event analysis. When describing spatial extremes, it is reasonable to use a lattice graph as shown in Fig. 1a. Unfortunately, construction of a cyclic graphical model by means of pairwise copulas is intractable; the corresponding log partition function cannot be computed in a closed form. However, the density function of a tree graphical model can be expressed by pairwise copulas. As such, our objective is to approximate the lattice graphical model using tree graphical models. A fruitful step directed to this end is taken in [10], [11], where a mixture of all the spanning trees in the lattice is utilized. Moreover, there exists tail dependence in trees under mild conditions [11].

In this paper, to better approximate the cyclic graph, we propose a novel extreme-value graphical model, i.e., ensemble-of-latent-trees of pairwise copulas (ELTPC). Specifically, instead of using the spanning trees of the observed variables, we introduce hidden variables and exploit latent trees (see Fig. 1b). As shown in Fig. 1c and 1d, by integrating out the hidden variables, the graphical model capturing the dependence between observed variables is cyclic. However, one issue that frequently arises for latent trees is the presence of blocky artifacts [9], [12], since pairs of variables with the same physical distance have different statistical distances in the graph. To overcome this difficulty, we use an ensemble of latent trees as shown in Fig. 2. More concretely, we first construct a pyramidal graph (see Fig. 2a) by associating the nodes in the bottom scale with the measuring stations. Next, we consider all the spanning trees of the pyramidal graph, which can be expressed by pairwise copulas. The resulting ELTPC model possesses superior modeling power over the ensemble of spanning trees of observed variables [10], [11], while at the same time guaranteeing the tail dependence, and avoiding the need to compute the log partition function as in cyclic graphs. We then develop an efficient triply stochastic

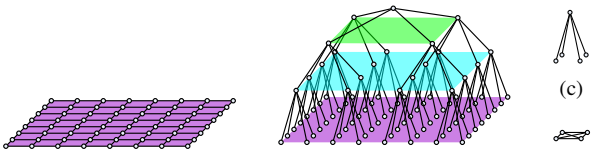


Fig. 1: A latent tree (c) can approximate the cyclic graph (d) after integrating out the latent variables

variational inference (TSVI) algorithm to learn the parameters of the proposed model. The results from synthetic and real data show that the ETPLC model can well describe extreme events with few parameters.

This paper is structured as follows. First we learn the extreme-value marginal distributions in Section II. Next, we present the ELTPC model in Section III, and develop the TSVI algorithm in Section IV. Numerical results are provided in Section V. Finally, we offer concluding remarks in Section VI.

II. Extreme-Value Marginal Distributions

In this study, we consider maxima over a particular time period, e.g., monthly or annual maxima. Suppose that we have N samples $x_i^{(n)}$ at each of P_O sites, where $i = 1, \cdots, P_O$ and $n = 1, \cdots, N$. The extreme value theory states that the block maxima x_i of i.i.d. univariate samples at each location converge to the three-parameter Generalized Extreme Value (GEV) distribution with CDF [1]:

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

where $\mu_i \in \mathbb{R}$, $\sigma_i > 0$, and $\xi_i \in \mathbb{R}$ denote the location, scale and shape parameter, respectively. To improve the estimation accuracy, we smooth the GEV parameters across space using a thin-membrane model as in [2].

III. Ensemble of Latent Trees

In the following, we first briefly introduce copulas, and then present the ELTPC model at length.

A. From Copulas to Trees

An arbitrary copula function C taking marginal distributions $F_i(x_i)$ as its arguments, defines a valid joint distribution with marginals $F_i(x_i)$ [13]:

$$F(x_1, \ldots, x_P) = C\left(F_1(x_1), \ldots, F_P(x_P)\right) = C(u_1, \ldots, u_P),$$

where $u_i = F_i(x_i)$ follows unit uniform distributions. The corresponding joint PDF can be written as:

$$f(x_1, \ldots, x_P) = c(F_1(x_1), \ldots, F_P(x_P)) \prod_{i=1}^{P} f_i(x_i), \quad (2)$$

where c is the copula density function. We can observe from (2) that a copula separates the marginal specification from dependence modeling.

As mentioned in the introduction, the density of a tree graphical model $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ can be written as an expression of pairwise copulas [10]:

$$f(\boldsymbol{x}|\mathcal{T}) = \prod_{j \in \mathcal{V}} f_j(x_j) \prod_{(j,k) \in \mathcal{E}} \frac{f_{jk}(x_j, x_k)}{f_j(x_j) f_k(x_k)} \quad (3)$$

$$= \prod_{j \in \mathcal{V}} f_j(x_j) \prod_{(j,k) \in \mathcal{E}} c_{jk}(F_j(x_j), F_k(x_k)), \quad (4)$$

where it follows from (2) that $f_{jk}/(f_j f_k) = c_{jk}$. As proven in [11], two variables in a tree are tail dependent if all the pairwise copulas on the path connecting them have tail dependence.

B. Ensemble of Latent Trees

In this section, we model the spatial dependence between different locations using an ensemble of latent trees. As shown in Fig. 2, we first introduce hidden variables by imposing a pyramidal graph (Fig. 2a) on the observed variables. To remove the blocky artifacts of one single latent tree [9], we further consider a mixture of all the spanning trees of the pyramidal graph. The proposed model has three pleasing properties: First, latent trees can be expressed by pairwise copulas (4) [14], resulting in tractable and compact representation of large-scale data. Second, latent trees can tackle tail dependence among the extremes by properly selecting the constituent pairwise copulas. Third and most importantly, after integrating out hidden variables, this model can better approximate the intractable cyclic graph than the mixture of spanning trees of the observed variables [10], [11].

Specifically, let $\boldsymbol{x}_O$ and $\boldsymbol{x}_H$ represent the observed and hidden variables respectively and $\boldsymbol{x} = [\boldsymbol{x}_O^T, \boldsymbol{x}_H^T]^T$. The dimension of $\boldsymbol{x}_O$, $\boldsymbol{x}_H$, and $\boldsymbol{x}$ is denoted by P_O, P_H and P respectively. The density of a latent tree $\mathcal{LT}_i$ can be written as:

$$p(\boldsymbol{x}|\mathcal{LT}_i) = \prod_{j \in \mathcal{V}_O} f_j(x_j) \prod_{(j,k) \in \mathcal{E}_i} c_{jk}(F_j(x_j), F_k(x_k)). \quad (5)$$

Note that we omit the marginals of the hidden variables since they are out of our concern; we introduce hidden variables only to help explain the dependence between observed variables. Next, we build the ELTPC model by computing the weighted sum over all possible spanning trees of the pyramidal graph:

$$p(\boldsymbol{x}) = \sum_{\mathcal{LT}_i} p(\mathcal{LT}_i) p(\boldsymbol{x}|\mathcal{LT}_i), \quad (6)$$

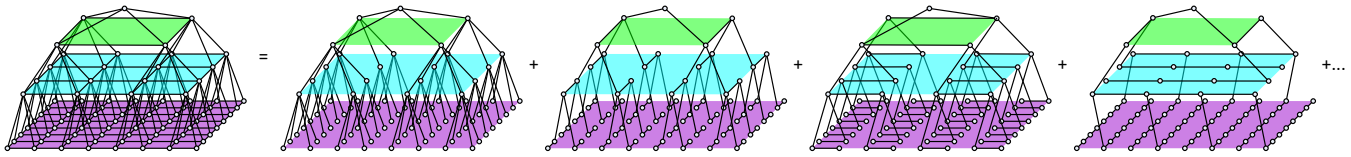
where $p(\mathcal{LT}_i)$ is a decomposable prior that assigns the weight of each latent tree $\mathcal{LT}_i$ as the product of the weights β_{jk} of the edges $(j,k) \in \mathcal{E}_i$ in the tree, i.e.,

$$p(\mathcal{LT}_i) = \frac{1}{Z} \prod_{(j,k) \in \mathcal{E}_i} \beta_{jk}, \quad (7)$$

and Z is the normalization constant. According to the matrix-tree theorem [10], $Z = \sum_{\mathcal{LT}_i} \prod_{(j,k) \in \mathcal{E}_i} \beta_{jk} = \det[Q(\boldsymbol{\beta})]$, where $\boldsymbol{\beta}$ is a symmetric edge weight matrix with zeros on the diagonal entries and with the (j,k)th entry being β_{jk}, and $Q(\boldsymbol{\beta})$ be the first $P-1$ rows and columns of the original $P \times P$ Laplacian matrix of the graph whose edge weight matrix is $\boldsymbol{\beta}$. By substituting (5) and (7) into (6), the ELTPC model can be succinctly formulated as:

$$p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\det[Q(\boldsymbol{\beta} \odot \boldsymbol{c})]}{\det[Q(\boldsymbol{\beta})]} \prod_{j \in \mathcal{V}_O} f_j(x_j), \quad (8)$$

where $\odot$ denotes componentwise multiplication, $\boldsymbol{c}$ is the copula density matrix whose (j,k)th entry equals $c_{jk}(u_j, u_k; \theta_{jk})$, and θ_{jk} is the dependence parameter of c_{jk}. In this setting, the



(a) $\mathcal{G}_{pyr} = (\mathcal{V}, \mathcal{E})$. (b) $\mathcal{LT}_1 = (\mathcal{V}, \mathcal{E}_1)$. (c) $\mathcal{LT}_2 = (\mathcal{V}, \mathcal{E}_2)$. (d) $\mathcal{LT}_3 = (\mathcal{V}, \mathcal{E}_3)$. (e) $\mathcal{LT}_4 = (\mathcal{V}, \mathcal{E}_4)$.

Fig. 2: ELT model: the pyramidal graph (a) and several decomposed spanning trees (b) - (g).

number of parameters in the ELTPC model increases linearly with P_O.

IV. Triply Stochastic Variational Inference

In this section, we focus on learning both the edge weight matrix $\boldsymbol{\beta}$ and the parameters of the pairwise copulas $\boldsymbol{\theta}$. Given N samples, our objective is to find $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ that maximizes the following log-likelihood:

$$\sum_{n=1}^{N} \log p(\boldsymbol{u}_O^{(n)}|\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \int_{\boldsymbol{u}_H} p(\boldsymbol{u}_O^{(n)}, \boldsymbol{u}_H|\boldsymbol{\beta},\boldsymbol{\theta}) d\boldsymbol{u}_H.$$

Unfortunately, the above integration is intractable due to the complex functional form of $p(\boldsymbol{u}_O, \boldsymbol{u}_H|\boldsymbol{\beta},\boldsymbol{\theta})$ as in Eq. (8). Instead, we maximize the lower bound of $\sum_{n=1}^{N} \log p(\boldsymbol{u}_O|\boldsymbol{\beta},\boldsymbol{\theta})$:

$$\mathcal{L} = \sum_{n=1}^{N} \int_{\boldsymbol{u}_H} q^{(n)}(\boldsymbol{u}_H) \log \frac{p(\boldsymbol{u}_O^{(n)}, \boldsymbol{u}_H|\boldsymbol{\beta},\boldsymbol{\theta})}{q^{(n)}(\boldsymbol{u}_H)} d\boldsymbol{u}_H. \quad (9)$$

Here, we need to choose a tractable variational distribution $q^{(n)}(\boldsymbol{u}_H)$ and such methods are referred to as variational Bayes [15]. Equivalently, we seek to minimize the KL divergence between $q^{(n)}(\boldsymbol{u}_H)$ and $p(\boldsymbol{u}_H|\boldsymbol{u}_O^{(n)};\boldsymbol{\beta},\boldsymbol{\theta})$. For simplicity, we specify $q^{(n)}(\boldsymbol{u}_H)$ to be a Gaussian copula with a diagonal covariance matrix, that is,

$$q^{(n)}(\boldsymbol{u}_H) = \frac{\phi_{P_H}\left(\Phi^{-1}(u_{H1}), \cdots, \Phi^{-1}(u_{HP_H}); \boldsymbol{m}_H^{(n)}, C_H^{(n)}\right)}{\prod_i \phi\left(\Phi^{-1}(u_{Hi})\right)}, \quad (10)$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denotes the CDF and PDF of a univariate standard Gaussian distribution, $\phi_{P_H}(\cdot; \boldsymbol{m}_H^{(n)}, C_H^{(n)})$ denotes a P_H-variate Gaussian distribution with mean $\boldsymbol{m}_H^{(n)}$ and covariance $C_H^{(n)} C_H^{(n)^T}$, and $C_H^{(n)}$ is a diagonal matrix with the standard deviation vector $\boldsymbol{v}_H^{(n)}$ on the diagonal. Note that different from a regular Gaussian copula with $\boldsymbol{m}_H^{(n)} = 0$ and $\boldsymbol{v}_H^{(n)} = 1$, we allow the mean and the covariance to change flexibly such that $q^{(n)}(\boldsymbol{u}_H)$ can better approximate $p(\boldsymbol{u}_H|\boldsymbol{u}_O^{(n)};\boldsymbol{\beta},\boldsymbol{\theta})$.

By substituting (10) into (9) and reparameterizing u_{Hi} by $\tilde{x}_{Hi} = \Phi^{-1}(u_{Hi})$, the lower bound can be simplified as:

$$\mathcal{L} = \sum_{n=1}^{N} \int \left[\phi_{P_H}\left(\tilde{\boldsymbol{x}}_H; \boldsymbol{m}_H^{(n)}, C_H^{(n)}\right) \cdot \log \frac{p\left(\boldsymbol{u}_O^{(n)}, \Phi(\tilde{x}_{H1}), \cdots, \Phi(\tilde{x}_{HP_H})|\boldsymbol{\beta},\boldsymbol{\theta}\right)}{\phi_{P_H}(\tilde{\boldsymbol{x}}_H; \boldsymbol{m}_H^{(n)}, C_H^{(n)})} \right] d\tilde{\boldsymbol{x}}_H.$$

To further simplify the problem, we change variables according to $\boldsymbol{z}_H = {C_H^{(n)}}^{-1}(\tilde{\boldsymbol{x}}_H - \boldsymbol{m}_H^{(n)})$ [15], and express $\mathcal{L}$ as:

$$\mathcal{L} = \sum_{n=1}^{N} \left\{ E_{\phi(\boldsymbol{z}_H)} \left[\log p\left(\boldsymbol{u}_O^{(n)}, \Phi\left(v_{H1}^{(n)} z_{H1} + m_{H1}^{(n)}\right), \cdots, \Phi\left(v_{HP_H}^{(n)} z_{HP_H} + m_{HP_H}^{(n)}\right)|\boldsymbol{\beta},\boldsymbol{\theta}\right) \right] - \frac{1}{2}\left[\mathrm{tr}\left({C_H^{(n)}}^T C_H^{(n)}\right) + {\boldsymbol{m}_H^{(n)}}^T \boldsymbol{m}_H^{(n)} \right] + \log\det C_H^{(n)} \right\},$$

where $\phi(\boldsymbol{z}_H)$ is a P_H-variate standard normal distribution.

In order to maximize the lower bound, we consider the gradients w.r.t. the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as well as the parameters of the variational distributions $\boldsymbol{m}_H^{(n)}$ and $\boldsymbol{C}_H^{(n)}$. For the sake of brevity, we focus on the gradients w.r.t. $\boldsymbol{\beta}$ and its stochastic variants in the sequel; other gradients can be derived in a similar spirit.

$$\frac{\partial \mathcal{L}}{\partial \beta_{jk}} = \sum_{n=1}^{N} E_{\phi(\boldsymbol{z}_H)} \left[\frac{\partial \log\det Q(\boldsymbol{\beta} \odot \boldsymbol{c}^{(n)})}{\partial \beta_{jk}} - \frac{\partial \log\det Q(\boldsymbol{\beta})}{\partial \beta_{jk}} \right], \quad (11)$$

where

$$\frac{\partial \log\det Q(\boldsymbol{\beta} \odot \boldsymbol{c})}{\partial \beta_{jk}} = c_{jk} e_{jk}^T Q(\boldsymbol{\beta} \odot \boldsymbol{c})^{-1} e_{jk}, \quad (12)$$

$$\frac{\partial \log\det Q(\boldsymbol{\beta})}{\partial \beta_{jk}} = e_{jk}^T Q(\boldsymbol{\beta})^{-1} e_{jk}, \quad (13)$$

and $\boldsymbol{e}_{jk}$ is a $(P-1)\times 1$ vector extracting elements in Q^{-1} that are related to β_{jk}. We emphasize that the computational complexity of evaluating the gradients is $\mathcal{O}(NP^3)$. Furthermore, the expectations in Eq. (11) are intractable. To address these issues, we exploit stochastic gradients. Stochastic gradients are unbiased estimates of the true gradients and they can be computed efficiently [15]. In what follows, we elaborate on how to compute stochastic gradients to achieve dramatic computational gain.

First, we make the computational cost independent of N by sampling i_κ uniformly from the set $\{1, \cdots, N\}$, and replacing the true gradient w.r.t. $\boldsymbol{\beta}$ with:

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \beta_{jk}} = N E_{\phi(\boldsymbol{z}_H)} \left[\frac{\partial \log\det Q(\boldsymbol{\beta} \odot \boldsymbol{c}^{(i_\kappa)})}{\partial \beta_{jk}} - \frac{\partial \log\det Q(\boldsymbol{\beta})}{\partial \beta_{jk}} \right].$$

Consequently, the computational complexity is reduced to $\mathcal{O}(P^3)$. Next, we approximate the intractable expectations by their unbiased estimates, that is, we draw a sample $\boldsymbol{z}_H^{\{\kappa\}}$ from the distribution $\phi(\boldsymbol{z}_H)$, and then compute the stochastic gradients as:

$$N \left[\frac{\partial \log\det Q(\boldsymbol{\beta} \odot \boldsymbol{c}^{(i_\kappa)})}{\partial \beta_{jk}} - \frac{\partial \log\det Q(\boldsymbol{\beta})}{\partial \beta_{jk}} \right] \Bigg|_{\boldsymbol{z}_H = \boldsymbol{z}_H^{\{\kappa\}}}.$$

Finally, we intend to reduce the computational complexity w.r.t. P such that the proposed model can be applicable to high-dimensional problems. The $\mathcal{O}(P^3)$ computational cost stems from the inverse of matrices in Eq. (12)-(13). Here,

instead of directly computing the inverse Q^{-1}, we seek to find a random low-rank $(P-1)\times M$ matrix L such that $E[LL^T]=I$. As a result, an unbiased estimate of the diagonal of $EQ^{-1}E^T$ can be computed with $\mathcal{O}(MP)$ complexity by first solving M sparse linear systems $QR=L$ and then compute the sum over rows in the matrix $(ER)\odot(EL)$, where E is comprised of e_{jk}. To this end, we borrow the idea in [16], [17], which was originally proposed to efficiently compute the approximate diagonal of Q^{-1} given Q, and design L in each iteration as follows: we partition all the nodes in the graph corresponding to Q into M color classes using the greedy multicoloring algorithm [17] 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. Other entries are set to be zero. We set $D=4$ in our experiments, and M is typically 64 regardless of the dimension of Q. Hence, the computational complexity is eventually reduced to $\mathcal{O}(P)$.

Based on the theory of stochastic approximation [15], using a schedule of learning rates $\{\rho_t\}$ such that $\sum\rho_t=\infty$, and $\sum\rho_t^2<\infty$, the triply stochastic variational inference algorithm will converge to a local maximum of the bound $\mathcal{L}$. Concretely, we set the learning rate using ADAM [18] in which the upper bound of learning rate is initialized as 0.01 and is multiplied by a factor of 0.9 every $50N$ iterations.

After learning the parameters, it is straightforward to further extend the algorithm to infer the variational distribution $q(\boldsymbol{x}_M)$ at unobserved locations given observed ones; we simply fix $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to their estimates and update $\boldsymbol{m}_M$ and $\boldsymbol{\nu}_M$ in each iteration.

V. Experimental Results

In this section, we test the proposed model against three other models, including copula Gaussian graphical models with a lattice structure (CGGM) [4], ensemble of trees of pairwise copulas (ETPC) [11], and regular vine copula models (R-vine) [7]. Specifically, the CGGM constructs a cyclic graph based on Gaussian copulas, the ETPC model considers an ensemble of all spanning trees of the observed variables in the lattice, and the R-vine employs a hierarchy of trees to capture all orders of dependence between observed variables. We choose all pairwise copulas to be Gumbel copula (i.e., an extreme-value copula with tail dependence) for the ETPC, the R-vine and the ELTPC. To access the performance of all four models, we consider both synthetic data sampling from max-stable random fields and real extreme precipitation in Japan. For each model, we evaluate the averaged log-likelihood (AvgLogLLH), the number of parameters (Prm No.), the score of Bayesian information criterion (BIC), and the computational time.

A. Synthetic Data

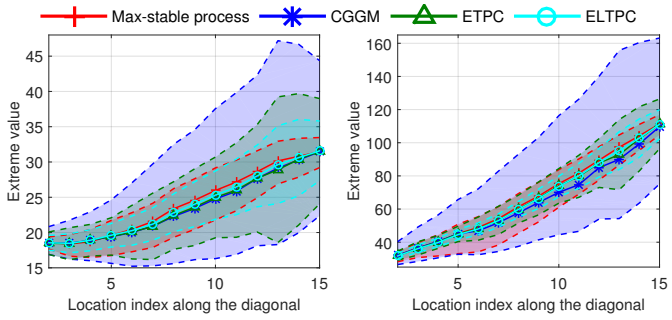
We generate synthetic data by sampling from max-stable random fields. Theoretically, max-stable processes can well characterize the joint behavior of pointwise maxima in a spatial domain. In particular, we consider a Schlather model with a powered exponential covariance function [19], and simulate 402 samples respectively from lattices with size 8×8, 16×16, 32×32 and 64×64. We use two samples corresponding to the 99th and 60th quantiles of the overall extremes in the spatial domain as the testing data, and the rest as training data.

TABLE I: 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	ELTPC	9.86×10^{1}	448	-7.62×10^{4}	3.69×10^{3}
	ETPC	9.52×10^{1}	224	-7.48×10^{4}	1.29×10^{2}
	CGGM	8.67×10^{1}	176	-6.83×10^{4}	7.28×10^{-1}
	R-vine	9.97×10^{1}	2016	-6.77×10^{4}	4.96×10^{1}
16×16	ELTPC	3.89×10^{2}	1920	-3.00×10^{5}	9.54×10^{3}
	ETPC	3.68×10^{2}	960	-2.88×10^{5}	2.53×10^{2}
	CGGM	3.19×10^{2}	736	-2.50×10^{5}	7.92
	R-vine	3.99×10^{2}	33895	-1.23×10^{5}	1.12×10^{3}
32×32	ELTPC	1.39×10^{3}	7936	-1.07×10^{6}	8.72×10^{4}
	ETPC	1.33×10^{3}	3968	-1.04×10^{6}	1.01×10^{3}
	CGGM	1.15×10^{3}	3008	-8.99×10^{5}	5.50×10^{3}
	R-vine	1.51×10^{3}	523776	1.93×10^{6}	7.56×10^{4}
64×64	ELTPC	5.42×10^{3}	32256	-4.15×10^{6}	2.93×10^{5}
	ETPC	5.26×10^{3}	16128	-4.11×10^{6}	1.24×10^{4}
	CGGM	4.60×10^{3}	12160	-3.60×10^{6}	2.54×10^{6}
	R-vine	N.A.	8386560	N.A.	$>$ two weeks

The model fitting results are summarized in Table I. As shown in the table, regardless of the grid size, the R-vine achieves the largest averaged log-likelihood at the expense of using the largest number of parameters. As a result, the BIC score of the R-vine pales in comparison with other models, suggesting that the R-vine overfits the data. In contrast, the CGGM utilizes fewest parameters, whereas it fails to describe the data accurately due to the tail independence property of Gaussian copulas. The corresponding BIC score is the second largest. Different from the R-vine and the CGGM, the ELTPC and the ETPC model successfully strike a balance between data fidelity and model complexity. When comparing the proposed ELTPC with the ETPC, we can see that the ELTPC can better fit the data without introducing many redundant parameters. Indeed, the BIC score of the proposed model is the smallest under all scenarios. As for the computational time, the R-vine and the CGGM become intractable with the increase of the dimension. This can be explained by the computational complexity of their learning algorithms, which is $\mathcal{O}(P^2)$ and $\mathcal{O}(P^3)$ respectively for the R-vine and the CGGM. On the other hand, the learning algorithms of both the ELTPC and the ETPC model scale linearly with P. Since there are hidden variables in the proposed ELTPC, the corresponding learning process takes a longer time than that of the ETPC.

Next, for two testing samples in the 16×16 grid, we consider the case where the data is missing in the 14×14 area in the center of the grid. We then infer the marginals at the missing sites given observed ones using different models. We also conduct conditional simulation in the underlying max-stable random field and regard the simulated distributions as the ground truth. Fig. 3 shows the estimated marginals along the diagonal of the missing region resulting from different models. We omit the results of the R-vine since imputation in this model is unwieldy slow. We can tell from the figure that the estimated distributions of missing data resulting from the ELTPC model follow the ground truth most closely. Indeed,



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

Fig. 3: Marginals at sites along the diagonal in the missing region conditioned on observed sites. The solid lines denote the median, and the dash lines denote the 0.025 and 0.975 quantile

TABLE II: Modeling extreme precipitation data.

Criteria	AvgLogLLH	Prm. No.	BIC	Running Time (s)
ELTPC	5.44×10^2	1920	-6.77×10^5	2.31×10^4
ETPC	5.33×10^2	960	-6.70×10^5	9.46×10^2
CGGM	5.35×10^2	736	-6.73×10^5	2.64×10^1
R-vine	6.07×10^2	32640	-5.58×10^5	1.42×10^3

the mean absolute error between the estimated and true median for the ELTPC, the ETPC, and the CGGM is respectively 0.37, 0.57 and 0.58 in the 60th-quantile case, while being 0.99, 1.06, and 2.00 in the 99th-quantile case. Moreover, the uncertainty level of the CGGM and the ETPC can be much larger than that of the ground truth, especially in the high-quantile case.

B. Real Data

The daily rainfall data from 1900–2011 is compiled and interpolated onto a grid with resolution 0.05 [20]. We select a 16×16 lattice with resolution 0.1 in South Japan, where heavy rainfall is often the cause of floods. We extract the extreme precipitation values from the sites 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 rain fall amount. The resulting sample size is 633.

We list the results in Table II. Again, the ELTPC model outperforms other models in terms of BIC score: it achieves a small negative log-likelihood without introducing too many parameters. Interestingly, the CGGM yields larger likelihood with fewer parameters when compared with the ETPC model, suggesting that using cyclic graph is necessary in this case. However, the performance of the proposed ELTPC model is even better than the CGGM, since it is a good approximation to the cyclic graph after integrating out the hidden variables and it can deal with tail dependence in the real data.

VI. Conclusions

In this paper, we proposed a novel statistical model to describe spatial extreme events. This model exploits a mixture of latent trees that can be built upon pairwise copulas. After integrating out the hidden variables in the latent trees, the resulting model can better approximate a cyclic graphical model, and therefore can capture the spatial dependence more flexibly and effectively. We further derive a triply stochastic variational inference algorithm to learn the model with low computational complexity. Numerical results demonstrate that the proposed ELTPC model achieves better performance than benchmarks when describing spatial extremes .

Acknowledgment

This research was supported by MOE (Singapore) ACRF Tier 2 grant M4020187.

References

[1] D. Cooley, J. Cisewski, R. J. Erhardt, S. Jeon, E. Mannshardt, B. O. Omolo, and Y. Sun, "A survey of spatial extremes: measuring spatial dependence and modeling spatial effects," *REVSTAT*, vol. 10, pp. 135-165, 2012.

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

[3] H. Sang, and A. E. Gelfand, "Continuous Spatial Process Models for Spatial Extreme Values," *J. Agr. Biol. Envir. St.*, vol. 15 ,pp. 49-65, 2010.

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

[5] A. C. Davison, S. Padoan, and M. Ribatet, "Statistical modeling of spatial extremes," *Statist. Sci.*, 27 (2):161-186, 2012.

[6] 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.

[7] J. Dissmann, E. C. Brechmann, C. Czado, and D. Kurowicka, "Selecting and Estimating Regular Vine Copulae and Application to Financial Returns," *Comput. Stat. Data An.*, vol. 59, no. 1, pp. 52-69.

[8] 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.

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

[10] S. Kirshner, "Learning with tree-averaged densities and distributions," *Proc. NIPS*, 2008.

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

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

[13] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de L'Université de Paris 8*, pp. 229-231, 1959.

[14] S. Kirshner, "Latent tree copulas," *6th European Workshop on Probabilistic Graphical Models*, 2012.

[15] M. Titsias, M. Lázaro-Gredilla, "Doubly Stochastic Variational Bayes for non-Conjugate Inference," *J. Mach. Learn. Res.*, W&CP, vol. 32, no. 1, 1971-1979, 2014.

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

[17] J. M. Tang, and Y. Saad, "A Probing Method for Computing the Diagonal of a Matrix Inverse,"*Numerical Lin. Alg. Appl.*, vol. 19, no. 3, pp. 485-501, 2012.

[18] D. P. Kingma, and J. L. Ba, "Adam: A method for stochastic optimization, " *Proc. ICLR*, 2015.

[19] M. Schlather, "Models for Stationary Max-Stable Random Fields," *Extremes*, vol. 5, no. 1, pp. 33-44, 2002.

[20] 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.