Copula-Based Semiparametric Models for Spatio-Temporal Data

Yanlin Tang\textsuperscript{1}, Huixia Judy Wang\textsuperscript{2,}\textsuperscript{*}, Ying Sun\textsuperscript{3}, and Amanda S. Hering\textsuperscript{4}

\textsuperscript{1}Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE, School of Statistics, East China Normal University, Shanghai, 200062, China
\textsuperscript{2}Department of Statistics, George Washington University, Washington, DC, 20052, USA
\textsuperscript{3}Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia
\textsuperscript{4}Department of Statistical Science, Baylor University, Waco, TX, 76798, USA

\textsuperscript{*}email: judywang@gwu.edu

SUMMARY: The joint analysis of spatial and temporal processes poses computational challenges due to the data’s high dimensionality. Furthermore, such data are commonly non-Gaussian. In this paper, we introduce a copula-based spatio-temporal model for analyzing spatio-temporal data and propose a semiparametric estimator. The proposed algorithm is computationally simple, since it models the marginal distribution and the spatio-temporal dependence separately. Instead of assuming a parametric distribution, the proposed method models the marginal distributions nonparametrically and thus offers more flexibility. The method also provides a convenient way to construct both point and interval predictions at new times and new locations, based on the estimated conditional quantiles. Through a simulation study and an analysis of wind speeds observed along the border between Oregon and Washington, we show that our method produces more accurate point and interval predictions for skewed data than those based on normality assumptions.

KEY WORDS: Copula; Markov process; Pseudo likelihood; Spatio-temporal.
1. Introduction

The main challenge in processing spatio-temporal data is accurately modeling the finite-dimensional joint and conditional distributions of the spatio-temporal process. These distributions can be used to describe the spatio-temporal variability of a process, to obtain forecasts at new time points and predictions at new locations, and to provide the prediction uncertainties. The model must appropriately characterize the marginal distribution for a given location while accounting for both spatial and temporal dependencies, which can be computationally challenging even for time series observed at a modest number of locations.

Gaussian processes are commonly used to model spatio-temporal data (Storlie et al., 2017; Ludwig et al., 2017; Datta et al., 2017). However, the Gaussianity assumption is known to be inappropriate or insufficient for representing many real-world datasets. When spatio-temporal processes exhibit non-Gaussian features, a Gaussian process model may be fitted to transformed data. However, direct transformations of the data cannot guarantee Gaussianity and may alter other features of the process. For example, the wind speed data studied in Section 5 have non-Gaussian distributions and remain skewed even after a power transformation (Figure 3). Wallin and Bolin (2015) and Schmidt et al. (2017) illustrated that transforming data with either a square root or a logarithm induces a link between the mean and the covariance, resulting in a non-stationary process. Therefore, transforming the data is not a desirable approach. Real data applications call for non-Gaussian models.

One aim of our work is to provide temporal forecasts and spatial predictions. To make one-step-ahead forecasts, many authors have built separate models for individual sites (Gneiting et al., 2006; Patton, 2012; Kazor and Hering, 2015). Such models cannot be used for predictions at new sites. To make predictions at new locations, the most common approach is to apply a variation of kriging (Krige, 1951; Cressie and Wikle, 2011; Gräler et al., 2016). Kriging is the best linear predictor; however, if the normality assumption fails, then the best predictor may be nonlinear, and thus kriging may no longer be optimal for predictions with a skewed distribution, such as the wind speed data. Various approaches have been proposed to relax the Gaussian assumption; see for instance Zhang and El-Shaarawi (2010), Bai et al. (2014), Quessy et al. (2015), Schmidt et al. (2017), Kang and Song (2017). These methods are either not flexible enough or computationally intensive.

Spatio-temporal predictions of wind speed are particularly challenging because the best performing model may depend on the number of locations available, the spatial scale,
or the frequency of temporal averaging (for example, daily averages versus monthly averages); see some relevant studies in Malmberg et al. (2005), Luo et al. (2008), and Lenzi et al. (2017). Spatio-temporal models are generally complex, and their likelihood-based inference is computationally expensive even for Gaussian process models (Smith and Wilen, 2003; Derado et al., 2010; Berrocal et al., 2012; Lindström et al., 2013; Krupskii and Genton, 2017).

For the analysis of non-Gaussian spatio-temporal data, we propose a new copula-based spatio-temporal (COST) model and a semiparametric maximum pseudo-likelihood estimator. Copula-based estimation methods were also considered in Chen and Fan (2006), Chen et al. (2009), and Rémillard et al. (2012), but these were developed for time series data. Compared to existing methods, the proposed approach has the following advantages. First, the COST framework creates a simplified temporal and spatial dependence structure by separating the modeling of dependence from the marginal distributions, which reduces the computational challenges. Second, instead of assuming a parametric distribution, the proposed method models the marginal distributions nonparametrically, thereby increasing the model’s flexibility. Third, it is straightforward to assess the uncertainty using the COST method. By generating multivariate random draws from the conditional prediction distribution and calculating the conditional quantiles, we can use the conditional median (which is more suitable for skewed distributions) for point forecasts/predictions and then use the conditional tail quantiles to construct prediction intervals. Finally, while in theory the COST method requires a large number of locations for spatial predictions, we find that the method performs very well even with a moderate number of irregularly spaced sites, as we demonstrate on a wind speed dataset in Section 5.

The rest of the paper is organized as follows. In Section 2, we introduce the COST model framework and the maximum pseudo-likelihood estimator of the copula parameters. In Section 3, we describe how COST can be used to construct one-step-ahead forecasts and new-location predictions. We assess the finite-sample performance of the COST method through a simulation study in Section 4 and apply the proposed method to an analysis of wind speed data in Section 5. Additional supporting information may be found online in the Supporting Information section at the end of the article.

This article is protected by copyright. All rights reserved.
2. Copula-based semiparametric model for spatio-temporal data

2.1 COST model

Let \( Y(s, t) \) be a continuous spatio-temporal process of the location \( s \in \mathcal{S} \subset \mathbb{R}^2 \) and time \( t \in \mathcal{T} \subset \mathbb{N} \), and \( \{Y(s, t), i = 1, \ldots, d, t = 1, \ldots, n\} \) be the observed data generated from this process at \( d \) locations and \( n \) time points. For notational simplicity, we denote \( Y_{i, t} = Y(s_i, t) \) and \( Y_t = (Y_{t, 1}, \ldots, Y_{t, d})^T \). We assume that the data are generated from the following copula-based spatio-temporal (COST) model.

(i) First, the model assumes that the process is \( p \)-order Markovian in time, which implies that \( P(Y_{t+p} \leq Y_{t+p-1}, \ldots, Y_1) = P(Y_{t+p} \leq Y_{t+p-1}, \ldots, Y_t) \), that is, the probabilistic properties of the process are completely determined by the joint distribution \( H(y_1, \ldots, y_{t+p}) = P(Y_{t, 1} \leq y_{t, 1}, \ldots, Y_{t, d} \leq y_{t, d}, \ldots, Y_{t+p, 1} \leq y_{t+p, 1}, \ldots, Y_{t+p, d} \leq y_{t+p, d}) \), where \( y_t = (y_{t, 1}, \ldots, y_{t, d})^T \).

(ii) Second, by Sklar’s theorem (Sklar, 1959), there exists a \((p + 1)d\)-dimensional copula \( C_{ST} \) such that

\[
H(y_t, \ldots, y_{t+p}) = C_{ST}\left\{F(y_{t, 1}|s_1), \ldots, F(y_{t, d}|s_d), \ldots, F(y_{t+p, 1}|s_1), \ldots, F(y_{t+p, d}|s_d); B_t, \theta\right\}, \tag{1}
\]

where \( F(\cdot|s) \) is the marginal cumulative distribution function (CDF) for \( \{Y(s, t), t = 1, \ldots\} \) at location \( s \), \( B_t = \{s_1, \ldots, s_d\} \times \{t, t + 1, \ldots, t + p\} \), and \( \theta \) is the copula parameter vector.

For ease of presentation, we focus on \( p = 1 \) hereafter, but the proposed method can be adapted for general \( p \) with some modification.

Remark 1: The copula in the above COST model is a special case of the spatio-temporal field copula defined in Definition 1 of the Supporting Information S1.1 with \( D = d \) and \( N = p + 1 \). The copula links the marginals together to form the joint distribution, and it allows us to model the marginal distributions \( F(\cdot|s_i) \) separately from the spatio-temporal dependence structure, which is expressed in \( C_{ST}(\cdot; B_t, \theta) \). By the COST model, data can be generated by first generating \( y_1, \ldots, y_p \) from \( C_{ST}\left\{F(y_{1, 1}|s_1), \ldots, F(y_{1, d}|s_d), \ldots, F(y_{p, 1}|s_1), \ldots, F(y_{p, d}|s_d); B_1, \theta\right\} \) and then generating \( Y_{t+p} \) sequentially from the conditional distribution \( Y_{t+p}|y_t, \ldots, y_{t+p-1} \).

2.2 Choice of spatio-temporal copula

Though any multivariate copula can be adopted in the COST model, for our purpose we need choose a copula that is convenient for specifying the spatio-temporal dependence. Two practical choices are the elliptic copula (Rémillard et al., 2012), including Gaussian.
2 COPULA-BASED SEMIPARAMETRIC MODEL FOR SPATIO-TEMPORAL

2.2 Choice of spatio-temporal copula

and Student $t$ copulas as special cases, and the $v$-transformed multivariate normal (VTMN) copula (Bárdossy and Li, 2008), which can capture the asymmetry in the tail dependence. For both copulas, the dependence can be specified through a correlation matrix $\mathbf{R}$. Considering such copulas, we can rewrite $C_{ST}(\cdot; \mathbf{B}_t, \mathbf{\theta})$ as $C_{ST}(\cdot; \mathbf{R})$, and rewrite (1) for $p = 1$ as

$$H(y_1, y_2) = C_{ST}\{F(y_{1,1}|s_1), \ldots, F(y_{1,d}|s_d), F(y_{2,1}|s_1), \ldots, F(y_{2,d}|s_d); \mathbf{R}\} = \Psi_{2d}\{\Psi_1^{-1}(u_{1,1}), \ldots, \Psi_1^{-1}(u_{1,d}), \Psi_1^{-1}(u_{2,1}), \ldots, \Psi_1^{-1}(u_{2,d}); \mathbf{R}\},$$

where $u_{t,i} = F(y_{t,i}|s_i), t = 1, 2, i = 1, \ldots, d$, and $\Psi_{2d}$ and $\Psi_1$ are the joint and marginal distributions of the $(2d)$-dimensional distribution associated with the copula, respectively. In general, the marginals of $\Psi_{2d}$ need not be the same, but here we assume that they are for simplicity. Using the Gaussian copula as an example, $\Psi_{2d}$ is the joint CDF of a $(2d)$-dimensional normal distribution with unit variance and correlation matrix $\mathbf{R}$, and $\Psi_1$ is the CDF of $N(0, 1)$; more details on the elliptic and VTMN copulas are given in the Supporting Information S1.2.

For the copula function to appropriately capture the spatio-temporal dependence, we must specify the structure of the matrix $\mathbf{R}$. In this paper, we assume second order stationarity so that the correlation between two locations only depends on the difference in their locations, not on the locations themselves. We adopt the non-separable spatio-temporal correlation matrix as in Gneiting (2002), which incorporates possible anisotropy in space (Zimmerman, 1993; Huser et al., 2017). To apply this to our wind speed data analysis, we let $h_{ij}$ denote the scaled Mahalanobis distance between $s_i$ and $s_j$; that is, $h_{ij}^2 = (s_i - s_j)^T \mathbf{V}^{-1} (s_i - s_j)$ with

$$\mathbf{V} = \begin{pmatrix} \cos(\zeta) & -\sin(\zeta) \\ \sin(\zeta) & \cos(\zeta) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \begin{pmatrix} \cos(\zeta) & -\sin(\zeta) \\ \sin(\zeta) & \cos(\zeta) \end{pmatrix}^T,$$

where $\lambda > 0$ is the length ratio of the two principal axes, and $\zeta \in [0, \pi/2]$ is the angle with respect to the west-east direction. When $\lambda = 1$, the Mahalanobis distance degenerates to the Euclidean distance. We let $X(s_i, t) = \Psi_1^{-1}[F\{Y(s_i, t)|s_i\}]$, and assume that

$$\text{corr}\{X(s_i, t), X(s_j, t + \delta)\} = \frac{1}{(1 + b_1 |\delta|^{2b_3})^{b_2}} \exp \left\{-\frac{c h_{ij}^{2\gamma}}{(1 + b_1 |\delta|^{2b_3})^{b_2}} \right\},$$

(2)

where $b_1 > 0, c > 0$ are the scale parameters of time and space, $b_3 > 0, 0 < \gamma < 1$ are the smoothness parameters, and $b_2 \in (0, 1]$ is the space-time interaction parameter. The purely spatial correlation function is of the powered exponential form, and the purely temporal correlation function belongs to the Cauchy class.

This article is protected by copyright. All rights reserved.
Since we consider Markov processes of order \( p = 1 \), we only need to consider the conditional distribution of \( Y_t \) given \( Y_{t-1} \), that is, \(|\delta| = 0,1 \) in (2), where \( \delta = 0 \) describes the variance of the process, and \( \delta = \pm 1 \) describes the covariance between adjacent time points. Define \( \eta = (1 + b_1)^{b_2} \geq 1 \). The correlation matrix \( \mathbf{R} \) can be written as

\[
\mathbf{R} = \begin{pmatrix}
R_{11} & R_{12} \\
R_{21} & R_{22}
\end{pmatrix},
\]

(3)

where \( R_{11} = R_{22} = \left(\exp(-ch_{ij}^{2\gamma})\right)_{i,j=1}^d \), and \( R_{12} = R_{21} = \left(\eta^{-1}\exp(-ch_{ij}^{2\gamma}/\eta^\gamma)\right)_{i,j=1}^d \). By Theorem 2 and Example 1 from Gneiting (2002), (2) is a stationary correlation function on \( \mathbb{R}^2 \times \mathbb{N} \); that is, \( \text{corr}\{X(s,t),X(s+h,t+\delta)\} \) only depends on \( h \) and \( \delta \) (Gneiting, 2002). Consequently, the correlation matrix defined in (3) is nonnegative definite.

Let \( \theta \) be the vector of parameters involved in the copula; for example, \( \theta = (\zeta, \lambda, c, \gamma, \eta) \) for the Gaussian copula, and \( \theta = (\zeta, \lambda, c, \gamma, \eta, \upsilon) \) for the \( t \) copula with degrees of freedom (df) \( \upsilon \). Since \( \mathbf{R} \) is determined by \( \theta \), we do not distinguish between \( \mathbf{R} \) and \( \theta \) when the context is clear. The COST framework can easily accommodate nugget effects by introducing some additional parameters in the correlation matrix; see Supporting Information S2.

### 2.3 Properties of the COST model

We present the mixing and stationary properties implied by the COST model and the condition A.1 as below.

**A.1** The spatio-temporal process \( Y(s,t) \) is induced by \( C_{ST}(\cdot; \mathbf{R}) \), where \( \mathbf{R} \) is the correlation matrix defined in (3).

**Proposition 1:** Assuming that the COST model and A.1 hold, we have (i) for elliptic and VTMN copulas, with a given \( s \), \( Y(s,t) \) is stationary and \( \beta \)-mixing with the decay rate \( \beta(t) = O\{\exp(-a_1t)\} \) for some \( a_1 > 0 \); (ii) for Gaussian and VTMN copulas, with a given \( t \), \( Y(s,t) \) is \( \alpha \)-mixing with the mixing coefficient \( \alpha(k) = O(k^{-a_2}) \) for some \( a_2 > 0 \) large enough.

Supporting Information S3.1 gives the definitions of different types of mixings, and S3.2 provides the proof of Proposition 1.

**Remark 2:** In our numerical studies, we focus on elliptic copulas as the computation for VTMN is complicated. For the Gaussian copula, the dependence between two distant locations diminishes if the correlation decays with \( h \). However, the spatial dependence of \( t \)-copula does not converge to zero as the distance increases (Røislien and Omre, 2006).

This article is protected by copyright. All rights reserved.
Despite this limitation, we still consider the $t$ copula in our numerical studies for several reasons. First, it is analytically tractable. Second, the rank correlation induced by the $t$ copula decays with $h$. More specifically, the prediction at new locations/times are based on the conditional $t$ distribution with df increasing with the number of locations, which means that locations far away from those of interest play negligible roles in the prediction, similar to standard kriging. In addition, the $t$ copula contains the Gaussian copula as a special case but provides more flexibility by leaving the df unspecified.

2.4 Semiparametric maximum pseudo-likelihood estimator of $\theta$

Let $f(\cdot|s_i)$ be the density function corresponding to $F(\cdot|s_i)$. The conditional density of $Y_t$, given $Y_{t-1} = (y_{t-1,1}, \ldots, y_{t-1,d})^T$ and evaluated at $y_t = (y_{t,1}, \ldots, y_{t,d})^T$, is

$$f_{Y_t|Y_{t-1}}(y_t|y_{t-1}) = \frac{c_{ST}(u_{t-1}, u_t; \theta)}{c_S(u_{t-1}; \theta)} \prod_{i=1}^d f(y_{t,i}|s_i),$$

where $u_i = \{F(y_{t,1}|s_i), \ldots, F(y_{t,d}|s_i)\}^T$, $c_{ST}(u, v; \theta) = \partial^2 C_{ST}(u, v; \theta) / \partial u \partial v$, $c_S(u; \theta) = \partial^d C_S(u; \theta) / \partial u$, and $C_S(u; \theta) = P(U_1 \leq u_1, \ldots, U_d \leq u_d, V_1 \leq 1, \ldots, V_d \leq 1; \theta) = C_{ST}(u, 1_d; \theta)$ is the $d$-dimensional purely spatial copula.

We propose a two-stage procedure for parameter estimation. In stage 1, at each location $s_i, i = 1, \ldots, d$, we estimate $F(\cdot|s_i)$ by the rescaled empirical CDF $\tilde{F}(y|s_i) = \frac{1}{n+1} \sum_{t=1}^n I(Y_{t,i} \leq y)$, and get the pseudo uniform variables $\tilde{u}_t = \{\tilde{F}(y_{t,1}|s_i), \ldots, \tilde{F}(y_{t,d}|s_i)\}^T, t = 1, \ldots, n$. In stage 2, we construct the pseudo likelihood based on the pseudo variables and define

$$\hat{\theta} = \arg \max_\theta \sum_{t=2}^n l(\tilde{u}_{t-1}, \tilde{u}_t; \theta),$$

where $l(u, v; \theta) = \log\{c_{ST}(u, v; \theta) / c_S(u; \theta)\}$. Following the copula literature (Joe, 1997; Rémillard et al., 2012), we refer to the proposed two-stage estimator $\hat{\theta}$ as the maximum pseudo-likelihood estimator (MPLE). Under the COST model, condition A.1, and some additional regularity conditions, we can establish the consistency and asymptotic normality of $\hat{\theta}$ as $n \to \infty$. The additional conditions and proofs can be found in Supporting Information S3.3.

3. Forecasts and predictions

3.1 Forecasts at future time points

With the estimation $\tilde{F}(\cdot|s_i), i = 1, \ldots, d$, and $\hat{\theta}$ (consequently $\hat{R}$), we can obtain a one-step-ahead forecast at time $n + 1$. For a first-order Markov series, we forecast $Y_{n+1}$ by
generating random draws from the conditional distribution of $Y_{n+1} | Y_n = y_n$. To simplify notation, denote $F(\cdot|s_i)$ by $F_i$, and $\bar{F}(\cdot|s_i)$ by $\bar{F}_i$. Below, we describe how to generate from the conditional distribution based on the Gaussian copula, following Algorithm 1 in Supporting Information S4.1; the computing details for the $t$ copula can be found in Supporting Information S4.2.

Define $X_n = (X_{n,1}, \ldots, X_{n,d})^T$ and $\tilde{x}_n = (\tilde{x}_{n,1}, \ldots, \tilde{x}_{n,d})^T$, where $X_{n,i} = \Phi^{-1}\{F_i(Y_{n,i})\}$, and $\tilde{x}_{n,i} = \Phi^{-1}\{\bar{F}_i(y_{n,i})\}$. We let $x_{n+1}^{(m)} = \{x_{n+1,1}^{(m)}, \ldots, x_{n+1,d}^{(m)}\}^T, m = 1, \ldots, M$, be random draws from $X_{n+1} | X_n = \tilde{x}_n \sim N(B\tilde{x}_n, \hat{\Omega})$, where $B = \hat{R}_{21}\hat{R}_{11}^{-1}$, $\hat{\Omega} = \hat{R}_{22} - \hat{R}_{21}\hat{R}_{11}^{-1}\hat{R}_{12}$, then the random draws of $Y_{n+1} | Y_n = y_n$ are generated by $y_{n+1}^{(m)} = \left(\bar{F}_1^{-}[\Phi\{x_{n+1,1}^{(m)}\}], \ldots, \bar{F}_d^{-}[\Phi\{x_{n+1,d}^{(m)}\}]\right)^T$ where $\bar{F}_i^{-}$ is the generalized inverse of $\bar{F}_i$. The $q$-th estimated quantile of $Y_{n+1,i} | Y_n = y_n$ is

$$
\hat{Q}_q(Y_{n+1,i} | y_n) = \bar{F}_i^{-}\left[\Phi\left(\hat{R}^{T}_{12,i}\hat{R}^{-1}_{11}\tilde{x}_n + \Phi^{-1}(q)\sqrt{1 - \hat{R}^{T}_{12,i}\hat{R}^{-1}_{11}\hat{R}^{T}_{12,1}}\right)\right],
$$

where $\hat{R}_{12,i}$ is the $i$-th column of $\hat{R}_{12}$. Thus, at each location $i$, we can construct a point forecast of $Y_{n+1}$ by using either the sample mean of the random draws, $M^{-1}\sum_{m=1}^{M} y_{n+1}^{(m)}$, or the conditional median estimator, $\hat{Q}_{0.5}(Y_{n+1,i} | y_n)$, which is more suitable for skewed distributions. We can also construct a 95% forecast interval using $(\hat{Q}_{0.025}(Y_{n+1,i} | y_n), \hat{Q}_{0.975}(Y_{n+1,i} | y_n))$.

The random draws $y_{n+1}^{(m)}$ are generated as a vector from the conditional distribution of $Y_{n+1} | Y_n$. Therefore, using the random draws for probabilistic forecasts can appropriately capture the spatial dependence of the forecasts; see numerical evidences given in Section 4.2.

### 3.2 Prediction at new locations

Let $Y_t(s^*) = (Y_{t,1}^*, \ldots, Y_{t,K}^*)^T$ be the unknown outcomes at $K$ new locations, denoted by $s^* = (s^*_1, \ldots, s^*_K)$, at time $t$. To construct the point and interval predictions, we generate random draws from the conditional distribution of $Y_t(s^*)|\{Y_t = y_t, Y_{t-1} = y_{t-1}\}$.

We demonstrate with the Gaussian copula; computation details for the $t$ copula are in Supporting Information S4.2.

Define the estimated $(2d + K) \times (2d + K)$ extended correlation matrix as

$$
\hat{R} = \begin{pmatrix}
\hat{R} & \hat{R}_* \\
\hat{R}^T & \hat{R}_{**}
\end{pmatrix},
$$

where $\hat{R}_*$ and $\hat{R}_{**}$ are the estimated correlation matrices of dimension $2d \times K$ and $K \times K$, constructed similarly as $\hat{R}$. For COST with the Gaussian copula, we have

$$
X_t^*|\{X_{t-1} = x_{t-1}, X_t = x_t\} \sim N\left(\hat{R}^T_t \hat{R}^{-1} x_{t-1,t}, \hat{R}_{**} - \hat{R}^T_t \hat{R}^{-1} \hat{R}_*\right),
$$

This article is protected by copyright. All rights reserved.
where \( x_{t-1:t} = (x_{t-1}^T, x_t^T)^T \). We denote the \( M \) random draws from the above multivariate normal distribution as \( x_t^{(m)}, m = 1, \ldots, M \), and the random draws from \( Y_t(\mathbf{g})|Y_t = y_t, Y_{t-1} = y_{t-1} \) as \( y_t^{(m)} \) as \( y_t^{(m)} = (\tilde{F}_1^{-1}[\Phi(x_t^{(m)}]), \ldots, \tilde{F}_K^{-1}[\Phi(x_t^{(m)})])^T \), where \( \tilde{F}_k^{-1}(\cdot) \) is the estimated marginal distribution of \( \{Y_{s_t}, t\} \), and \( \tilde{F}_k^{-1} \) is the corresponding generalized inverse. Then, the point prediction of \( Y_t(\mathbf{g}) \) at a single time point \( t \geq 2 \) can be constructed by \( \hat{Y}_t(\mathbf{g}) = y_t, Y_{t-1} = y_{t-1} \) as \( M^{-1} \sum_{m=1}^M y_t^{(m)} \). For an interval prediction at the new location \( s_k^* \), we derive the \( q \)-th conditional quantile of \( Y_{t,k}^* \) as

\[
\hat{Q}_q(Y_{t,k}^*|y_t, y_{t-1}) = \tilde{F}_k^{\ast\ast} \left[ \Phi \left( \hat{R}_{sk} \hat{R}^{-1}x_{t-1:k} + \Phi^{-1}(q) \sqrt{1 - \hat{R}_{sk} \hat{R}^{-1}\hat{R}_{sk}} \right) \right],
\]

where \( \hat{R}_{sk}, \tilde{R}_{kk} \) are the \( k \)-th columns of \( \hat{R}_s \) and \( \tilde{R}_k \), respectively. The 95% prediction interval can be constructed by \( \hat{Q}_{0.025}(Y_{t,k}^*|y_t, y_{t-1}), \hat{Q}_{0.975}(Y_{t,k}^*|y_t, y_{t-1}) \), and if the data are skewed, we can construct the point prediction using \( \hat{Q}_{0.5}(Y_{t,k}^*|y_t, y_{t-1}) \). To estimate \( F_k^*(\cdot) = F(\cdot|s_k^*) \), we need to use information from neighboring sites, because no data have been observed at the new locations; see estimation details and the consistency in the Supporting Information S3.4.

4. Simulation study

4.1 Simulation design and parameter estimation

The data are generated from the COST model using Algorithm 1 (Supporting Information S4.1) with

\[
\mathbf{R}_{11} = \mathbf{R}_{22} = \left( \exp \left(-c_0 h_{ij}^{2\alpha} \right) \right)_{i,j=1}^d, \quad \mathbf{R}_{12} = \mathbf{R}_{21} = \left( \frac{1}{\gamma_0} \exp \left(-\frac{c_0 h_{ij}^{2\beta}}{\eta_0} \right) \right)_{i,j=1}^d,
\]

where \( h_{ij} \) is the Mahalanobis distance between \( s_i \) and \( s_j \). We consider four cases. In all cases, \((c_0, \gamma_0, \eta_0) = (1, 0.5, 1.5)\). In Cases 1–3, \((\zeta_0, \lambda_0) = (0, 1)\), that is, \( h_{ij} = ||s_i - s_j|| \), so the true correlation structure is isotropic. In Case 4, \((\zeta_0, \lambda_0) = (0, 5)\), so the correlation structure is anisotropic. The copulas used for generating data in Cases 1–4 are Gaussian, \( t_3 \), \( t_{10} \), and \( t_3 \), respectively. The time series at each location has a Gamma marginal distribution, with location-dependent shape and scale parameters, \( \alpha(s) = 2s_1 + s_2^2 \) and \( \beta(s) = s_1 + s_2 \), respectively, where \( s_1, s_2 \) are the location coordinates. For each case, we consider 1000 replicates.

To mimic the wind speed data in Section 5, we assume that the data are observed at \( d = 9 \) fixed locations, \( \{s_1, \ldots, s_9\} = \{(1/6, 1/6), (1/6, 3/6), (1/6, 5/6), (3/6, 1/6), (3/6, 3/6), (3/6, 5/6), (5/6, 1/6), (5/6, 3/6), (5/6, 5/6)\} \), and we make predictions at this article is protected by copyright. All rights reserved.
4 SIMULATION STUDY

4.2 Forecasts at future time points

We consider the following metrics to assess the forecasting performance of different methods at time $t = n + 1$: (i) the mean squared error of the estimator of $E(Y_{n+1,j} | Y_n) \doteq \mu_j$, that is, $E(\hat{\mu}_j - \mu_j)^2$ for $j = 1, \ldots, 9$; (ii) the empirical coverage percentages (ECPs) and mean lengths (MLs) of 95% forecast intervals; and (iii) the multivariate rank. Let $y_{n+1,j}^{(0)}$ be the observed response at time $n + 1$ and location $s_j$. The multivariate rank is an assessment tool for probabilistic forecasts of vector-valued quantities (Gneiting et al., 2008). Here it measures the rank of $y_{n+1} = (y_{n+1,1}, \ldots, y_{n+1,d})^T$, the observed verifying vector at time $n + 1$ among the $M$ random draw from $Y_{n+1} | Y_n = y_n$; see Supporting Information S4.4 for the definition of multivariate rank. If the probabilistic forecasts ensemble the verifying vectors well, the multivariate rank follows a uniform distribution.

Figure 1 shows that the GP and logGP methods give worse estimation than COST$^t$, likely due to their misspecified distributions. COST$^G$ performs almost the same as COST$^t$ in Case 1, but it is less efficient when the data are generated from the $t$ copula in Cases 2–3.

Taking spatial information into account clearly improves the efficiency of COST$^t$ over its location-specific counterpart method CF. The prediction intervals from all methods have similar performance.

This article is protected by copyright. All rights reserved.
4.3 Predictions at new locations

To estimate the marginal distribution at a new location, we consider the kernel-type estimator with $m_0 = 4$ nearest neighbors, corresponding to setting $h_1 = \sqrt{2}/6$, for all four new locations; the kernel-type estimator can be found in Supporting Information S3.4. For the GP method, we estimate the mean and variance at a new location by averaging the sample means and variances of the four nearest neighbors. We do not include kriging predictions because kriging captures the prediction uncertainty based on the kriging variance, which is only a function of measurement density and spatial configuration of data points, while in COST, uncertainty estimation at the unsampled location is based on the full conditional distribution calculated as a conditional copula (Li et al., 2011). Furthermore, in our numerical analysis, the number of locations is small, so accurately characterizing the spatial dependence structure needed for kriging will be difficult, and with the large number of time points, space-time kriging is computationally infeasible.

Table 1 presents the ECP and ML of the 95% prediction intervals from the COST$_t$, GP, and logGP methods at four new locations. The prediction intervals from all methods have coverage close to the nominal level. However, the intervals from COST$_t$ are consistently

Figure 2 presents histograms of the multivariate ranks across 1000 replicates for Case 1; results for Cases 2-4 are in Figure 1 of Supporting Information S4.5. The histograms of the multivariate ranks from the COST$_G$ and COST$_t$ forecasts have nearly uniform distributions, indicating that the random draws from the proposed predictive distribution ensemble the verifying vectors well, even when the copula function is misspecified. The histograms for the CF ranks are far from uniform, as the forecasts are obtained by analyzing each location separately and thus fail to capture the spatial dependence. The GP and logGP histograms also show clear deviation from uniform, reflecting the bias in ensembles obtained from Gaussian/log-Gaussian processes for non-normal data. In practice, it can be difficult to find a unique transformation to make data from all locations normal due to the non-stationarity, making it impossible to predict at new locations.
shorter than those from GP and logGP. This, together with findings in Section 4.2, suggests that the proposed semiparametric method gives more accurate predictions and is more suitable for analyzing non-normal, spatio-temporal data than Gaussian-based methods.

5. Analysis of the hourly wind speed data

The data contain the average hourly wind speeds recorded from January 1, 2012 to December 31, 2014 at twenty meteorological towers along the border between Oregon and Washington state, a region with large wind-energy resources. The data are maintained and archived by the Bonneville Power Administration and were presented previously in Kazor and Hering (2015). To minimize the imputation required to compensate for missing data, we focus on ten sites that have less than 3.5% missing data during each of the three years. The missing values were imputed using multiple linear regression of the wind speed on neighboring horizontal and vertical wind components. The dataset contains 26,304 hourly observations for each site.

First, we perform some preprocessing. We transfer the latitude and longitude of the ten sites into a Cartesian coordinate system, assigning the Hood River, OR, site as the origin with 100-kilometer axes. Our preliminary analysis at each site showed that the wind speed distribution is highly non-normal; see Figure 3 for an example. Applying the log transformation aligns the distribution most closely with normality, but it still differs. We apply COST$_G$ and COST$_t$ methods to the untransformed wind speed data, but apply the GP approach to the log-transformed data (denoted as logGP). Following the de-trending method in Zhu et al. (2014), we model and remove the diurnal trend with the average hourly wind speed in a rolling window so the detrended data can be treated as mean-zero processes.

We find that the first-order partial correlation of the time series at each location is around 0.9, and the largest second-order partial correlation is smaller than 0.2. Therefore, for computational convenience, we assume a first-order Markov process when fitting the data for the COST$_G$, COST$_t$ and logGP methods. For the one-hour-ahead forecasts, we also include the CF method, which is the separate time series analysis method of Chen and Fan (2006). For all methods, we update the parameter estimation progressively in a rolling-window format: we incorporate the year 2014 data as it becomes available, with
a window length of 90 days, resulting in data vectors featuring \( n = 2,160 \) time points, and update the parameters in each method every 24 hours. For \( \text{COST}_t \), the estimated df ranges from 15.9 to 28.7, with mean 21.3 and standard deviation 3.0.

For short-term wind speed forecasting, Kazor and Hering (2015) proposed a regime-switching (RS) forecasting method, where the regimes are identified by a Gaussian mixture model (GMM). Within each regime, a multiple regression model is built with variables selected from among four lags for wind speed, and three lags for the cosine and sine of the wind direction across multiple locations. Compared to our proposed model that relies on only one regime and one lag for wind speed, the RS model is much more complex and data adaptive; thus, it provides excellent results for one-hour-ahead forecasts. Therefore, we use the forecast results from the RS method as a gold standard, where the median of the predictive distribution from the chosen regime is used for point forecasts.

Table 2 summarizes the mean absolute error (MAE) of different methods for one-hour-ahead point forecasts, based on hourly averages from year 2014, and the ECP and ML of the 95% prediction intervals. In terms of point forecasting, \( \text{COST}_G \), \( \text{COST}_t \) and CF methods all perform comparably to the RS method and better than logGP, and \( \text{COST}_t \) is slightly better than \( \text{COST}_G \). The forecast intervals from all four methods have coverage probabilities close to 95%, but those from logGP are generally wider. The multivariate rank histograms in Figure 4 indicate that the ensembles generated from both \( \text{COST}_G \) and \( \text{COST}_t \) capture the features of the observed spatio-temporal data well, but those from logGP and CF are either biased due to the violation of normality or failure to capture the spatial dependence.

We now assess prediction performance at the Hood River testing site. The RS and CF methods are not included as they cannot be used for prediction at new sites. In terms of point prediction, the MAE of the logGP predictions is slightly smaller than \( \text{COST}_G \) and \( \text{COST}_t \). However, the prediction intervals from logGP have much lower coverage probabilities and larger widths (ECP=73.8%, ML=15.39) than those of \( \text{COST}_G \) (ECP=94.8%, ML=14.83) and \( \text{COST}_t \) (ECP=94.9%, ML=14.48). The prediction intervals from \( \text{COST}_t \) are slightly narrower than those from \( \text{COST}_G \).

This article is protected by copyright. All rights reserved.
6. Discussion

In our numerical studies, we considered Gaussian and $t$ copulas, which are tail-symmetric copulas. The dependence in some studies may be asymmetric in the upper and lower tails. To increase the modeling flexibility, our future research may draw on ideas from spatial models developed to handle tail asymmetry, namely the factor copula models (Hua et al., 2017; Krupskii et al., 2018), vine models (Erhardt et al., 2015), and Tukey $g$-$h$ random fields (Xu and Genton, 2017). Another alternative is to adopt asymmetric copulas in our proposed procedure, for instance, the skewed $t$-copula considered in Yoshiba (2018), which is a useful extension of the $t$ copula. Extending these methods to analyze spatio-temporal data and conduct probabilistic predictions require further investigation both numerically and theoretically.

Acknowledgements

The authors would like to thank two reviewers, an associate editor, and the editor for constructive comments and helpful suggestions. The work is partially supported by King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research award OSR-2015-CRG4-2582, the National Science Foundation grant DMS-1712760, the IR/D program from the National Science Foundation, National Natural Science Foundation of China grants 11871376 and 11801355, Shanghai Pujiang Program 18PJ1409800, and Key Laboratory for Applied Statistics of MOE, Northeast Normal University 130028849. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


Figure 1. The mean squared error of $\hat{E}(Y_{n+1,j}|Y_n)$ at nine locations (top row), and the empirical coverage percentages (middle row) and mean lengths (bottom row) of 95% forecast intervals. COST$_G$: line with solid circle (blue), the proposed method assuming a Gaussian copula; COST$_t$: solid (black), the proposed method assuming a $t$ copula; GP: line with triangle (green), the method assuming a Gaussian process; logGP: line with open circle (purple), the method assuming a Gaussian process after log transformation; CF: line with diamond (red), a separate time series analysis based on Chen and Fan (2006) with a $t$ copula. The shaded areas in the first row correspond to the 95% pointwise confidence band for the mean squared error of the COST$_t$ estimator. This figure appears in color in the electronic version of this article, and color refers to that version.
Figure 2. Histograms of multivariate ranks from Case 1. $\text{COST}_G$ and $\text{COST}_t$: the proposed method assuming Gaussian and $t$ copulas, respectively; CF: a separate time series analysis based on Chen and Fan (2006) with a $t$ copula; GP: the method assuming a Gaussian process; logGP: the method assuming a Gaussian process after log transformation. This figure appears in color in the electronic version of this article.
Figure 3. Histograms of the untransformed and log-transformed hourly wind speeds observed at the Biddle Butte site in 2014. This figure appears in color in the electronic version of this article.
Figure 4. Histograms of multivariate ranks of nine sites over 8760 hours determined by the rolling-window method from the wind speed data analysis; for each hour, 500 ensembles were used to calculate the multivariate rank. COST\textsubscript{G}: the proposed method with a Gaussian copula; COST\textsubscript{t}: the proposed method with a \textit{t} copula; CF: a separate time series forecast based on Chen & Fan (2006) with a \textit{t} copula; logGP: the method based on a Gaussian process assumption, applied to the log-transformed data. This figure appears in color in the electronic version of this article.
### Table 1

Empirical coverage percentage (ECP) and the mean length (ML) of the 95% prediction interval at four new locations.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Method</th>
<th>s₁</th>
<th>ML</th>
<th>s₂</th>
<th>ML</th>
<th>s₃</th>
<th>ML</th>
<th>s₄</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>COSTₜ</td>
<td>95.5</td>
<td>1.23</td>
<td>96.7</td>
<td>2.09</td>
<td>96.8</td>
<td>2.29</td>
<td>96.3</td>
<td>3.33</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>94.1</td>
<td>1.29</td>
<td>94.4</td>
<td>2.15</td>
<td>95.9</td>
<td>2.38</td>
<td>95.1</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>logGP</td>
<td>94.6</td>
<td>2.31</td>
<td>96.7</td>
<td>2.81</td>
<td>94.5</td>
<td>2.70</td>
<td>96.5</td>
<td>3.74</td>
</tr>
<tr>
<td>2</td>
<td>COSTₜ</td>
<td>93.8</td>
<td>1.12</td>
<td>96.3</td>
<td>1.92</td>
<td>96.7</td>
<td>2.14</td>
<td>96.1</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>93.2</td>
<td>1.27</td>
<td>94.1</td>
<td>2.12</td>
<td>95.1</td>
<td>2.34</td>
<td>93.8</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td>logGP</td>
<td>93.8</td>
<td>2.29</td>
<td>95.6</td>
<td>2.93</td>
<td>94.1</td>
<td>2.70</td>
<td>93.3</td>
<td>3.80</td>
</tr>
<tr>
<td>3</td>
<td>COSTₜ</td>
<td>94.6</td>
<td>1.15</td>
<td>97.0</td>
<td>2.01</td>
<td>95.9</td>
<td>2.23</td>
<td>98.0</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>94.7</td>
<td>1.27</td>
<td>94.4</td>
<td>2.12</td>
<td>95.3</td>
<td>2.35</td>
<td>94.6</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td>logGP</td>
<td>93.5</td>
<td>2.27</td>
<td>94.3</td>
<td>2.89</td>
<td>95.9</td>
<td>2.63</td>
<td>95.6</td>
<td>3.75</td>
</tr>
<tr>
<td>4</td>
<td>COSTₜ</td>
<td>94.7</td>
<td>1.50</td>
<td>97.6</td>
<td>2.55</td>
<td>97.4</td>
<td>2.86</td>
<td>96.7</td>
<td>4.12</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>95.3</td>
<td>1.65</td>
<td>94.7</td>
<td>2.75</td>
<td>95.1</td>
<td>3.05</td>
<td>93.9</td>
<td>4.41</td>
</tr>
<tr>
<td></td>
<td>logGP</td>
<td>95.4</td>
<td>3.58</td>
<td>95.7</td>
<td>4.16</td>
<td>94.4</td>
<td>3.66</td>
<td>93.8</td>
<td>5.15</td>
</tr>
</tbody>
</table>

COSTₜ: the proposed method assuming a t copula; GP: the method assuming a Gaussian process; logGP: the method assuming a Gaussian process after log transformation.
The mean absolute error (MAE) and corresponding standard error, empirical coverage percentages (ECPs), and mean lengths (MLs) of 95% forecast intervals, for the one-hour-ahead 2014 wind forecasts from different methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>BB</th>
<th>FG</th>
<th>HH</th>
<th>ME</th>
<th>NR</th>
<th>RO</th>
<th>SH</th>
<th>SU</th>
<th>TI</th>
</tr>
</thead>
<tbody>
<tr>
<td>COST\textsubscript{G}</td>
<td>MAE</td>
<td>0.908</td>
<td>0.502</td>
<td>0.979</td>
<td>0.888</td>
<td>0.915</td>
<td>1.026</td>
<td>1.041</td>
<td>1.499</td>
<td>0.655</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.010)</td>
<td>(0.005)</td>
<td>(0.011)</td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.015)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>COST\textsubscript{t}</td>
<td>MAE</td>
<td>0.895</td>
<td>0.500</td>
<td>0.977</td>
<td>0.882</td>
<td>0.908</td>
<td>1.204</td>
<td>1.038</td>
<td>1.494</td>
<td>0.654</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.010)</td>
<td>(0.005)</td>
<td>(0.011)</td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.015)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>CF</td>
<td>MAE</td>
<td>0.841</td>
<td>0.512</td>
<td>1.000</td>
<td>0.881</td>
<td>0.877</td>
<td>1.194</td>
<td>1.049</td>
<td>1.485</td>
<td>0.679</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.009)</td>
<td>(0.005)</td>
<td>(0.011)</td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.015)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>logGP</td>
<td>MAE</td>
<td>1.109</td>
<td>0.507</td>
<td>1.027</td>
<td>0.936</td>
<td>0.983</td>
<td>1.337</td>
<td>1.061</td>
<td>1.585</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.013)</td>
<td>(0.005)</td>
<td>(0.012)</td>
<td>(0.009)</td>
<td>(0.010)</td>
<td>(0.013)</td>
<td>(0.011)</td>
<td>(0.016)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>RS</td>
<td>MAE</td>
<td>0.739</td>
<td>0.496</td>
<td>0.934</td>
<td>0.839</td>
<td>0.819</td>
<td>1.133</td>
<td>1.024</td>
<td>1.433</td>
<td>0.652</td>
</tr>
</tbody>
</table>

COST\textsubscript{G}: the proposed method with a Gaussian copula; COST\textsubscript{t}: the proposed method with a $t$ copula; CF: a separate time series forecast based on Chen and Fan (2006) with a $t$ copula; logGP: the method based on a Gaussian process assumption, applied to the log-transformed data; RS: the GMM.r.s method from Kazor and Hering (2015). Values in parentheses are the corresponding standard errors of the mean absolute errors; the standard errors of RS were not reported in Kazor and Hering (2015). BB: Biddle Butte; FG: Forest Grove; HH: Horse Heaven; ME: Megler; NR: Naselle Ridge; RO: Roosevelt; SH: Shaniko; SU: Sunnyside; TI: Tillamook.