

RESEARCH ARTICLE

A multivariate framework to study spatio-temporal dependency of electricity load and wind power

Swasti R. Khuntia^{1,2}  | Jose L. Rueda² | Mart A.M.M. van der Meijden^{2,3}

¹ Department Intelligent and Autonomous Systems, Centrum Wiskunde and Informatica, Amsterdam, The Netherlands

² Department Electrical Sustainable Energy, Delft University of Technology, Delft, The Netherlands

³ TenneT TSO B.V., Arnhem, The Netherlands

Correspondence

Swasti R. Khuntia, Department Intelligent and Autonomous Systems, Centrum Wiskunde and Informatica, Science Park 123, Amsterdam 1098 XG, The Netherlands.
Email: swastigunu@gmail.com

Funding information

FP7 Energy, Grant/Award Number: 608540; European Union Seventh Framework Programme (FP7/2007-2013), Grant/Award Number: 608540

Abstract

With massive wind power integration, the spatial distribution of electricity load centers and wind power plants make it plausible to study the inter-spatial dependence and temporal correlation for the effective working of the power system. In this paper, a novel multivariate framework is developed to study the spatio-temporal dependency using vine copula. Hourly resolution of load and wind power data obtained from a US regional transmission operator spanning 3 years and spatially distributed in 19 load and two wind power zones are considered in this study. Data collection, in terms of dimension, tends to increase in future, and to tackle this high-dimensional data, a reproducible sampling algorithm using vine copula is developed. The sampling algorithm employs *k*-means clustering along with singular value decomposition technique to ease the computational burden. Selection of appropriate clustering technique and copula family is realized by the goodness of clustering and goodness of fit tests. The paper concludes with a discussion on the importance of spatio-temporal modeling of load and wind power and the advantage of the proposed multivariate sampling algorithm using vine copula.

KEYWORDS

dependence modeling, electricity load, multivariate model, spatio-temporal modeling, vine copula, wind power

1 | INTRODUCTION

In accordance to the Paris Agreement 2016,¹ the move from conventional energy (fossil fuel and nuclear power) towards renewable energy (RE) is adopted globally. As a result, massive RE in-feed to the existing electric grid network at both transmission and distribution level is noticeable and is further increasing. Trying to tap more of RE into the existing grid is however challenging, pertaining to its irregular availability, variability, and link to varying atmospheric factors. To give an example, in an attempt to increase the utilization of RE, it is understood that investments in wind power plants (WPPs) are concentrated at locations with higher-average wind speeds and solar farms at locations with higher-average solar insolation. The transmission system operators (TSOs) have to cautiously evaluate operation as well as future planning when power output fluctuations occur for such spatially distributed systems. An accurate knowledge of spatial and temporal characteristics is thus beneficial to model the behavior of the power system under different RE penetration.

Wind generation is driven by wind patterns, which tend to follow certain geographical spatial correlations. Till date, modeling of wind power focused on a single wind farm (or aggregation of wind power in single WPP). In this manner, they do not account for potential information from neighboring sites, for example, other WPPs or meteorological stations. Spatial modeling is vital when wind power error in a WPP might propagate to WPPs in other locations during the following period when they are affected by the same meteorological factors. As a massive integration of wind power is witnessed in Europe and the United States, considering inter-spatial dependence along with temporal correlation is important

for wind power modeling. Load modeling follows the same path, and to understand, we consider aggregated values of load. While forecasting load, it is probable to obtain the forecast between the maximum and minimum of the historical values. Between the two extrema, there might be some hot or cold days that went unnoticed in the training set. But, if we include the historical data from the closest load zone, there is a high possibility that the hot and cold days are taken into consideration. This is because electricity load is affected by weather parameters, and hence, the energy usage of adjacent load zone is also taken into the new training set. So, it is evident that incorporating inter-spatial dependence can help improve the modeling accuracy, and thus, the trend has been towards exploiting all of the available data in modeling. And, the first step is to obtain a tractable model that captures the uncertainties and correlations among the exogenous variables*.

Before we dive into details, it is to be noted that this study

- is performed at the transmission level and assumed that WPPs are connected to the transmission system;
- aims at short-term operational planning purposes (eg, hourly resolution data is considered in this study);
- excludes solar power and other distributed energy resources that are prevalent more at the distribution level;
- excludes the interaction of transmission and distribution system operators.

The key contributions of this paper are as follows:

- It reviews the advancements in spatio-temporal modeling of electricity load and wind and the challenges associated with multivariate modeling.
- A first-hand application of vine copula for spatio-temporal modeling with multivariate framework is presented, ie, a reproducible sampling algorithm is developed using C-vine to model the spatio-temporal dependency.
- It is intended that the future power system will be data-centric, and data collection from stochastic sources in terms of spatial and temporal resolution will result in a high-dimensional database of varied features. To tackle this and ease the computational burden, our multivariate framework is first-of-its-kind to employ clustering and feature extraction technique.

The rest of the paper is organized as follows: Section 2 presents a background of univariate and multivariate frameworks in spatio-temporal modeling. Section 3 explains concept of copula and vine copula in terms of spatio-temporal modeling. The developed framework is detailed in Section 4 and simulation results in Section 5. Finally, Section 6 concludes the paper.

2 | REVISITING SPATIO-TEMPORAL STUDY OF ELECTRICITY LOAD AND WIND POWER

2.1 | Brief background on spatio-temporal study of load and wind power as univariate models

The spatio-temporal features of electricity load can be explained by two underlying spatio-temporal processes, namely, weather and human activities.² The weather of adjacent neighborhoods or cities tends to be more alike than those far apart. Similarly, human activities in adjacent neighborhoods tend to be highly correlated. Electricity load has a long-anticipated factor because of its very strong seasonality feature, ie, daily, weekly, and monthly. This is referred to as seasonal or temporal component. Because of the fact that electricity load has seasonal or temporal component, most previous studies aimed at modeling temporal correlation while overlooking spatial correlation among load variation in different zones. Since weather plays an important part, weather-based variables that affect load can differ according to location. Considering spatio-temporal aspect of electricity load will result in more accurate modeling, it will be possible to realize extreme values beyond a fixed target load from the training spatio-temporal dataset, which results from different load profiles in different zones. Most of spatio-temporal studies focused at distribution level where the target is residential homes.^{3,4} A dynamic spatio-temporal model was developed in² taking Southern California's electricity load time series. However, no study has been reported yet for spatio-temporal modeling at the transmission level.

Stochasticity of wind makes it difficult to predict accurate wind power output by only considering temporal wind behavior when it is affected by other geographical and technical factors like wind farm topology and wind turbine characteristics.⁵ As such, it is a common belief among researchers that for the spatial pattern, the dependence is relatively strong for elements that are closer to each other.⁶ Literature study reveals that inter-spatial dependence and temporal correlation were studied separately until recently.⁷⁻¹⁰ A spatial study of wind power in different zones in the United Kingdom was studied in Miranda and Dunn¹¹ using a multivariate regression model. The study showed a multivariate time-series model for real wind speed data from multiple sites is complicated in nature, particularly the presence of a large number of wind sites. Maisonneuve and Gross¹² proposed a wind regime model for planning studies. The study aimed at modeling both seasonal and diurnal trends of wind power and its correlation to the same trends of electricity load. In the temporal aspect, transformation and standardization of non-Gaussian and

*Meaning that they are determined by someone else than the TSO, and the TSO will have to adapt its behavior accordingly.

nonstationary characteristics of wind power are studied in the previous studies^{13,14} by application of regression models. The use of copula for spatio-temporal scenario studies is reported in Tastu et al,¹⁵ and more discussion is followed in Section 3. Papavasiliou et al¹⁶ addressed spatial correlation for wind power modeling using a noise vector-based regression model in an attempt where a single-multivariate time-series model is decoupled into distinct univariate time-series models.

2.2 | Need for multivariate framework for spatio-temporal study

It is to be noted that meaningful correlation exists between load and wind power because both are significantly affected by weather. There is an immediate need for the development of spatio-temporal modeling of load and wind power as joint normal distribution for three reasons. Firstly, inter-spatial dependence and temporal correlation of load and wind power in any considered site are important. From Section 2.1, the literature study revealed consideration of load and wind power as independent variables and also some instances of temporal or spatial correlation. However, there are no significant findings that investigate the spatio-temporal dependence of these two exogenous variables. Secondly, a suitable spatio-temporal modeling approach will facilitate improving both short-term operational planning and long-term grid development of power grids. To give an example, in short-term operational planning, an accurate spatio-temporal modeling can help the TSO in assessing system security in terms of asset overloading or reducing operational costs by using forecast values for unit commitment or reducing wind curtailment. Similarly, in terms of long-term planning, accurate modeling can result in assessing grid development plans to answer the load growth or massive generation of wind power. For example, in this study, the control area of US regional transmission operator is considered (more details about the dataset follow later in Section 3). As of 2016, 1GW of installed wind power along with other generation sources serve large load centers along the east coast and mid-western region.¹⁷ Under the renewable portfolio standard, 11.3 % (32GW) and 13.9 % (42GW) of the total load are expected in the years 2021 and 2026 in the form of massive integration.¹⁸ A map of future wind farm projects is shown in Figure 1.¹⁹ Lastly, a suitable spatio-temporal multivariate model can generate a rich synthetic database of normally distributed load and wind power data. Such a database will be of immense help to research community and industry as well to develop other statistical tools. Examples of online tools to visualize the most promising areas where wind farms can be profitably installed are IRENA global atlas²⁰ and NREL's wind prospector.²¹ However, the data extracted from these wind atlases lack the clarity when they are used to learn the behavior of power system operation in shorter time horizons (15 min or 1 h).

3 | SPATIO-TEMPORAL MODELING USING VINE COPULA

A suitable spatio-temporal model should be able to capture both inter-spatial dependence and temporal correlation embedded in the multivariate dataset. To realize such a model, the modeling framework can be divided into three steps:

- Modeling one-dimensional marginal distributions;
- Modeling stochastic dependence using copula;
- And spatio-temporal modeling using vine copula.

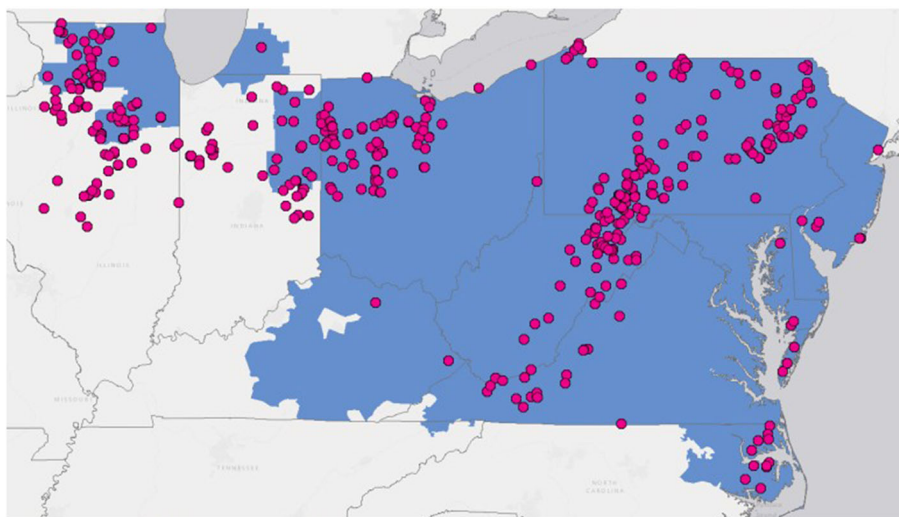


FIGURE 1 Map of future wind farm projects under one of the US regional transmission operator (blue shade is control area, and red dots are planned wind farms) [Colour figure can be viewed at wileyonlinelibrary.com]

In a multivariate framework, modeling stochastic dependence is a challenging task because the variables within the multivariate framework do not always have standardized marginal distributions. To answer such challenge, one solution is that the dependence between multiple correlated random variables can be captured by different measures of dependence. For general multivariate random variables, Spearman's rank correlation coefficient can be used to study the nonlinear, monotonic relationship between two random variables.²² It is the nonparametric statistical measure used to study the strength of association between the two ranked variables and helps in defining the dependence structure based on rank with specific functions, called copula functions. Copulas, Latin word for "bond" or "couple," are functions that couple the multivariate distribution functions to their marginal distribution functions and, therefore, describe the dependence structure between these random variables.²³ Using copula functions, it is possible to simulate two random variables that are correlated according to rank correlation by first simulating a copula and later transforming the obtained ranks into respective marginals. The advantage is that the joint distribution function is built based on two independent tasks comprising the modeling of the dependence and the modeling of the marginal distribution functions. The use of copula is not new in the field of electric power systems. Literature study reveals the use of Gaussian copulas to evaluate short-term scenarios for wind power generation,^{24,25} wind power forecasting error,¹⁰ transmission network planning,²⁶ and empirical copulas for modeling the dependence structure between the wind speed and the wind power output.²⁷ By definition, two random variables P and Q with CDF_P, CDF_Q are joint by copula C if their joint distribution CDF_{PQ} can be written as²⁸

$$CDF_{PQ}(p, q) = C(CDF_P(p), CDF_Q(q)). \quad (1)$$

The function C is therefore defined on uniformly random variables, and the $CDFs$ can be used to map the uniformly random variables to P and Q . Let $P = \{P_1, P_2, \dots, P_n\}$ be an n -dimensional random vector with a continuous marginal $CDF \{F_1, F_2, F_3, \dots, F_n\}$. The relationship between $CDF F$ of P is written as

$$F(P) = C(F_1(P_1), F_2(P_2), \dots, F_n(P_n)) \quad P \in R^n, \quad (2)$$

where unique function $C : [0,1]^d \rightarrow [0,1]$ is called the copula. A function $C : [0,1]^n \rightarrow [0,1]$ is called an n -dimensional copula if it satisfies the following conditions²⁹:

- i. $C(u_1, \dots, u_n)$ is increasing in each component u_i .
- ii. $C(u_1, \dots, u_{k-1}, 0, u_{k+1}, \dots, u_n) = 0$ for all $u_i \in [0,1], i \neq k, k = 1, \dots, n$.
- iii. $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ for all $u_i \in [0,1], i = 1, \dots, n$.
- iv. For all $(a_1, \dots, a_n), (b_1, \dots, b_n) \in [0,1]^n$ with $a_i \leq b_i$,

$$\sum_{i_1=1}^2 \dots \sum_{i_n=1}^2 (-1)^{i_1 + \dots + i_n} C(p_{1i_1}, \dots, p_{ni_n}) \geq 0, \quad (3)$$

where $p_{j1} = a_j$ and $p_{j2} = b_j$ for all $j \in \{1, \dots, n\}$. Condition (iii) follows from the fact that the marginals are uniform in the range $[0,1]$ and that condition (iv) is true because the sum in Equation (3) can be interpreted as $P[a_1 \leq p_1 \leq b_1, \dots, a_n \leq p_n \leq b_n]$, which is non-negative. For further understanding, it is advised to follow Patton,³⁰ which reviews the use of copulas in econometric modeling and for an elaborate bibliometric overview of copulas.³¹

However, copula estimation is tricky in higher dimensions. The selection of an appropriate copula function is very important, as inappropriate selection can lead to unacceptable errors. Of all copulas, the Gaussian copula is the most commonly used copula because of its computational convenience. However, for this study, a more comprehensive approach is adopted by first testing a number of standard copulas on multivariate dataset as presented in Louie.³² The flexibility with copula is encountered with a bottleneck and that is they perform better for bivariate distributions and that the individual pattern of chosen random variables must be described by the same parametric family of univariate distributions. Furthermore, multivariate copulas are neither good. Although the number of parametric multivariate copula families with flexible dependence is limited, there are many parametric families of bivariate copulas that can be modeled as a vine. Hence, vine copulas are preferred as they allow a more flexible dependence structure.

Vines are a representation of high-dimensional copulas that are constructed from a sequence of nested bivariate copula components called "pair-copulas." They are flexible because any combination of bivariate copulas can be used for the pair-copulas. Vine copula models decompose a multivariate copula into a set of bivariate copulas, and each bivariate copula can be described as a branch of a graph connecting two consecutive marginal distributions or their conditional bivariate distributions. This is a more practical way to represent high-dimensional copula problems, thus, allowing to build a model that is aware of separating distances across space and time. To achieve this, the building blocks of vine copula are composed out of convex combinations of bivariate copulas. The weights of the convex combination as well as the copulas' parameters are defined by the distance over space and time, thus modeling spatial and temporal correlation. This explains the motivation to use vine copula for spatio-

temporal modeling of non-Gaussian datasets, where the non-Gaussianity refers not only to marginal distributions at one location but also to the dependence structure between locations.

Vine copula follows a nested tree structure with edges and nodes as seen in Figure 2. By definition, a vine copula on n variables is a nested set of trees T_j , where the edges of the j^{th} tree become the nodes of the $(j+1)^{st}$ tree for $j = 1, \dots, n$. In general, vine decompositions are referred to as regular vines. A regular vine on n variables is defined as a vine in which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node. Each edge in the regular vine may be associated with a conditional rank correlation and a copula and each node with a marginal distribution. All assignments of rank correlations to edges of a vine are consistent, and each one of these correlations may be realized by a copula. Based on the bivariate and conditional bivariate distributions, the joint distribution can be constructed. The use of vine copulas to tackle power system uncertainty is reported in the previous studies³³⁻³⁵ and probabilistic forecast for multiple wind farms in Wang et al.³⁶

A regular vine can be decomposed to either

- i. D (drawable)—vine where each node in T_j has a degree of at most 2, and conditioning is done sequentially; or
- ii. C (canonical)—vine in which each tree T_j has a unique node of degree $n - i$, where the first variable is used as a conditioning variable for the following ones.

In D -vine, conditioning is performed sequentially, whereas in C -vine, the first variable is used as conditioning variable for the following ones. Figure 2 shows a graphical representation to represent the joint density decomposition using C -vine. In general, a C -vine copula selects a root node in each tree, and all pair-wise copulas connecting with this node are modeled and conditioned on all of the previous root nodes. The nodes of tree 1 correspond to marginal distribution functions while each edge corresponds to pair-copula density given as $C_{2,3|1}$ in tree 2. The notation means that the copula model between variables 2 and 3 is conditional on 1. The full density of this C -vine copula is given by

$$C(U_1, \dots, U_6) = C_{5,6|1234}(U_{4,5|123}, U_{4,6|123}) C_{4,5|123}(U_{3,4|12}, U_{3,5|12}) \cdot C_{4,6|123}(U_{3,4|12}, U_{3,6|12}) C_{3,4|12}(U_{2,3|1}, U_{2,4|1}) \cdot C_{3,5|12}(U_{2,3|1}, U_{2,5|1}) \cdot C_{3,6|12}(U_{2,3|1}, U_{2,6|1}) C_{1,2}(U_1, U_2) \cdot C_{1,3}(U_1, U_3) \cdot C_{1,4}(U_1, U_4) \cdot C_{1,5}(U_1, U_5) \cdot C_{1,6}(U_1, U_6). \tag{4}$$

The conditioned variables $U_{i|v}$, $i \in \{2, \dots, 6\}$, and $v \in \{\{1\}, \{1, \dots, 4\}\}$ are derived through the copulas in the preceding tree (eg, from tree 1):

$$u_{i|1} := F_{i|1}(u_i|u_1) = \frac{\partial C_{1i}(u_1, u_i)}{\partial u_1} \text{ at } u_1, i \in \{2, \dots, 6\} \tag{5}$$

Such a modeling scheme is appropriate for modeling complex multivariate dependence structures with mixed types of dependencies, such as asymmetries and tail dependencies, since each pair-copula can belong to a different parametric copula function.

To understand the spatio-temporal modeling using vine copula model, a spatio-temporal random field H is considered such that

$$H: S \times T \times \Phi \rightarrow \mathbb{R}, \tag{6}$$

where S corresponds to a spatial domain, T corresponds to temporal domain, and both with an underlying probability space Φ . For a section of the spatio-temporal random field defined as $H = (h(s_0, t_0), h(s_1, t_1) \dots h(s_n, t_n))$ of size $n+1$, the section consists of one pivotal location and its n -neighbors

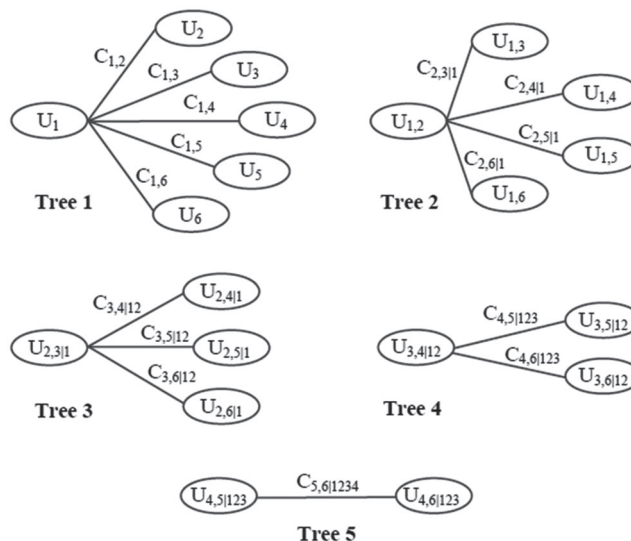


FIGURE 2 C-vine of CDF ($u \in [0,1]$) with five trees

in distinct spatio-temporal locations $(s_0, t_0), (s_1, t_1) \dots (s_n, t_n) \in S \times T$. Normally, some spatial locations would be sampled at multiple time instances. And as the dependence structure changes over space and time, the first section of the vine is realized by spatio-temporal bivariate copulas. The rest of the vine, ie, the vine of the variables conditioned under the value of the central location, is modeled as a n -dimensional C-vine. To understand the functional capability of C-vine, Figure 3 shows the spatio-temporal n -dimensional C-vine copula. The temporal extension of the spatial copula at different time lags for three spatial locations with Euclidean distance defined as $h_E := \|s_i - s_j\|$, $s_i \forall i, j \in \{0, 1, 2, 3\}$ & $t_C = 1 \dots n$. This n can represent a mix of load centers and WPPs, and essentially, we follow the same methodology. Every curved connection in Figure 3 is modeled by the same spatio-temporal copula C_{h_E, t_C} but with different spatial and temporal distances h_E and t_C deduced from the indicated spatio-temporal locations. It is already assumed that marginals are stationary and combining them with multivariate copula results in a multivariate distribution of the spatio-temporal random field. And this multivariate distribution is later used for application studies like simulation or prediction.

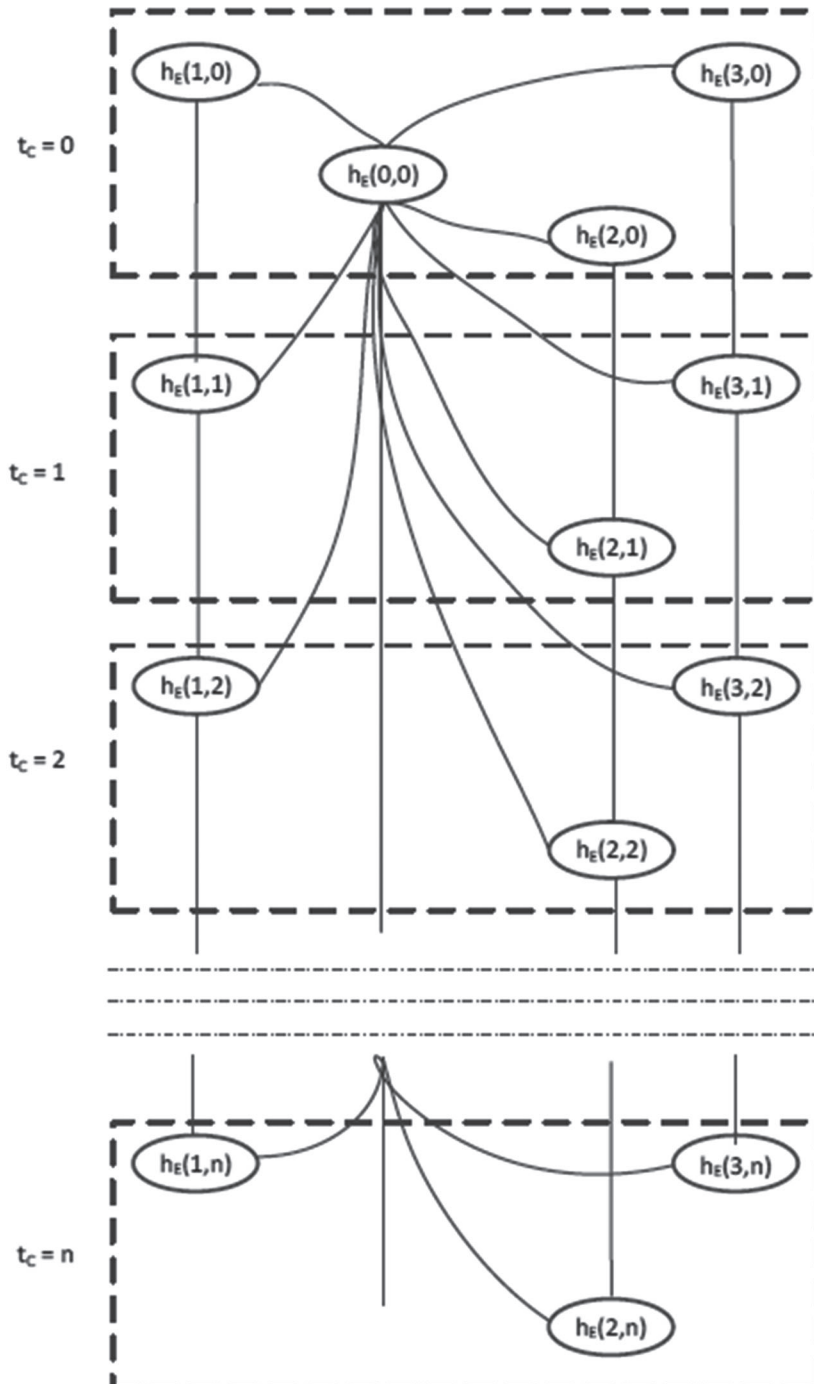


FIGURE 3 Spatio-temporal n -dimensional C-vine copula

4 | MODELING FRAMEWORK AND ASSESSMENT BASED ON REAL DATA

This section describes the preprocessing steps for the real data followed by the sampling algorithm. Algorithm 1 explains the developed algorithm step-by-step, and each step of the algorithm is further explained in subsections. All computation is performed in MATLAB (version 2017b) environment on an Intel Core i7 with 8 cores and 8-GB RAM.

Algorithm 1 Spatio-temporal modeling for high-dimensional data using vine copula

Inputs: High-dimensional dataset of size $(M \times N)$, representing M data points and N features.

Outputs: $S \times N$ dimensional sampled dataset

1 Perform clustering to partition the high-dimensional data. k number of clusters are selected after performing GoC test on sample size S

2 Feature extraction of k clusters using singular value decomposition (SVD)

3 Calculate copula function and construct vine copula models for k clusters. Choice is different copula functions can be tested and GoF test is performed to select the best copula function

4 Simulate the copula function for k clusters using cluster weight obtained in Step 1

5 Reconstruct dataset from low to high-dimension with all features using eigenvectors from Step 2

6 Output as $S \times N$ dimensional sampled dataset

7 **End**

4.1 | Inputs

For this study, publicly available load and wind power data are taken from one of the US regional transmission operator.³⁷ Aggregated zonal load (19 numbers) and wind power (2 numbers) data spanning 3 years with hourly resolution is used in this study and is shown in Figure 4. The three market regions are *MIDATL*, *WEST*, and *SOUTH*, and a detailed composition of these regions with load and wind power zones is shown in Table 1. The load and wind power from each zone are described by a distinct time series corresponding to a distinct position in space defined as the *weighted centroid*. Such a weighted centroid is required to calculate the spatial correlation using the geographical coordinates of zones. Since the exact coordinates are treated confidentially by utilities, an approximated weighted centroid is defined in this study to locate an approximated “center” of load zones and wind power generation zones. A detailed explanation of the latitudes and longitudes to calculate the approximated weighted centroid is available in Khuntia.³⁸ The weighted centroid approach is able to describe the approximate dependencies between zones and a more realistic relationship is determined by the actual size of the zone. It can be argued that the resulting dependency will be affected

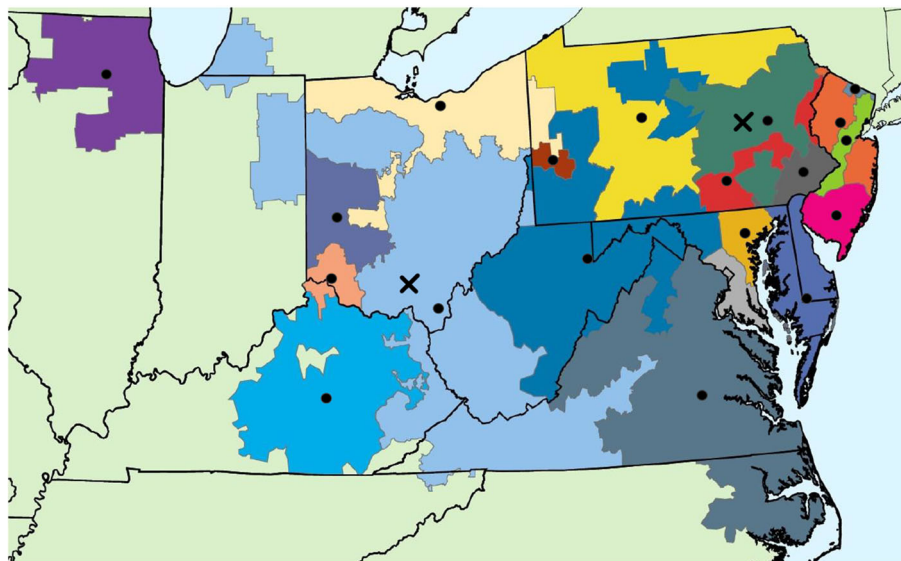





















FIGURE 4 Map showing the control areas of PJM with load and wind power zones' approximated weighted load centroid (•) and wind power centroid (X) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Load and wind power zones as seen in Figure 4

Market Region	Load Zone	Wind Power Zone
MIDATL	AE 	MIDATL
	BC 	
	DPL 	
	JC 	
	ME 	
	PE 	
	PL 	
	PN 	
	PS 	
	RECO 	
WEST	AEP 	WEST
	AP 	
	ATSI 	
	CE 	
	DAY 	
	DEOK 	
	DUQ 	
	EKPC 	
SOUTH	DOM 	-

by such aggregated data and approximated weighted centroid approach. However, such an approach still provides valuable information about dependencies.

To visualize the complexity, scatter plot with marginal histograms of four load zones (*AP*, *CE*, *DAY*, and *DUQ*) and one wind power zone (*WEST*) under the market zone *WEST* is shown in Figure 5. The marginal histograms (in the diagonal) reveal non-Gaussian nature while the scattered plots reveal the nonlinear dependencies and also suggest a weak correlation. This does not mean lack of relationship but rather a lack of linear relationship. In such cases, the marginal distributions do not conform to normality assumption, and the dependencies between the variables are misleading too. The presence of nonlinear dependencies is valid for dependence studies between individual load and wind power, between different load zones and even between the output of two wind power zones.

The m load zones (ten *MIDATL*, eight *WEST*, and one *SOUTH*) and n wind power zones (one *MIDATL* and one *WEST*) for t time length (hourly resolution with a horizon of 3-y) are written as

$$\begin{bmatrix} L_1(t_1) & \cdots & L_m(t_1) & W_1(t_1) & \cdots & W_n(t_1) \\ L_1(t_2) & \cdots & L_m(t_2) & W_1(t_2) & \cdots & W_n(t_2) \\ \cdots & \ddots & \cdots & \cdots & \ddots & \cdots \\ L_1(t_i) & \cdots & L_m(t_i) & W_1(t_i) & \cdots & W_n(t_i) \end{bmatrix}, \quad (7)$$

where N refers to total measurements ($m+n$). For spatio-temporal modeling, we first visualize spatial correlation for the original data. The correlation calculation uses hourly interval aggregated load on individual load zone for the years 2014 to 2016. The correlation coefficient for A and B is calculated using covariance normalized, written as

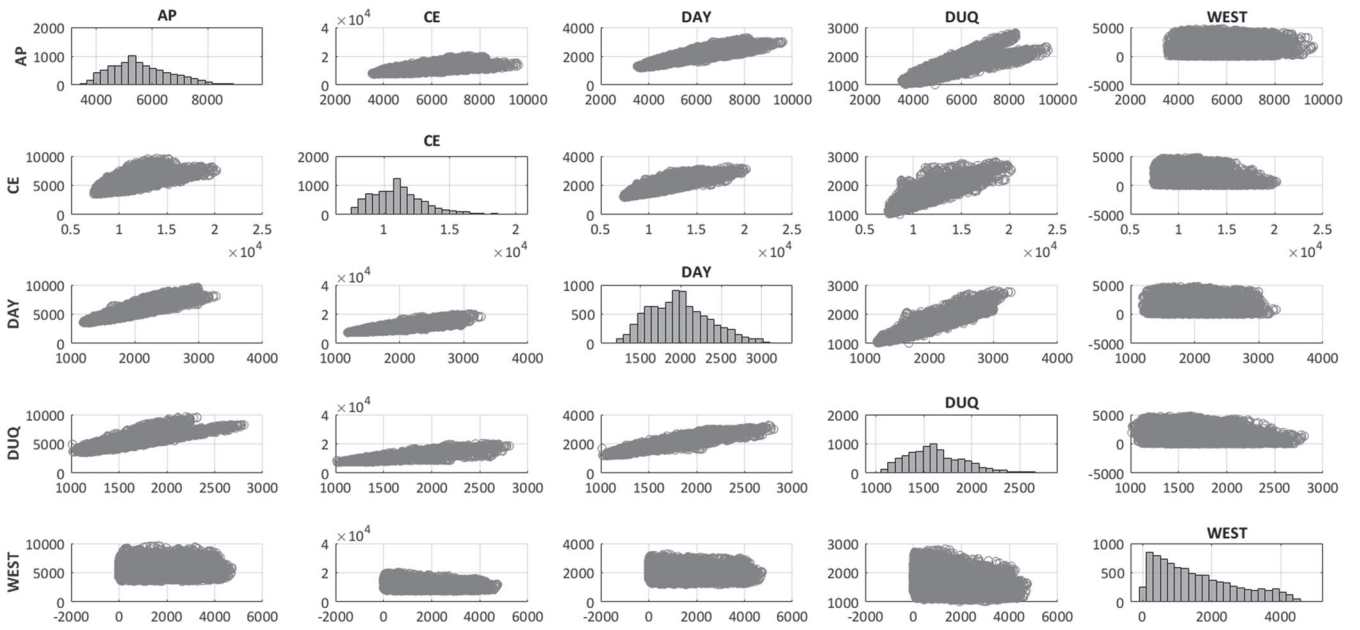


FIGURE 5 Scatter plot with marginal histograms of original data of four load zones and one wind power zone

$$\rho_{AB} = \frac{\text{Cov}(A, B)}{\sigma_A \sigma_B}, \quad (8)$$

where, σ_A, σ_B are the standard deviation and $\text{Cov}(A, B)$ is the covariance of A and B .

To account for temporal correlation, the time series is checked for seasonality. As an example, the original time series from load zone AE of MIDATL for the year 2014, shown in Figure 6, is studied. A closer look in Figure 7 reveal the peaks corresponding to weekly trend. It is understood that electricity load shows daily and weekly periodicity. Thus, the data need to be differenced at both 24 and 168 lags, which is performed for the rest of load time series. Backward differencing is normally used, and the 24th and 168th difference address the periodicity. For a time series y_t , the transformation is written as

$$\Delta_{24} \Delta_{168} y_t = (1 - L^{24})(1 - L^{168}) y_t, \quad (9)$$

where Δ is the difference operator, and L is the lag operator. After the lag operator polynomials $((1 - L^{24})(1 - L^{168}))$ are created, both are multiplied to get the desired lag operator polynomial. The differenced and deseasonalized time series for zone AE of MIDATL is shown in Figure 8. Similarly,

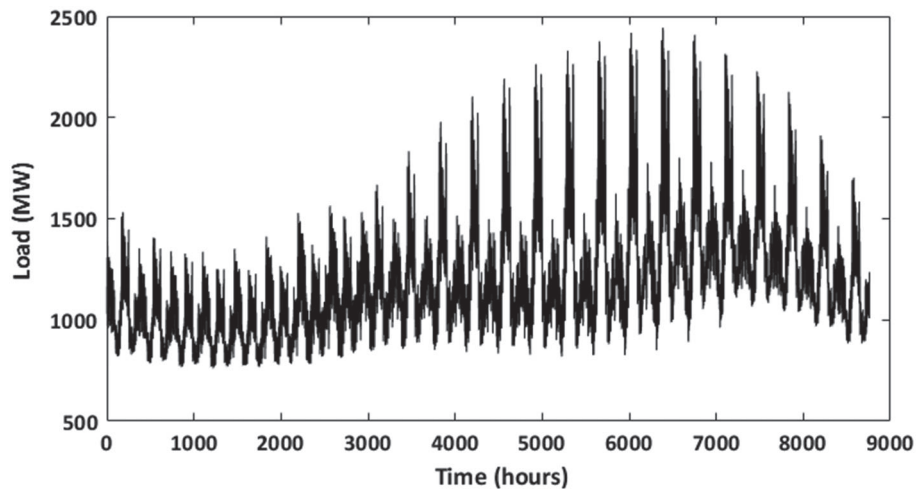


FIGURE 6 Original load time series with hourly resolution (year 2014 and zone AE)

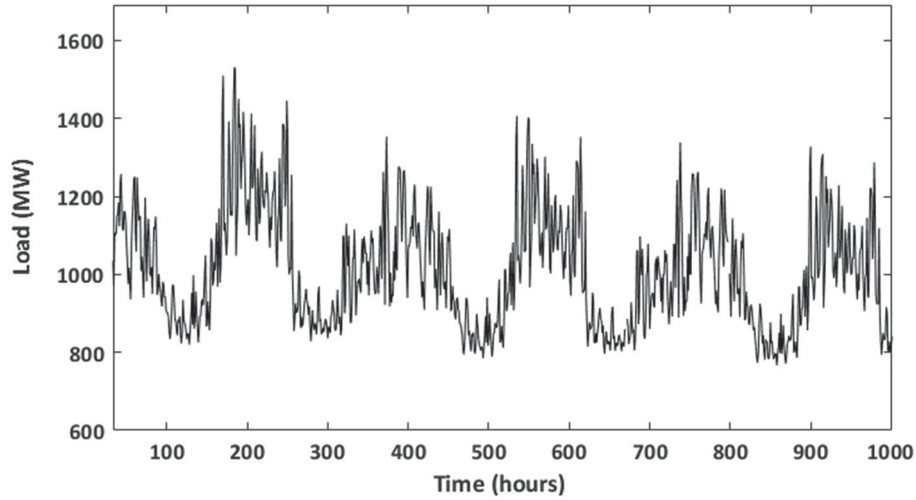


FIGURE 7 Original load time series showing the weekly periodicity (year 2014 and zone AE)

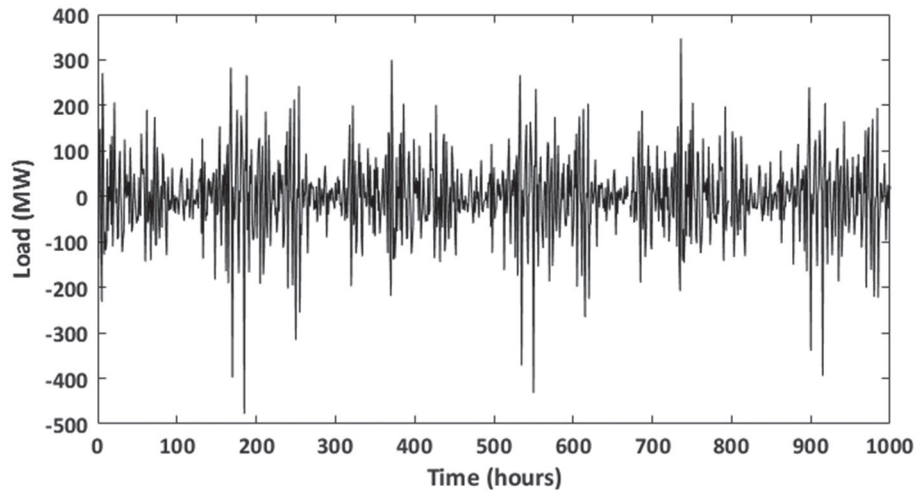


FIGURE 8 Load time series of zone AE after seasonal differencing at lags 24 and 168

temporal correlation is checked for wind power data. While analyzing wind power time series in Figure 9 and zoomed Figure 10, it shows distinctive seasonal and diurnal patterns. After checking for lags, seasonal differencing is performed for 24 lags by modifying Equation (9) as

$$\Delta_{24}y_t = (1 - L^{24})y_t. \quad (10)$$

As the multivariate dataset is preprocessed with seasonal differencing to remove the periodicity, the second step is data normalization. Normalization serves the purpose of bringing the multivariate variables into the same scale. For the load data, Z-score scaling is introduced to standardize the data for each zone as represented in Shi et al.² Z-score scaling is the most commonly used method, and it converts all indicators to a common scale with an average of zero and standard deviation of one. However, the wind power was normalized with respect to the installed wind capacity by comparing each of the datasets from different zones. Thus, the normalized wind power (W_{norm_i}) for each zone and hour i is calculated as

$$W_{norm_i} = \frac{W_i}{Cap_{inst}}, \quad (11)$$

where W_i is the actual wind power produced for hour i , and Cap_{inst} is the installed wind capacity of the zone.

The preprocessing steps result in a multivariate dataset, which is normalized and is free of any trend and seasonality. To alleviate computational burden for the high-dimensional dataset, the sampling procedure starts with data clustering and followed by feature extraction to reduce data

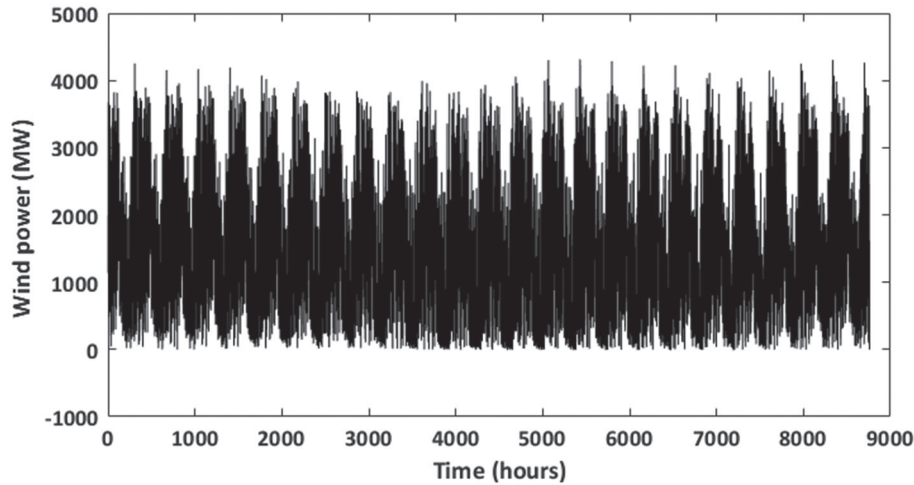


FIGURE 9 Original wind power time series with hourly resolution (year 2014 and zone WEST)

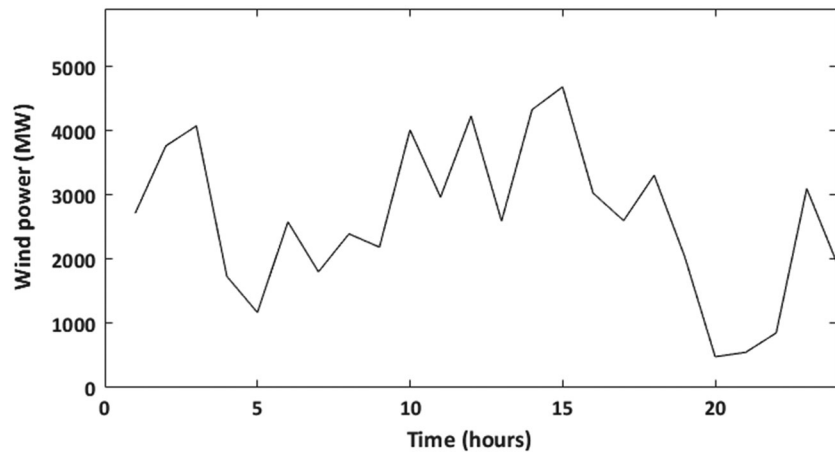


FIGURE 10 Original wind power time series for 24-hour duration (year 2014 and zone WEST)

dimensionality. Since we have mixed numerical data with different values based on location, these values are not really comparable anyway, and hence, we normalized to give equal weight to them. Normalizing the data improves convergence of clustering algorithms.³⁹ The idea is that if different components of data (features) have different scales, then derivatives tend to align along directions with higher variance, which leads to poorer/slower convergence.⁴⁰ Such an approach eases the problem as the aim is extracting the required features when the number of clusters is unknown.

4.2 | Step 1 (data clustering)

Both electricity load and wind power generation patterns are determined by different drivers depending on time and location as both vary with respect to time and space. Such varied and high-dimensional data often come with many highly correlated variables and succeeding in selecting all variables in a group of correlated variables can be very difficult. Clustering helps in partitioning the M -data points into groups of similar statistical characteristics or k clusters. The aim of data clustering is to discover the “natural” group(s) of a set of patterns in a multivariate dataset. As clustering algorithms use a distance measure of some sort to determine if object i is more likely to belong to the same cluster as object j than the same cluster as object k . These distance measures are affected by the scale of the variables. That is, when computing the distance between two objects, each with a length and a weight, the distance will change dramatically if you change the units. By putting all variables into the same range, we weigh the variables equally. In fact, the use of cluster analysis is widespread in any discipline that involves a study of multivariate data. For an overview of different clustering algorithms, readers can refer to the previous studies,^{41,42} where the different clustering algorithms are discussed. In this study, we will examine three widely used clustering algorithms used in analysis of multivariate datasets: k -means, Gaussian Mixture Model (GMM), and Hierarchical Linkage (HL) algorithms.

k -means is the most widely used unsupervised clustering technique as it is easy to understand and implement.⁴³ k -means clustering aims at partitioning M -data points into k -clusters, where each data point belongs to the cluster with its nearest mean. The k -means clustering works as an objective function, where the aim is to minimize a squared error function,

$$\min \sum_{j=1}^k \sum_{i=1}^M \|x_{i(j)} - c_j\|^2, \quad (12)$$

where $\|x_{i(j)} - c_j\|^2$ is a chosen distance measure between a data point $x_{i(j)}$ and the cluster center c_j . In contrast to simplicity, k -means has problems in discovering clusters that are not spherical in shape. It also encounters some difficulties when different clusters have a significantly different number of points. Since it is a minimization function, k -means also requires a good initialization to avoid getting trapped in a poor local minimum. It makes a number of assumptions about the data, and it does not search through every possible partitioning of the data; hence, it was opted to test GMM and HL techniques. GMM is a clustering algorithm to estimate probability function by using a finite linear combination of Gaussian model in which the weights of each Gaussian component are defined as a prior probability of each component. The concept of prior probability in context of expectation-maximization (EM) algorithm helps to select the right number of clusters.⁴⁴ The optimal model parameters are obtained by EM algorithm, and a step-by-step explanation of GMM clustering is available online.⁴⁵ GMM clustering technique uses the probability of a sample to determine the feasibility of it belonging to a cluster.

Compared with both k -means and GMM, HL clustering technique builds clusters incrementally. The clustering technique begins by assigning each sample to its own cluster (top level), and at each step, the two clusters that are most similar are merged. It continues until all of the clusters have been merged. In comparison with k -means, there is no need to specify a k parameter as one can navigate the layers of hierarchy to see which number of clusters is optimum. In addition, k -means clustering is usually more efficient run-time wise compared with GMM and HL clustering since k value is usually specified.

After deciding the clustering algorithms, choosing the right number of clusters is important to validate the chosen clustering algorithm. To select the right number of clusters, GoC test is performed using Davies-Bouldin index (DBI) and gap statistics index (GSI). A detailed explanation about the GoC tests is included in Khuntia.³⁸ k values ranging $\{2, \dots, 10\}$ were assessed using DBI and GSI statistics to choose the number of clusters. DBI quantifies the average similarity between the chosen number of clusters.⁴⁶ In theory, it is desirable for the clusters to be as distinct from each other as possible, and hence, the clustering technique, which minimizes the DBI value, is the ideal one for GoC test. Lower values of DBI correspond to better clustering validity. The results from GoC test for DBI is shown in Figure 11A. From the figure, values of $\{k = 2, 2, 2\}$ are obtained for the three clustering techniques. The second GoC test is GSI, which compares the within-cluster dispersion to its expectation under

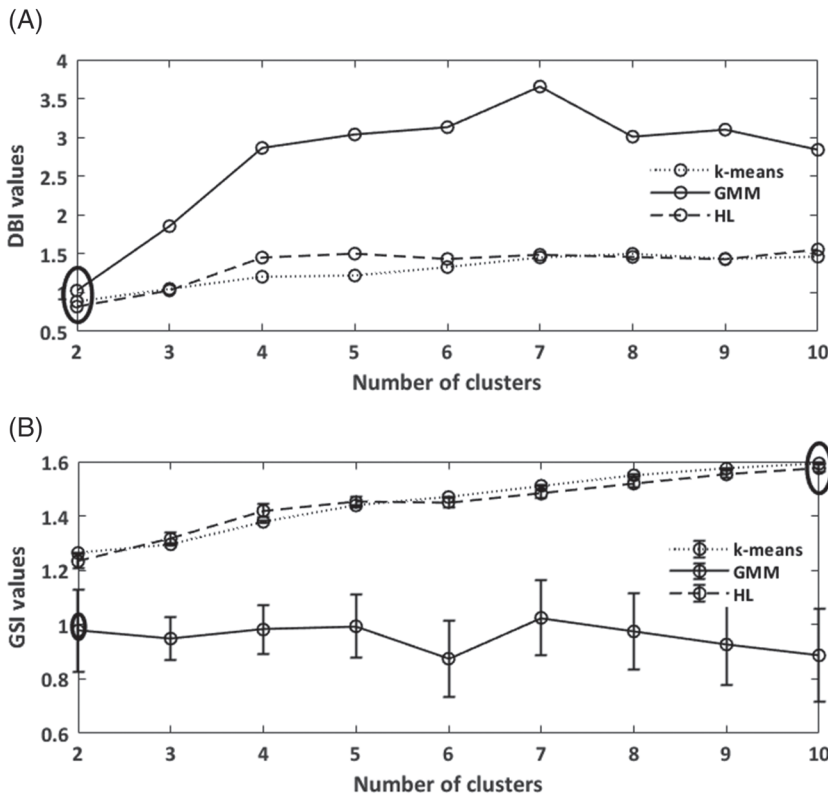


FIGURE 11 GoC test and corresponding (A) DBI values (B) GSI values

an appropriate null reference distribution.⁴⁷ Figure 11B shows the GoC test and corresponding GSI values. Gap statistics is maximized at $k = 10$ for k -means and HL clustering. For GMM, gap statistics is maximized at $k = 2$.

GoC test is required to carefully select the number of clusters k . If k is too small, the GMM is unable to well capture the distribution of data, (especially when $k = 1$, GMM will degenerate to Gaussian maximum likelihood). On the other hand, if k is too large, except for computation complexity, a severe overfitting problem will result. Based on the GoC tests, $k = 2$ is chosen in this study to generate a large sample size of $S = 30\,000$ from the preprocessed historical dataset. Such a large sample size is chosen to guarantee a good accuracy of the estimated value. Clustering results in sorting the multivariate dataset (X) into homogeneous clusters ($X^k \subset X$, for $k = 1, 2$), which are strongly related to each other and, thus, provide similar information. Next step is feature extraction from the clusters, which selects a small subset of actual features and remove redundant features.

4.3 | Step 2 (feature extraction)

In spatio-temporal modeling, feature extraction in the form of dimension reduction is reasonable given that the true spatio-temporal feature often exists on a lower dimensional structure.⁴⁸ As the name suggests, dimensionality reduction is the process to transform a high-dimensional dataset into a low dimensional space, while retaining most of the useful information from the original data. The principle behind such a transformation is that the useful information in the original high-dimensional dataset can be represented by a small number of features. Generally, in a high-dimensional space, the data points do not spread-out randomly but, rather, in a certain structure that can be easily exploited. Thus, dimensionality reduction can circumvent this problem by reducing the number of features in the dataset before the training process. Doing this reduces the computation time, and the resulting features in low dimension take less space to store as well as avoids overfitting. Other advantages of dimensionality reduction are easy interpretation and visualization, because of the low-dimensional space. However, if not performed correctly, there are high chances that dimensionality reduction will result in information loss. And there is no way to extract the lost information from low-dimension to high-dimensional space. Law⁴² classifies dimensionality reduction into three types: *feature selection and feature weighting*, *feature extraction*, and *feature grouping*. For this study, feature extraction using singular value decomposition (SVD) is chosen. SVD is a computational method often employed to calculate principal components for a dataset. The set of principal components for each cluster will represent the original data, and they are determined by computing the eigenvalues and eigenvectors of the corresponding correlation matrix. Before performing feature extraction, the clustered observations using k -means from the original domain are transformed to the rank-uniform domain via the empirical cumulative distribution function (ECDF). Thus, each clustered dataset X^k is transformed to Y^k in the $[0, 1]^N$ domain. Such a transformation helps in reducing the sensitivity of feature extraction techniques.

SVD is employed to perform principal component analysis since it is efficient and numerically robust.⁴⁹ Given an arbitrary $p \times q$ matrix $P \in \mathbb{R}^{p \times q}$, then there exist matrices U and V (both with orthogonal columns), and positive numbers $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ (where $r = \min(p, q)$), such that

$$P = \sum_{k=1}^r \sigma_k U_k V_k^T = U D V^T \quad (13)$$

with U_k and V_k denoting the k^{th} column of U and V , respectively, and D is a $p \times q$ matrix for which the numbers σ_k (the singular values) are placed on the main diagonal and are arranged in descending order: $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$. For a proof, see Wall et al.⁴⁹ For a given matrix X , we use the notation $\sigma_i(P)$ or $\lambda_i(P)$ to denote the i -th (ordered) singular or eigenvalue, respectively. If there is no danger of confusion, the explicit reference to the matrix will be suppressed. Recall that there is a useful relationship between the singular values of a matrix $P \in \mathbb{R}^{p \times q}$ and the eigenvalues of the related matrices PP^T and $P^T P$:

$$\sigma_i(X) = \sqrt{\lambda_i(PP^T)} = \sqrt{\lambda_i(P^T P)}, \quad (14)$$

where $i = 1 : \min(p, q)$. This connection will be used extensively in the analysis below.

SVD determines an optimal linear transformation: $y = Ax$, for p -dimensional data x into another q -dimensional transformed vector y for each cluster k . The linear transformation matrix A is optimal from maximal information retention criterion (IRC) viewpoint, given as

$$IRC = \frac{\sum_{j=m+1}^p \lambda_j}{\sum_{i=1}^p \lambda_i}. \quad (15)$$

This q -dimensional transformed vector ($q < t$) defines the reduced number of variables. For each cluster k obtained in Step 1, PCA is performed in three steps, (a) *centralize the data and compute the mean*, (b) *then generate scatter matrix and compute eigenvalues (λ) and eigenvectors (m)*, and (c) *project the data comprising principal components*. While the eigenvectors represent the principal components of the clustered dataset, the eigenvalues indicate the total variance accounted for by each principal component. The descending order of q -dimensional transformed vector y allows for straightforward feature extraction as well as dimensional reduction by discarding elements with lowest information content. Thus, feature

extraction on each Y^k results in low-dimensional dataset $\mathcal{H}^k \in \mathbb{R}^{t^k \times q}$, where t^k represents the number of observations in cluster k . It is to be noted that eigenvectors are later used in Step 5 (Section 4.6) in the resampling step.

4.4 | Step 3 (vine copula construction)

For each cluster, we have extracted the features based on which vine copula will be constructed. The next step is describing a dependence structure and selecting a bivariate copula family for each edge in the vine as well as estimating its parameters. In vine copula modeling, the obtained observations in Step 2 should be fitted in the uniform domain. Such a uniform transformation is achieved by ECDFs of \mathcal{H}^k to obtain C^k . Families of bivariate copula, namely, Gaussian copula, Student's t copula, asymmetric Clayton, and its corresponding 90° , 180° , and 270° rotational copula types are tested. The first step in constructing a C -vine starts with selecting a root node, which is achieved by generating Kendall rank correlation matrix and adding the correlations across each location with respect to other locations.⁵⁰ The location with the highest value of Kendall rank correlation coefficient is chosen as root node followed by other nodes in the tree.

After selecting the root node, estimating the conditional copulas in the tree is performed. To select the appropriate copula, GoF test is performed to check the copula that can be rejected. The GoF test is used for assessing whether a generated ECDF is suitable to describe a dataset or not. The null hypothesis for the GoF test states that the data are sampled from a normal distribution. When the P value is greater than the predetermined critical value, the null hypothesis is accepted, and thus, we conclude that the data fit well or is normally distributed. Two GoF tests used in this study are Kolmogorov-Smirnov (K-S) and Cramer-von Mises (CvM) test. Both the tests are based on estimated ECDF. The idea of using ECDF in testing the normality of data is to compare the ECDF, based on the data with the CDF of the normal distribution, to see if there is a good agreement between them. The K-S test is a nonparametric test and is used to check if a sample comes from a hypothesized continuous

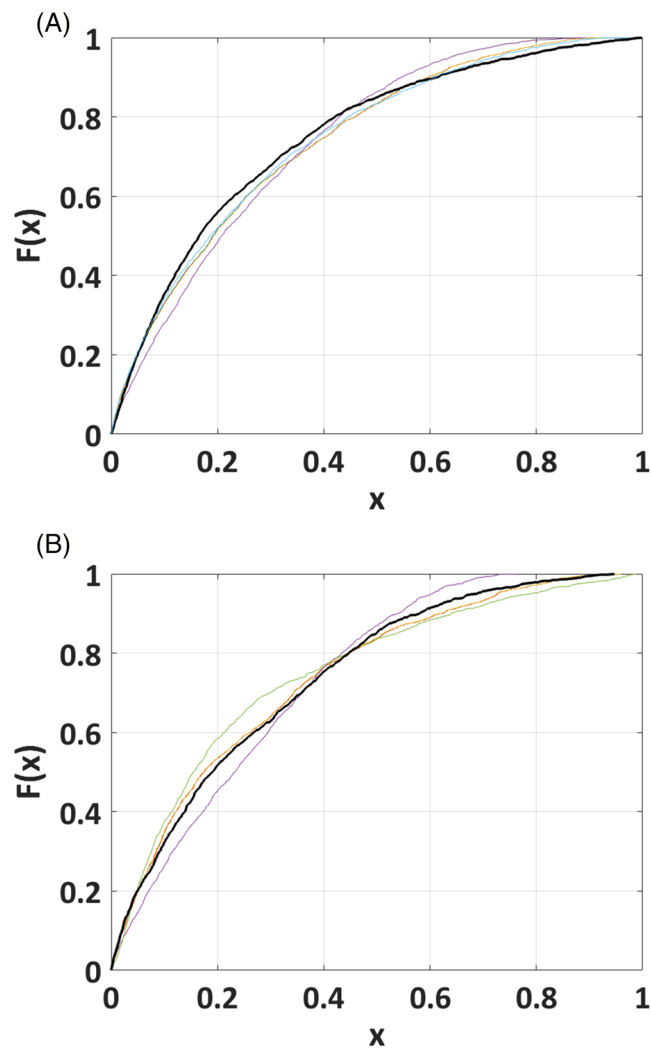


FIGURE 12 ECDF for estimated copulas (black line represents the selected copula) [Colour figure can be viewed at wileyonlinelibrary.com]

distribution. The K-S statistic (D) is written as⁵¹

$$D = \sup_x |F_h(x) - F_o(x)|, \quad (16)$$

where $F_h(x)$ is the ECDF of hypothesized distribution, and $F_o(x)$ is the ECDF of observed distribution. A powerful and refined version of K-S test, called the Cramer-von Mises (CvM) test, is also used in this study. The CvM statistic (ω^2) is written as⁵¹

$$\omega^2 = \int_{-\infty}^{\infty} [F_h(x) - F_o(x)]^2 dF_o(x). \quad (17)$$

Following the GoF test, ECDFs of different copula selection for each cluster at one branch of tree is shown in Figure 12. For $k = 1$, P value is .8950 (K-S test) and .8765 (CvM test), and 270° rotated Clayton copula is selected. Similarly, for $k = 2$, P value is .2441 (K-S test) and .2422 (CvM test), and 180° rotated Clayton copula is selected. A high P value indicates that it is a good fit since the acceptable level is greater than or equal to .05. This is repeated for each bivariate copula in rest of the vine.

4.5 | Step 4 (vine copula simulation)

Step 4 operates in coordination with step 3 in terms of vine copula simulation. For the chosen sample S , each parametric cluster-model generates samples C^k of size $t_s^k \times q$, where $t_s^k = S \times W^k$, and W^k is the weight of cluster k . Construction of C-vine is based on W^k obtained in Step 1 (Section 4.2). Once the model has been stated and estimated, a key question is to check whether the initial model assumptions are realistic. Again, a GoF test is performed on C-vine for each cluster. The chi-square (χ^2) test is used to graphically represent GoF. Likewise, K-S and CvM tests, the chi-

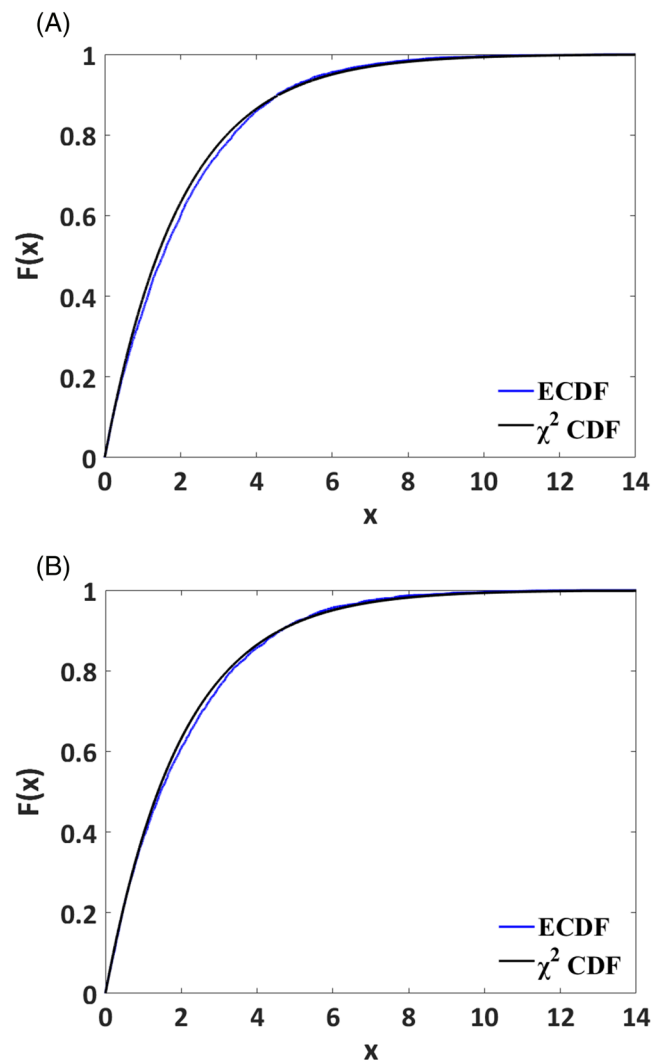


FIGURE 13 ECDF of K-S test and chi-square (χ^2) CDF plot for (A) $k = 1$, and (B) $k = 2$ showing the GoF in vine copula simulation [Colour figure can be viewed at wileyonlinelibrary.com]

square test is used to test if a sample of data came from a population with a specific distribution. At one branch of the tree, for $k = 1$, P value is .9290 (K-S test) and .9122 (CvM test), and for $k = 2$, P value is .1145 (K-S test) and .1095 (CvM test) is calculated. A graphical comparison of GoF for K-S test and chi-square test is shown in Figure 13.

4.6 | Step 5 (resampling)

Performing inverse ECDF transformation ($ECDF^{-1}$) on the sampled output to retrieve high-dimensional data is achieved in this step. In the first step, samples of each cluster \mathcal{C}^k are transformed back to the domain of \mathcal{H}^k of size $t_s^k \times q$ by transforming \mathcal{C}^k through $ECDF^{-1}$ of original dataset \mathcal{C}^k . The second step is transforming \mathcal{H}^k to high-dimensional space $\mathbb{R}^{t_s^k \times N}$, denoted by the dataset Y^k . And the last step involves the transformation of Y^k in the $[0,1]^N$ domain to original dataset X^k through $ECDF^{-1}$. In the end, the high-dimensional sampled dataset is $\bar{X} = \bar{X}^1 \cup \bar{X}^2 \cup \dots \cup \bar{X}^k \in \mathbb{R}^{S \times N}$, where \bar{X}^k corresponds to the sampled dataset for cluster k .

In order to evaluate the performance of C-vine sampling, two-sample K-S test is employed. The null hypothesis for such a test is to check if the historical and simulated dataset for each variable in the multivariate dataset if they are drawn from the same marginal distribution. The two-sample K-S test statistic quantifies a distance between the empirical distribution functions of two samples. Theoretical description of the two-sample K-S test is included in Khuntia.³⁸ A resampling method is employed to randomly generate comparison samples from the historical and sampled datasets based on Sun et al.³³ For each of the 21 variables, 200 data points from the historical dataset and 400 data points from the sampled dataset were drawn in random. And the process was iterated for 500 times. Thus, the total number of times the K-S test performed is (500×21) times. Figure 14 shows the ECDF plot against the P values calculated from K-S test. In the test, reference dataset is the historical dataset plotted against itself, which has a uniform distribution. And the sampled dataset from the sampling procedure is also perfectly uniform, aligning to reference dataset P values. Thus, we can conclude that the sampled dataset does not suffer from information loss.

5 | SIMULATION RESULTS

Figures 15 and 16 show the sampled output of load zone AE and wind power zone WEST, respectively. A visual inspection of Figure 16 reveals the presence of negative data points, which is explained by the fact that wind power is treated as negative load in the sampling procedure. As in Figure 6, pair-wise comparison of histograms and scatter plots for the marginal distributions of sampled dataset for the same load and wind zones is shown in Figure 17. The histograms (diagonal) of the figure show the normally distributed marginals with heavy tail characteristic while the scattered plots reveal nonlinear dependence between the variables. Now, we compare the spatial correlation plot for the load zones in the sampled dataset with respect to the original data for zones MIDATL and WEST and shown in Figure 18. For the sake of visual comparison, the correlation color map is adjusted to the same scale ($\rho = [0,1]$). While Figure 18C accounts for more positive correlation, Figure 18D reveals some very strong correlation and some zero correlation. Understanding the variety in spatial correlation in WEST is understood by the long separation (distances) between the load zones.

To further check the dependency in the sampled output, correlation coefficients are calculated using Equation (8) and included in Table 2. Figure 19 shows the overall spatial correlation including all load and wind power zones: Figure 19A for original dataset and Figure 19B for sampled

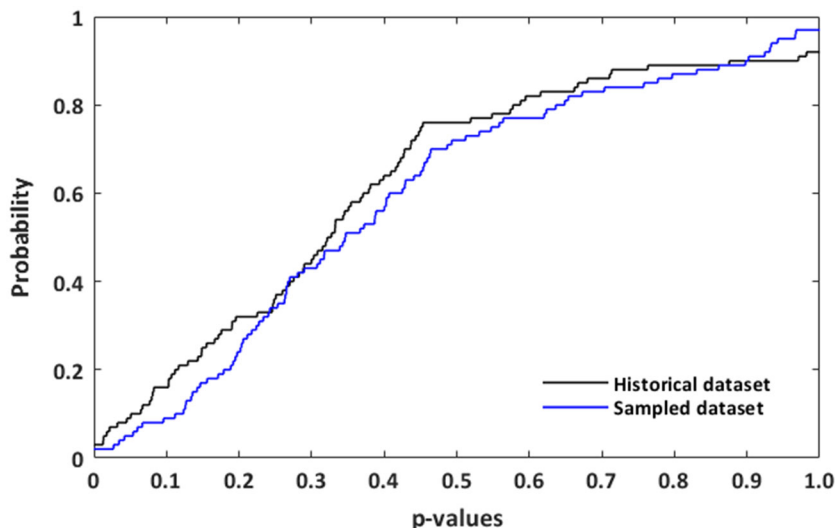


FIGURE 14 Two-sample K-S test for historical and sampled dataset [Colour figure can be viewed at wileyonlinelibrary.com]

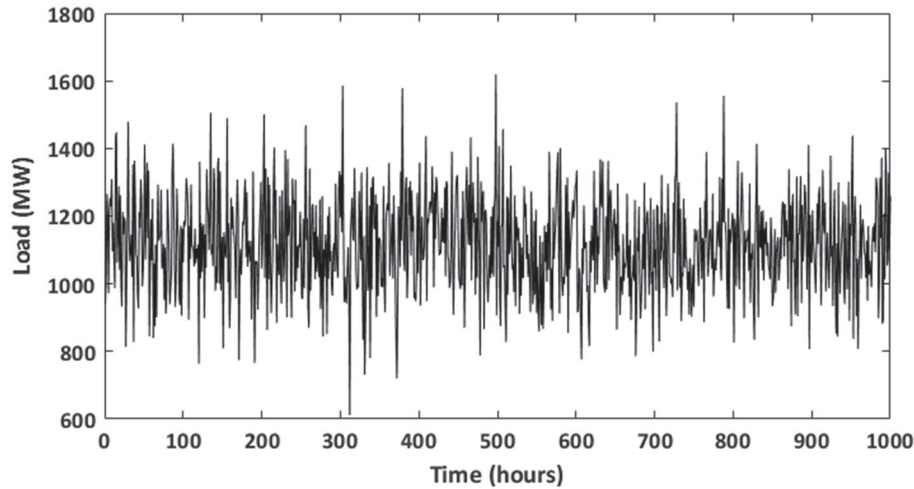


FIGURE 15 Sampled load time series with hourly resolution of zone AE

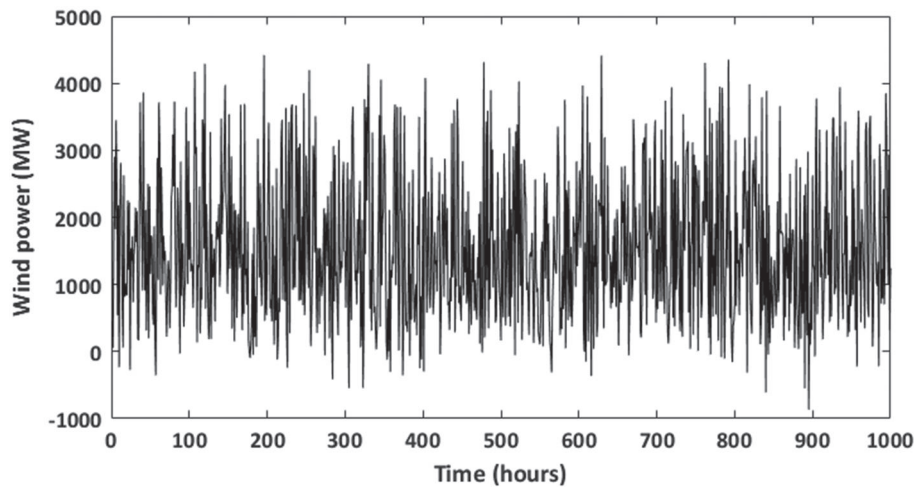


FIGURE 16 Sampled wind power time series with hourly resolution of zone WEST

dataset. Inferencing the correlation coefficients suggests negative, zero, and positive correlation. Negative correlation means that high values of one dataset are matched with low values of the other while positive correlation means that high values of one dataset are always matched with high values of the other. While minimum variance is obtained for maximum negative correlation ($\rho = -.2675$), maximum variance is seen in case of maximum positive correlation ($\rho = 1$). This matching has a direct impact on the behavior of the sum: Negative correlation prevents extreme values from happening at the same time, while positive correlation urges coincidence of extreme events. In terms of physical significance, a positive correlation between load and wind power explains the fact that both tend to increase and decrease at the same time, thereby facilitating load following task of the power system. On contrary, a negative correlation suggests that increase in load demand is identified by a decrease in wind power generation (and vice-versa), thereby asking for more production from load following plants in the power system. Also, it explains the need to balance out the wind power fluctuations in different zones with corresponding load fluctuations to maintain a steady supply. To suffice the negative correlation, the zones need to have adequate transmission connection so as to utilize the generated wind power at the location with high demand. In the context of this study, it is fairly understandable that more WPPs will be integrated under the control area as shown in Figure 1.

Because of the location constraint of wind power, WPPs are usually far away from load centers. The knowledge about the correlation between wind power and electricity load will help in ascertaining the capability of wind power to equalize the changes in load fluctuation. Figure 20 shows the overall correlation between each zone pair (both load and wind power) as a function of their approximate distance. The zone pair correlation coefficient corresponds to Table 2, and the location details are provided in Khuntia.³⁸

The correlation between zone pairs could vary because of changes in their installed wind power capacity. Since an approximated centroid approach is employed in this study, the exact relationship will also be influenced by the actual size of the zones. Although the correlation patterns

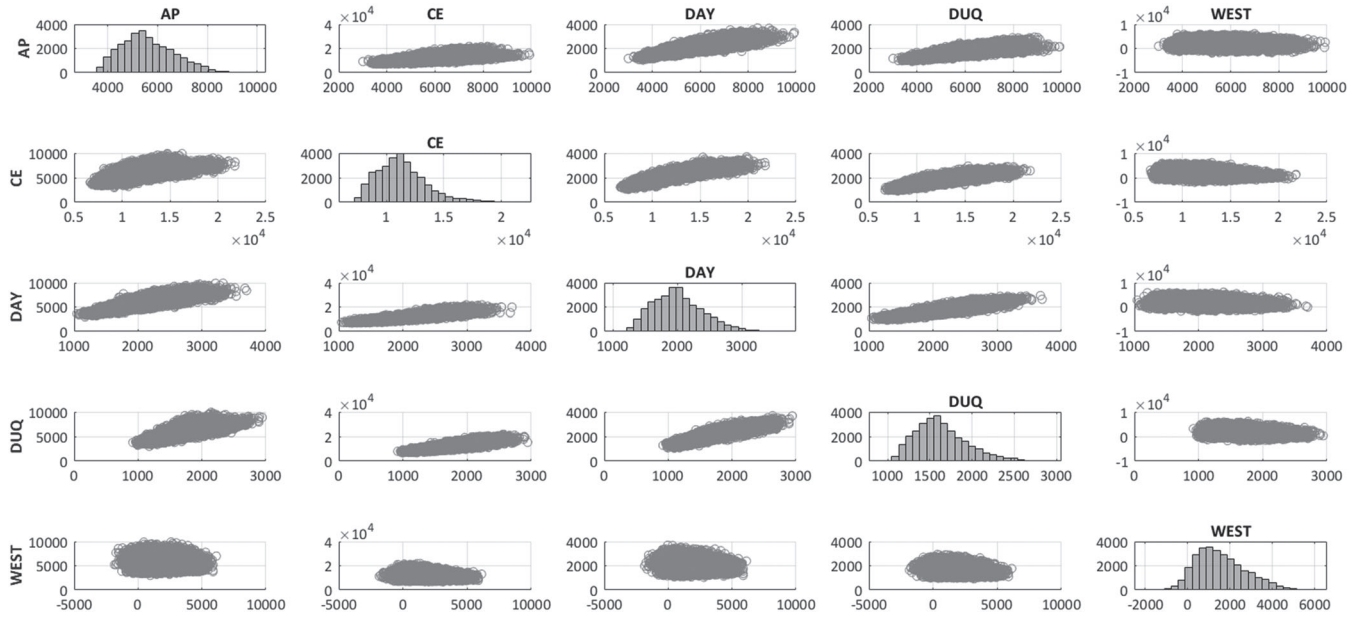


FIGURE 17 Scatter plot with marginal histograms of sampled output of four load zones and one wind power zone

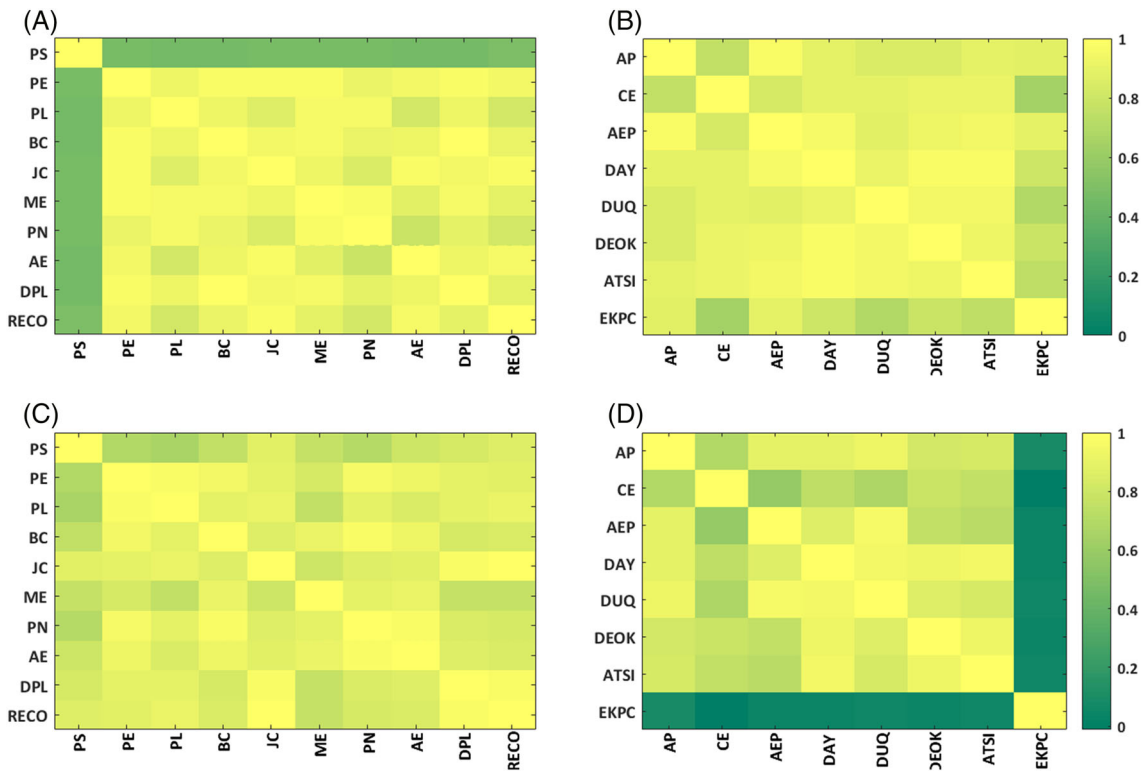


FIGURE 18 Spatial correlation plot of different load zones (A) original dataset *MIDATL* zone, (B) original dataset *WEST* zone, (C) sampled dataset *MIDATL* zone, and (D) sampled dataset *WEST* zone [Colour figure can be viewed at wileyonlinelibrary.com]

will be affected, these patterns still provide useful information. The load-wind power zone pairs in *MIDATL* show significant weak and negative correlation when compared with *WEST*. Thus, in order to meet the increased demand in one load zone within the *MIDATL* market zone, a potential interconnector will benefit the system operator to supply the increased demand by wind power generated in another zone. The notion of weak correlation with the increase in distance does not hold valid for the sampled dataset. While in *MIDATL*, there are instances of zero correlation for the load-load pair, the zero correlation is independent of the distance between the zones. Surprisingly, *WEST* has no instances of zero correlation

TABLE 2 Correlation table for load and wind zones after sampling

	PS	PE	PL	BC	JC	ME	PN	AE	DPL	RECO	AP	CE	AEP	DAY	DUQ	DEOK	ATSI	EKPC	DOM	MIDATL	WEST
PS	1	.69	.67	.75	.88	.77	.72	.8	.83	.87	.85	.53	.96	.78	.91	.67	.61	.02	.93	-.21	-.28
PE	.69	1	.97	.94	.9	.83	.96	.92	.9	.89	.89	.9	.77	.92	.85	.91	.93	.04	.7	-.07	-.07
PL	.67	.97	1	.9	.91	.76	.89	.85	.9	.91	.85	.88	.76	.94	.85	.94	.94	.03	.66	-.03	-.03
BC	.75	.94	.9	1	.87	.91	.97	.94	.84	.85	.94	.74	.83	.93	.89	.87	.92	.08	.81	-.09	-.12
JC	.88	.9	.91	.87	1	.8	.86	.89	.97	.98	.9	.78	.91	.94	.96	.88	.84	.03	.83	-.13	-.16
ME	.77	.83	.76	.91	.8	1	.9	.91	.77	.77	.89	.65	.82	.81	.84	.73	.76	.09	.82	-.1	-.16
PN	.72	.96	.89	.97	.86	.9	1	.97	.84	.83	.91	.8	.8	.91	.86	.85	.9	.07	.77	-.07	-.12
AE	.8	.92	.85	.94	.89	.91	.97	1	.87	.86	.94	.79	.85	.88	.89	.8	.82	.07	.82	-.11	-.19
DPL	.83	.9	.9	.84	.97	.77	.84	.87	1	.97	.86	.8	.86	.9	.92	.86	.81	.03	.79	-.09	-.15
RECO	.87	.89	.91	.85	.98	.77	.83	.86	.97	1	.87	.78	.9	.92	.95	.88	.82	.02	.81	-.12	-.15
AP	.85	.89	.85	.94	.9	.89	.91	.94	.86	.87	1	.69	.9	.9	.92	.82	.84	.09	.88	-.18	-.2
CE	.53	.9	.88	.74	.78	.65	.8	.79	.8	.78	.69	1	.59	.75	.67	.78	.76	-.01	.47	.01	-.01
AEP	.96	.77	.76	.83	.91	.82	.8	.85	.86	.9	.9	.59	1	.87	.96	.76	.73	.05	.97	-.15	-.21
DAY	.78	.92	.94	.93	.94	.81	.91	.88	.9	.92	.9	.75	.87	1	.94	.94	.94	.05	.81	-.06	-.09
DUQ	.91	.85	.85	.89	.96	.84	.86	.89	.92	.95	.92	.67	.96	.94	1	.86	.83	.06	.92	-.12	-.19
DEOK	.67	.91	.94	.87	.88	.73	.85	.8	.86	.88	.82	.78	.76	.94	.86	1	.93	.04	.68	-.09	-.05
ATSI	.61	.93	.94	.92	.84	.76	.9	.82	.81	.82	.84	.76	.73	.94	.83	.93	1	.05	.67	-.05	-.02
EKPC	.2	.04	.03	.08	.03	.09	.07	.07	.03	.02	0.09	-.01	.05	.05	.06	.04	.05	1	.09	-.01	-.03
DOM	.93	.7	.66	.81	.83	.82	.77	.82	.79	.81	0.88	.47	.97	.81	.92	.68	.67	.09	1	-.19	-.25
MIDATL	-.21	-.07	-.03	-.09	-.13	-.1	-.07	-.11	-.09	-.12	-.18	.01	-.15	-.06	-.12	-.09	-.05	-.01	-.19	1	.54
WEST	-.28	-.07	-.03	-.12	-.16	-.16	-.12	-.19	-.15	-.15	-.2	-.01	-.21	-.09	-.19	-.05	-.02	-.03	-.25	.54	1

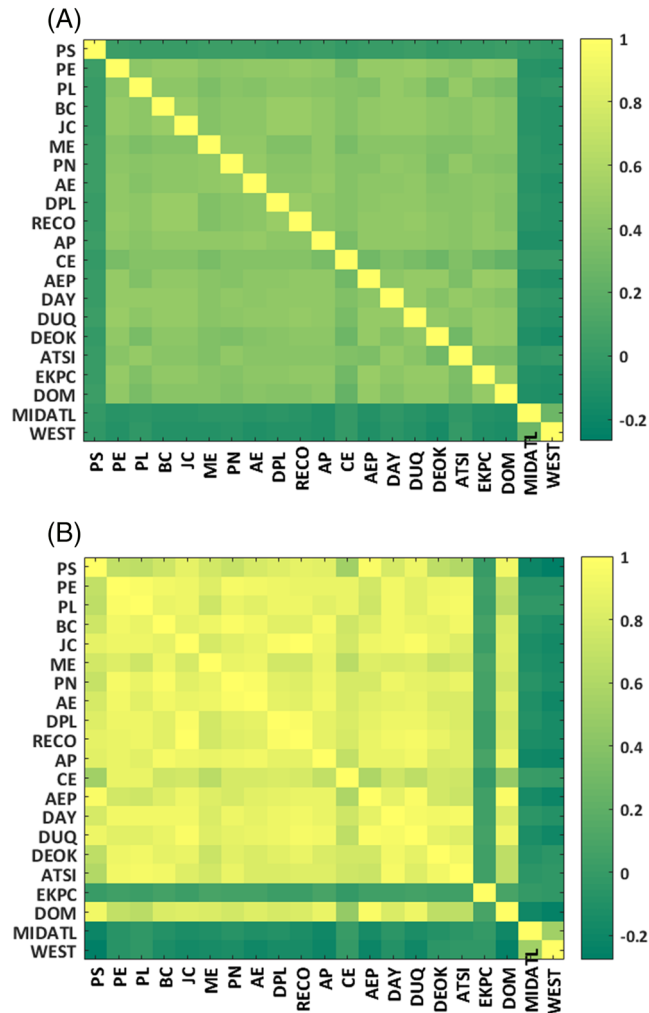


FIGURE 19 Spatial correlation plot of load and wind power zones of (A) original dataset, (B) sampled dataset [Colour figure can be viewed at wileyonlinelibrary.com]

between the load zones even when the separation is approximately 800 kms. The variations will sometime occur in the same directions and help the system and on other times in opposite directions making load following more difficult. Although understanding the correlation of wind power output is important for the incorporation of wind, it will be more critical for determining the regulation reserves necessary as the correlation of changes in wind power and load. In such a case, system operators will seek to supplement the peak demand from conventional generation sources or energy storage if available.

When handling high-dimensional dataset, computational efficiency is a vital concern. In this study, a mix of k -means clustering and C-vine sampling accounted for a computation time of 25.88 seconds. This study aimed at addressing spatio-temporal correlation from TSO's point of view. And, that is the reason for choosing aggregated zonal load and wind power. In terms of future work, to include solar power, this sampling methodology can be extended and reproduced if distribution feeder data are available for load, wind, and solar power.

6 | DISCUSSIONS

In this study, a C-vine-based multivariate framework is developed for spatio-temporal modeling of electricity load and wind power. With increased penetration of wind power into the existing grid, it was deemed vital to model the complex interdependencies introduced by it along with electricity load. This study revealed that chronological simulation of the multivariate dataset using vine copula is possible with conditional distribution calculated by multivariate copula to model the inter-spatial dependence and temporal correlation simultaneously. Such a modeling framework is realized with the developed sampling algorithm, and it can be employed for power system studies (such as security assessment studies, generation and transmission expansion planning, optimal outage scheduling, and stochastic unit commitment) involving a massive integration of stochastic generation. The developed sampling algorithm introduces a systematic way of reducing the original high-dimensional dataset to low-dimensional

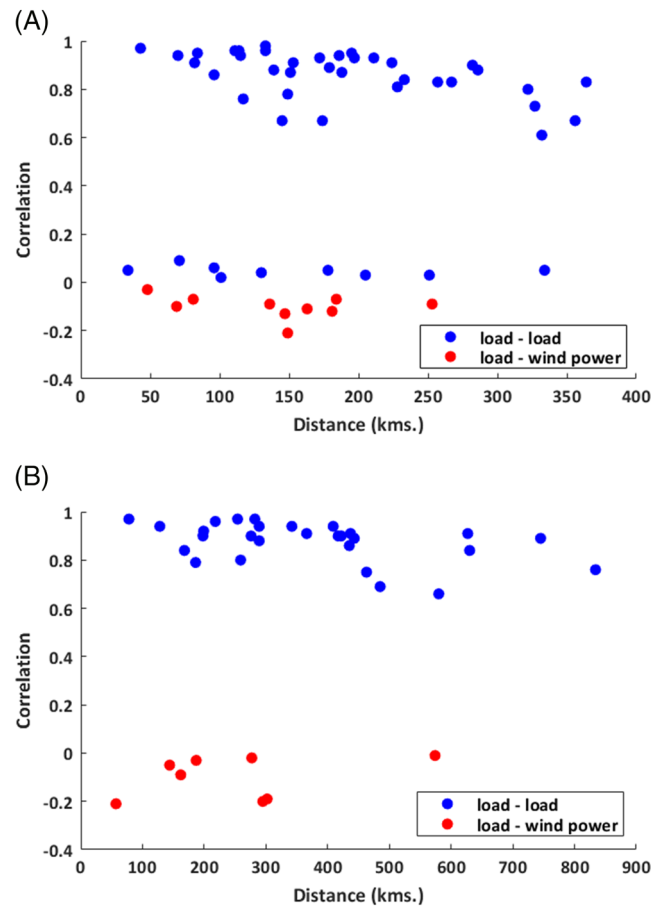


FIGURE 20 Spatial correlation of load and wind power zone pairs under (A) MIDATL and (B) WEST [Colour figure can be viewed at wileyonlinelibrary.com]

space while maintaining essential properties of the original dataset. With TSOs collecting a large amount of data, the future can be seen as data-centric in terms of large-sized high-dimensional data with various features that can surely challenge computational efficiency. To tackle the high dimensionality and variability of datasets, the developed multivariate framework is able to ease the computation burden by employing clustering and feature extraction techniques. In this direction, the sensitivity of clustering algorithms to data normalization is identified as a future work.

ACKNOWLEDGEMENT

The research leading to this result has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 608540 GARPUR project <http://www.garpur-project.eu>. The scientific responsibility rests with the authors.

ORCID

Swasti R. Khuntia  <https://orcid.org/0000-0002-5060-0608>

REFERENCES

- http://unfccc.int/paris_agreement/items/9485.php. Accessed December 26, 2018.
- Shi J, Liu Y, Yu N. Spatio-temporal modeling of electric loads. *Proc. 2017 IEEE North American Power Symposium (NAPS)*.
- Tascikaraoglu A, Sanandaji BM. Short-term residential electric load forecasting: a compressive spatio-temporal approach. *Energ Build.* 2016;111:380-392.
- Shin JH, Yi BJ, Kim YI, Lee HG, Ryu KH. Spatiotemporal load-analysis model for electric power distribution facilities using consumer meter-reading data. *IEEE Trans Power Delivery.* 2011;26(2):736-743.
- Lenzi A, Pinson P, Clemmensen LH, Guillot G. Spatial models for probabilistic prediction of wind power with application to annual-average and high temporal resolution data. *Stochastic Environ Res Risk Assess.* 2017;31(7):1615-1631.
- Osborn D, Henderson MI, Nickell BM, et al. Driving forces behind wind. *IEEE Power Energy Mag.* 2011;9(6):60-74.

7. Louie H. Correlation and statistical characteristics of aggregate wind power in large transcontinental systems. *Wind Energy*. 2014;17(6):793-810.
8. Haghi HV, Lotfifard S. Spatiotemporal modeling of wind generation for optimal energy storage sizing. *IEEE Trans Sustain Energy*. 2015;6(1):113-121.
9. Malvaldi A, Weiss S, Infield D, Browell J, Leahy P, Foley AM. A spatial and temporal correlation analysis of aggregate wind power in an ideally interconnected Europe. *Wind Energy*. 2017;20(8):1315-1329.
10. Wei HU, Yong MIN, Yifan ZHOU, Qiuyu LU. Wind power forecasting errors modelling approach considering temporal and spatial dependence. *J Modern Power Syst Clean Energy*. 2017;5(3):489-498.
11. Miranda MS, Dunn RW. Spatially correlated wind speed modelling for generation adequacy studies in the UK. *Proc. 2007 IEEE Power Engineering Society General Meeting*.
12. Maisonneuve N, Gross G. A production simulation tool for systems with integrated wind energy resources. *IEEE Trans Power Syst*. 2011;26(4):2285-2292.
13. Brown BG, Katz RW, Murphy AH. Time series models to simulate and forecast wind speed and wind power. *J Climate Appl Meteor*. 1984;23(8):1184-1195.
14. Torres JL, Garcia A, De Blas M, De Francisco A. Forecast of hourly average wind speed with ARMA models in Navarre (Spain). *Solar Energy*. 2005;79(1):65-77.
15. Tastu J, Pinson P, Madsen H. Space-time scenarios of wind power generation produced using a Gaussian copula with parametrized precision matrix. *Tech. Univ. Denmark, Tech. Rep.*; 2013.
16. Papavasiliou A, Oren SS, Aravena I. Stochastic modeling of multi-area wind power production. *Proc. 2015 48th Hawaii International Conference on System Sciences (HICSS)*.
17. <https://www.pjm.com/library/reports-notice/rtep-documents/2016-rtep.aspx>. Accessed December 26, 2018.
18. <https://www.pjm.com/committees-and-groups/subcommittees/irs/pris.aspx>. Accessed December 26, 2018.
19. <https://www.pjm.com/renewables/default.html>. Accessed December 26, 2018.
20. <https://irena.masdar.ac.ae/gallery/#gallery>. Accessed December 26, 2018.
21. <https://maps.nrel.gov/wind-prospector/>. Accessed December 26, 2018.
22. Embrechts P, Lindskog F, McNeil A. Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*; 2001.
23. Sklar M. Fonctions de repartition an dimensions et leurs marges. *Publ Inst Statist Univ Paris*. 1959;8:229-231.
24. Pinson P, Girard R. Evaluating the quality of scenarios of short-term wind power generation. *Appl Energy*. 2012;96:12-20.
25. Hagspiel S, Papaemannouil A, Schmid M, Andersson G. Copula-based modeling of stochastic wind power in Europe and implications for the Swiss power grid. *Appl Energy*. 2012;96:33-44.
26. Park H, Baldick R, Morton DP. A stochastic transmission planning model with dependent load and wind forecasts. *IEEE Trans Power Syst*. 2015;30(6):3003-3011.
27. Gill S, Stephen B, Galloway S. Wind turbine condition assessment through power curve copula modeling. *IEEE Trans Sustain Energy*. 2012;3(1):94-101.
28. Morales O, Kurowicka D, Roelen A. Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Eng Syst Safety*. 2008;93(5):699-710.
29. Embrechts P, McNeil A, Straumann D. Correlation and dependence in risk management: properties and pitfalls. *Risk manag: value risk beyond*. 2002;1:176-223.
30. Patton AJ. Copula-based models for financial time series. In: *Handbook of financial time series*. Berlin, Heidelberg: Springer; 2009:767-785.
31. Genest C, Gendron M, Bourdeau-Brien M. The advent of copulas in finance. *Eur J Financ*. 2009;15(7-8):609-618.
32. Louie H. Evaluation of bivariate Archimedean and elliptical copulas to model wind power dependency structures. *Wind Energy*. 2014;17(2):225-240.
33. Sun M, Konstantelos I, Tindemans S, Strbac G. Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems. *Proc. 2016 Power Systems Computation Conference (PSCC)*.
34. Sun M, Konstantelos I, Strbac G. C-Vine copula mixture model for clustering of residential electrical load pattern data. *IEEE Trans Power Syst*. 2017;32(3):2382-2393.
35. Wang Y, Zhang N, Kang C, Miao M, Shi R, Xia Q. An efficient approach to power system uncertainty analysis with high-dimensional dependencies. *IEEE Trans Power Syst*. 2018;33(3):2984-2994.
36. Wang Z, Wang W, Liu C, Wang Z, Hou Y. Probabilistic forecast for multiple wind farms based on regular vine copulas. *IEEE Trans Power Syst*. 2018;33(1):578-589.
37. <https://www.pjm.com/markets-and-operations/ops-analysis/>. Accessed December 26, 2018.
38. Khuntia SR. Probabilistic security assessment of sustainable power grids: Multivariate analysis and dependence modeling for risk-based security assessment. PhD Thesis. TU Delft; 2018.
39. Mohamad IB, Usman D. Standardization and its effects on K-means clustering algorithm. *Res J Appl Sci, Eng Technol*. 2013;6(17):3299-3303.
40. Fan J, Han F, Liu H. Challenges of big data analysis. *Natl Sci Rev*. 2014;1(2):293-314.
41. Berkhin P. A survey of clustering data mining techniques. In: *Grouping multidimensional data*. Berlin, Heidelberg: Springer; 2006:25-71.
42. Law HC. *Clustering, dimensionality reduction, and side information*. PhD Thesis. Michigan State University. 2006.
43. Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. *J Royal Stat Soc Series C (Applied Statistics)*. 1979;28(1):100-108.

44. Bishop CM. Mixture models and EM. *Pattern Recogn Mach Learning*. 2006;1:423-460.
45. <https://people.csail.mit.edu/rameshvs/content/gmm-em.pdf>. Accessed December 26, 2018.
46. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224-227.
47. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J Royal Stat Soc: Series B (Statistical Methodology)*. 2001;63(2):411-423.
48. Cressie N, Wikle CK. *Statistics for spatio-temporal data*. New York: John Wiley & Sons; 2015.
49. Wall ME, Rechtsteiner A, Rocha LM. Singular value decomposition and principal component analysis. In: *A practical approach to microarray data analysis*. Boston, MA: Springer; 2003:91-109.
50. Genest C, Favre AC. Everything you always wanted to know about copula modeling but were afraid to ask. *J Hydrol Eng*. 2007;12(4):347-368.
51. Stephens MA. Tests based on EDF statistics. *Goodness-of-fit Techniq*. 1986;68:97-193.

How to cite this article: Khuntia SR, Rueda JL, van der Meijden MAMM. A multivariate framework to study spatio-temporal dependency of electricity load and wind power. *Wind Energy*. 2019. <https://doi.org/10.1002/we.2407>