Forecasted Scenarios of Regional Wind Farms Based on Regular Vine Copulas

Zhao Wang, Weisheng Wang, Chun Liu, and Bo Wang

Abstract—Owing to the uncertainty and volatility of wind energy, forecasted wind power scenarios with proper spatio-temporal correlations are needed in various decision-making problems involving power systems. In this study, forecasted scenarios are generated from an estimated multi-variate distribution of multiple regional wind farms. According to the theory of copulas, marginal distributions and the dependence structure of multi-variate distribution are modeled through the proposed distance-weighted kernel density estimation method and the regular vine (R-vine) copula, respectively. Owing to the flexibility of decomposing correlations of high dimensions into different types of pair-copulas, the R-vine copula provides more accurate results in describing the complicated dependence of wind power. In the case of 26 wind farms located in East China, high-quality forecasted scenarios as well as the corresponding probabilistic forecasting and point forecasting results are obtained using the proposed method, and the results are evaluated using a comprehensive verification framework.

Index Terms—Forecasted scenarios, wind power, distance-weighted kernel density estimation (KDE), regular vine (R-vine) copula, spatio-temporal correlation.

I. INTRODUCTION

WIND power, as a promising type of renewable energy, has been integrated into power grids with increasing capacities in recent years. However, compared with conventional power generation, the inherent variability and the intermittence of wind power can introduce uncertainties to power systems. Therefore, it is a challenge for the system operators to schedule a reliable and economical plan considering the uncertainties caused by wind power.

Wind power forecasting (WPF) is a powerful tool that can provide decision makers, including system operators and participants in the electricity market, with the forecasting of future power outputs. Deterministic forecasting, i.e., point forecasting, is the main type of WPF and has been studied worldwide for decades, and a detailed review is presented in [1]. Point forecasting only gives the expected value of future power generation, while probabilistic forecasting has attracted increasing attention recently. It can provide situation-specific uncertainty information not included by the point forecasting results. Various types of probabilistic forecasting methods have been proposed in the field of WPF [2]–[9]. For example, a convenient adapted-resampling method based on empirical distributions of forecasting errors is proposed [2]. Quantile regression is another research focus [3], [4]. Prediction interval estimation methods employing machine-learning algorithms are introduced [5], [6]. For providing a continuous probability density function (PDF), kernel density estimation (KDE) methods are presented [7], [8]. A method based on meteorological ensembles of numerical weather prediction (NWP) systems is discussed in [9].

The aforementioned probabilistic forecasting methods give the forecasting distributions of each lead time and wind farm separately. The interdependence among these forecasting distributions is unclear. Therefore, it is difficult to generate the scenarios of either accurate time dependence (across the different lead time of a single wind farm) or accurate spatial dependence (among different wind farms). For the various decision-making problems based on stochastic programming, the wind power scenario set is indispensable for analyzing the effects of uncertainty. Stochastic programming is applied in many fields such as trading in the electricity market [10], calculating reserve requirements [11], and solving unit commitment and economic dispatch problems [12], [13].

There are two types of methods for obtaining the requested wind power scenarios. One comes from the ensemble NWP provided by NWP service providers. The ensemble NWP provides the scenarios of the concerned meteorological variables by perturbing the initial conditions and model parameters of NWP models. A review of ensemble forecasting is presented in [14]. With the forecasted wind speed ensembles, the wind power scenarios can be achieved through a wind-to-power conversion. The other method involves extracting the wind power scenarios from the multi-variate joint distribution whose marginal distributions correspond to the WPF distributions of different lead time and wind farms. A good way to model the multi-variate distribution is through the copula theory, which is capable of describing a multi-variate distribution with the marginal distributions and their dependence structure constructed independently. In
[15], scenarios based on a Gaussian copula model that summarizes the multi-variate joint distribution of a single wind farm forecasting are generated. Under the same Gaussian copula assumption, precision matrices are used to describe the spatiotemporal dependence for the situation of multiple wind farms [16]. To generate the multi-dimensional scenarios, the Gaussian copula model is combined with the forecasting error distributions of the point forecasting based on a support vector machine [17]. In [18], a new dependence model based on vine copulas is introduced to describe the spatiotemporal dependence of the wind data, and the model has higher accuracy than the Gaussian copula model for modeling the complex wind data. However, the model is utilized to generate synthetic wind data for a long-term application that only considers the historical data without the information of WPF.

On the basis of the aforementioned works, an improved scenario forecasting method involving the modeling of the forecasting multi-variate distribution is proposed in this paper. The main contributions of this paper are as follows.

1) The regular vine (R-vine) copula is introduced to describe the complicated dependence structure of high-dimensional wind data. It is more accurate and flexible than the Gaussian copula model. The runtime of the more complicated copula model is limited within a reasonable range through the selection of the vine structures and the application of the independence test before the estimation.

2) A distance-weighted KDE method is proposed to model the conditional forecasting distributions, which are the marginal distributions of the multi-variate distribution. High-quality probabilistic forecasting results can be achieved through this method.

3) An improved point forecasting result, as a byproduct, can be extracted from the conditional forecasting distributions.

4) Sufficient wind power scenarios with the proper temporal and spatial correlations can be easily obtained through the proposed model.

II. DATA

The data for 26 wind farms located in East China are used in this study, as shown in Fig. 1. Most of these wind farms are built near the coastline. They are easily affected by a weather front wiping through the whole area, resulting in significant spatiotemporal correlations among the wind power outputs. As shown in Fig. 1, four wind farm groups are obtained via the k-means clustering algorithm [19] according to the geographical coordinates (pairs of longitude and latitude) of the wind farms. In practice, the wind farm groups can be divided according to the network structure of the power grid by the system operators. This study focuses on the forecasting results of the total output for each group. The total capacity of the 26 wind farms is approximately 1.93 GW. The NWP data and wind power measurements are corresponding to the period between May 5, 2014 and March 24, 2016. The time range of the training set is May 5, 2014 to May 5, 2015, and the remaining data are used as the test set. The temporal resolution of the data is 1 hour. In this study, the proposed model is designed for day-ahead application, and the forecasting horizon is 24 hours.

III. METHODOLOGY

In the copula theory, Sklar’s theorem [20] explains the relationship among the multi-variate distribution function, the copula function, and the marginal distribution function. Let $F$ be a $d$-dimensional joint cumulative distribution function (CDF) whose marginal CDFs are $F_1, F_2, \ldots, F_d$ and $F$ can be determined as (1) by a unique copula function $c$, if the marginal functions are all continuous.

$$F(x) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$$  \hspace{1cm} (1)

A PDF version of (1) is as follows:

$$f(x) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) f_1(x_1)f_2(x_2)\ldots f_d(x_d)$$  \hspace{1cm} (2)

In the case of the WPF, marginal distributions of $F$ correspond to the conditional forecasting distributions of four wind farm groups with 24 time points per day. Therefore, the dimensionality $d$ of $F$ is 96.

The overall framework of the proposed methodology is summarized in Fig. 2. The whole dataset is divided into a training dataset and a test dataset. In each dataset, NWP data and power outputs are included. The test dataset is used to evaluate the out-of-sample performance of the proposed method. A distance-weighted KDE method is employed to estimate the marginal distributions in Section III-A, and a flexible and accurate dependence model—the R-vine copula—is introduced to model the dependence structure in Sections III-B to III-D. The multi-variate distribution is the combination of the marginal distributions and the dependence structure. Three types of forecasting results can be easily obtained using the estimated model. The probabilistic forecasting of each group corresponds to the estimated marginal distribution. Forecasting scenarios can be sampled from the estimated multi-variate distribution. The improved point forecasting results are represented by the median of the marginal distributions. Finally, a detailed evaluation is conducted on the test dataset of the 26 wind farms in Fig. 1.

A. Conditional Probabilistic Forecasting

Let $P_{i,t}$ be the total output of a wind farm group at time point $t+k$, where $k$ is the lead time of forecasting. If there
The conditional forecasting density of \( P_{i,k} \) is denoted as \( \hat{f}_P(P_{i,k} \mid v_{i,k}) \), and the symbol “\(^{\wedge}\)” on the top indicates that the object is estimated. For a region with a large amount of wind farms, the elements in the explanatory variable vector can be optimized via either feature selection or feature extraction, as described in [21].

Since \( \hat{f}_P(P_{i,k} \mid v_{i,k}) \) is constructed using the historical sample set \( \{P\} \) with similar conditions as \( v_{i,k} \), a similarity measure should be defined to select the proper set \( \{P\} \). In this study, the Euclidean distance \( d \) defined in (4) is used to quantify the similarity of the target vector \( v_{i,k} \) and a historical record \( v_i \). \( \omega_i \) (\( i = 1, 2, \ldots, M \)) is used to emphasize the different importance of each dimension in \( v_i \); the weight can be set as 1 if no emphasis is needed.

\[
d_{i,k} = \sqrt{(v_{i,k} - v_i)^T \text{diag}(\omega_1, \omega_2, \ldots, \omega_M)(v_{i,k} - v_i)}
\]  

(4)

Owing to the nonlinearity of wind power generation, the assumption that the conditional forecasting distributions of the wind power \( P \) follow a known parametric distribution proves to be weak [2]. Therefore, we follow the nonparametric framework, and a distance-weighted KDE method is proposed. This method is summarized as follows.

Since the normalized wind power \( P \) is bounded in the range of \([0, 1]\) and its distribution is skewed. A logarithm transformation (5) is applied to facilitate the estimation of the Gaussian kernel (6), which changes the domain of \([0, 1]\) to \((-\infty, +\infty)\).

\[
g(P) = \ln \frac{P}{1-P}
\]  

(5)

\[
G(z) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z^2}{2} \right)
\]  

(6)

Then, the distance-weighted KDE of \( \hat{f}_{i,k}(P_{i,k} \mid v_{i,k}) \) is obtained as follows:

\[
\hat{f}_{i,k}(P_{i,k} \mid v_{i,k}) = \frac{1}{\eta_G} \sum_n \omega_n(d_n) G \left( \frac{g(P_{i,k}) - g(P_n)}{\eta_G} \right)
\]  

(7)

where \( \eta_G \) represents the smoothing bandwidths of the Gaussian kernel; subscript \( n \) represents the \( n^{th} \) observation in the training dataset.

\[
\omega_n(d_n) = \frac{\exp \left( -d_n^2/\eta_d \right)}{\sum \exp \left( -d_k^2/\eta_d \right)}
\]  

(8)

As indicated by (8), the weight \( \omega \) decays as \( d \) increases, and the parameter \( \eta_d \) controls the slope of the decay.

Forecasting quantiles and intervals can be easily extracted from the continuous distribution function \( \hat{f}_{i,k} \) or \( \hat{F}_{i,k} \).

The forecasting quantile with the nominal proportion \( \alpha \) in \([0, 1]\) is expressed as:

\[
\hat{q}_{i,k}^{(\alpha)} = \hat{F}_{i,k}^{-1}(\alpha)
\]  

(9)

A forecasting interval with the nominal coverage rate \((1-\beta)\) is expressed as:

\[
\hat{I}_{i,k}^{(1-\beta)} = \left[ \hat{q}_{i,k}^{(\alpha)} - \hat{q}_{i,k}^{(1-\alpha)} \right]
\]  

(10)

The main difference between the proposed method and the traditional KDE method is that the distance-weighted KDE can reward more relevant samples (samples with a smaller distance \( d \)) with higher weights, which can improve the forecasting accuracy and robustness against extreme outliers. Additionally, the multiple conditions described by the explanatory variables can be easily reduced to a single measure of similarity \( d \).

B. Copulas of High-dimensional Space

It is computationally efficient for Gaussian copula to describe the high-dimensional dependence structure, because only the correlation coefficient matrix needs to be estimated. However, the Gaussian copula fails to describe the asymmetric tail dependence, which can be estimated precisely via some Archimedean copulas such as the Gumbel copula and Clayton copula. However, different Archimedean copulas cannot be used to simultaneously model the dependence of more than two dimensions. Different shapes of bivariate copula functions are clearly shown in Fig. 3.

In contrast to the two aforementioned classes of copulas, the R-vine copula provides a flexible way to decompose the high-dimensional dependence structure into many pair-copulas. And for each pair-copula, a best-fit copula function can be selected from different copula functions, which makes the R-vine copula model more accurate than the Gaussian copula in the case of a complicated dependence structure. An R-vine copula model is employed to estimate the dependence structures for conditional probabilistic forecasting in [21-22],
demonstrating its advantages for modeling multi-variate distributions.

where \( V(\cdot) = \{ f(\cdot), k(\cdot), D(\cdot) \} \) represents the universal set for an edge. Then, we can associate each edge \( e \) with a bivariate copula \( c_{f_0(x),y_0(x)}(D(e)) \). Let \( X \) be a \( d \)-dimensional random vector. \( x_{X(e)} \) represents the subset of \( X \) indicated by the elements in \( D(e) \), as for \( x_{X(e)} \) and \( x_{X(D(e))} \). With the structure determined uniquely by a specific R-vine and marginal distributions \( f_e, k = 1, 2, \ldots, d \), the joint PDF of \( X \) can be expressed as:

\[
 f(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} f_i(x_i) \times \prod_{i=1}^{d-1} c_{f_i(x_i), A_i(D(e))} \left( F(x_{i+1}(x_i)|x_{D(e)}) \right) \left( F(x_{i+1}(x_i)|x_{D(e)}) \right)
\]

where the conditional CDF \( F(x_{i+1}(x_i)|x_{D(e)}) \) is calculated as follows:

\[
 F(x_{i+1}(x_i)|x_{D(e)}) = F(x_{i+1}(x_i)|x_{D(e)}|x_{A_i}) = \frac{c_{f_i(x_i), A_i(D(e))} \left( F(x_{i+1}(x_i)|x_{D(e)} \right)}{\partial F(x_{i+1}(x_i)|x_{D(e)})}
\]

Similar to (14), \( F(x_{i+1}(x_i)|x_{D(e)}) \) can be obtained using the corresponding copula function. Equation (14) indicates that conditional CDF can be calculated recursively from the conditional CDF with a lower-dimensional conditioning set.

D. R-vine Copula Model of Multiple Wind Farms

The sequential method proposed in [24] and [25] is employed to establish the R-vine copula model of wind farms. Taking the \( d \)-dimensional random vector \( X \) for example, the modeling process is as follows.

**Step 1:** calculate the empirical Kendall’s \( \tau \) for all the variable pairs. For the \( d \)-dimensional case, the total number of variable pairs is \( C_d^2 = d(d-1)/2 \). The dimensions of the modeling data \( X \) are the actual wind power outputs of the four groups of wind farms with the horizon ranging from 1 to 24 hours ahead.

\[
 X = \left[ \begin{array}{cccc} W_{t,1} & W_{t,2} & \ldots & W_{t,24} \\
 W_{t,1} & W_{t,2} & \ldots & W_{t,24} \\
 W_{t,3} & W_{t,2} & \ldots & W_{t,4} \\
 W_{t,1} & W_{t,2} & \ldots & W_{t,24} \\
 \end{array} \right]
\]

**Step 2:** the tree structure with the highest value of the sum of all the Kendall’s \( \tau \) calculated in **Step 1** is selected, and the searching method is implemented according to the algorithm of the maximum spanning tree [26].

**Step 3:** each edge of the tree corresponds to a pair-copula, and a best-fit copula function is selected for each edge according to the Akaike information criterion (AIC) [27]. AIC shown in (16) can measure the quality of a statistical model.

\[
 \text{AIC} = 2k - 2\ln(\hat{L})
\]

In this study, there are abundant bivariate copula functions of the asymmetric copulas to choose from, including the Gaussian copula, t copula, Frank copula, Clayton copula, Gumbel copula, Joe copula, and rotated versions (rotated by 90°, 180°, and 270°) of the asymmetric copulas. The parame-
ters are estimated when selecting different copula functions. To reduce the computational burden, an independence test described in [28] is performed before selecting copula functions. For the pair of variables without significant dependence, no comparison among copula functions is needed, and an independence copula is used instead. The conditional variable pairs are calculated using the estimated conditional CDF, as (14), for the next tree to use.

Step 4: the three aforementioned steps are iterated from tree to tree. And the variable pairs of the trees, except for the first tree, are the conditional variable pairs transformed from the last tree.

E. Generating Forecasting Scenarios of Multiple Wind Farms

Scenarios can be easily obtained from the estimated R-vine copula model \(C\) and the conditional forecasting marginal distributions \(F_1, F_2, \ldots, F_d\) estimated in Section III-A via inverse transformation sampling. The detailed procedures are as follows.

1) Generate the \(d\)-dimensional and random vector \(W = [W_1, W_2, \ldots, W_d]\) following the uniform distribution \(U(0,1)^d\).

2) To the inverse functions of the conditional CDF \(F_{d|d-1,\ldots,1}(x)\), \(i = 1, 2, \ldots, d\), defined by copula model \(C\), we can obtain the correlated rank random vector \(U = [U_1, U_2, \ldots, U_d]\) as (17), whose marginal function is the uniform distribution \(U(0,1)\).

\[
\begin{align*}
U_1 & = W_1 \\
U_2 & = F_{d|d-1,\ldots,1}^{-1}(W_2|U_1) \\
& \vdots \\
U_d & = F_{d|d-1,\ldots,1}^{-1}(W_d|U_{d-1},U_{d-2},\ldots,U_1)
\end{align*}
\] (17)

3) Finally, scenarios of wind power outputs are given by:

\[
P = \left[ F_1^{-1}(U_1) \ F_2^{-1}(U_2) \ \ldots \ F_d^{-1}(U_d) \right] \] (18)

IV. FRAMEWORK OF EVALUATION

Using the methodology described in Section III, the probabilistic forecasting results and forecasting scenarios can be achieved. To evaluate the effectiveness of these results, a complete framework is provided as follows.

A. Evaluation of Probabilistic Forecasting Results

In this section, the quality of the forecasting marginal distribution is evaluated without considering the dependence structure. Three measures are introduced to evaluate the performance in the nonparametric framework, as described in [29].

1) Reliability

The reliability \(R_{\alpha}^{(i)}\) (19) is given by the deviation between the nominal proportion \(\alpha\) and the observed frequency of the data below the quantile forecasting \(\hat{q}^{(i)}\) in the test set.

\[
R_{\alpha}^{(i)} = \frac{\hat{n}^{(i)} - \alpha}{N}
\] (19)

where \(\hat{n}^{(i)}\) is the number of the observations located below \(\hat{q}^{(i)}\), and \(N\) is the length of the test set.

2) Sharpness

The sharpness \(S_{\alpha}^{(i)}\) is evaluated through the mean width of forecasting intervals.

\[
S_{\alpha}^{(i)} = \frac{1}{N} \left( \hat{q}^{(1-\beta/2)} - \hat{q}^{(\beta/2)} \right)
\] (20)

3) Skill score

The skill score introduced in [29] is an overall measure designed for the quantile forecasting set, which considers both the reliability and sharpness. The skill score is defined as:

\[
S_{\alpha} = \sum_{k=1}^{m} \left( \hat{q}^{(k)} - \alpha \right) \left( x - \hat{q}^{(k)} \right)
\] (21)

where \(x\) is the corresponding observation; and \(\hat{q}^{(k)}\) is an indicator variable that equals to 1 when \(x \leq \hat{q}^{(k)}\), and equals to 0 otherwise. The scoring rule is proper, and a method with higher quality is rewarded with a higher skill score.

Another scoring rule called the continuous rank probability score (CRPS) is a proper, negatively oriented score. CRPS is designed for continuous forecasting density results and is calculated using (22) to measure the dissimilarity between the forecasting distribution \(\hat{F}_x\) and the observation \(X_x\). The \(\xi\) in (22) is the same as that in (21), which is an indicator variable. The mean value of CRPS over a specific data set can be used to verify the effectiveness of the forecasting method.

\[
CRPS = \int_{-\infty}^{\infty} \left( \hat{F}_x - \xi \right)^2 dx
\] (22)

The computation of the integral in (22) is complicated. Therefore, an alternative representation (23) introduced in [30] is applied, which can be easily solved using the Monte Carlo method.

\[
CRPS = E_x \left| X - x \right| - \frac{1}{2} E_{\epsilon_x} \left| X - X' \right|
\] (23)

where \(X\) and \(X'\) are the independent random variables of the forecasting distribution \(\hat{F}_x\) and \(E_{\epsilon_x}\) is the expectation.

B. Evaluation of Forecasting Scenarios

The information involved in the forecasting scenarios includes not only the marginal forecasting distribution of each look-ahead time but also the dependence structure among these marginal distributions. Therefore, the energy score \(ES\), which is a generalization of CRPS designed for the multi-variate situation, is applied to evaluate the quality of the forecasting scenarios [31].

\[
ES = E_x \left| X - X' \right| - \frac{1}{2} E_{\epsilon_x} \left| X - X' \right|
\] (24)

where \(X\) and \(X'\) are the independent random vectors of the multi-variate forecasting distribution \(\hat{F}_x(x)\); and \(\left| \cdot \right|\) represents the Euclidean norm. When the energy score is designed for a single wind farm group, the dimensions of \(X\) correspond to the forecasted wind power outputs of each look-ahead time up to 24 hours. Considering the spatial correlation, the forecasted wind power outputs of other groups should be included in the random vector \(X\).

Taking the forecasting scenarios with \(m\) time series for ex-
ample, the energy score can be calculated as:

$$ES = \frac{1}{m} \sum_{j=1}^{m} \left\| X_j - x \right\| - \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\| X_i - X_j \right\|$$ (25)

In Section V, a case study is presented to demonstrate the performance in the evaluation framework introduced in this section.

V. CASE STUDY

In the case study, the four groups of wind farms in Fig. 1 are used to verify the effectiveness of the proposed method. By decomposing the forecasting multi-variate distribution into two parts, the marginal distributions and the dependence structure are evaluated in Sections V-A and V-B, respectively. The quality of the generated forecasting scenarios is evaluated in Section V-C.

A. Performance of Conditional Forecasting Distributions

In this section, the proposed weighted KDE (WKDE) method is compared with a frequently used probabilistic forecasting method called the empirical quantile forecasting (denoted as Emp). The Emp method selects the samples according to similar weather conditions in the historical dataset, similar to WKDE, and then calculates the quantiles of the selected samples without estimating the underlying distributions. Similar to the method introduced in [2], a re-sampling process based on the bootstrap method is adopted to make the forecasting quantiles more robust. The available data for the two methods are exactly the same.

Figure 4 shows the forecasting intervals of 3 days for Group 2. The yellow dotted line is the observed power output of Group 2. The white line corresponds to the point forecasting results. Prediction intervals (PIs) with a nominal coverage rate ranging from 0.1 to 0.9 are depicted with different depths of blue. The 0.5-quantile (median of the distribution) is noted as Emp. The Emp method selects the samples according to similar weather conditions in the historical dataset, similar to WKDE, and then calculates the quantiles of the selected samples without estimating the underlying distributions. Similar to the method introduced in [2], a re-sampling process based on the bootstrap method is adopted to make the forecasting quantiles more robust. The available data for the two methods are exactly the same.

Figure 4 shows the forecasting intervals of 3 days for Group 2. The yellow dotted line is the observed power output of Group 2. The white line corresponds to the point forecasting results. Prediction intervals (PIs) with a nominal coverage rate ranging from 0.1 to 0.9 are depicted with different depths of blue. The 0.5-quantile (median of the distribution) is noted as Emp. The Emp method selects the samples according to similar weather conditions in the historical dataset, similar to WKDE, and then calculates the quantiles of the selected samples without estimating the underlying distributions. Similar to the method introduced in [2], a re-sampling process based on the bootstrap method is adopted to make the forecasting quantiles more robust. The available data for the two methods are exactly the same.

Most of the PIs obtained via WKDE are sharper than those of Emp, as shown in Fig. 5(b), because of the distance-weighted mechanism of WKDE, which can make the distribution concentrated on the samples with higher similarity to the target point. Figure 5(c) depicts the sharpness for the 0.9-PIs based on the forecasting horizon. The skill-score results are presented in Fig.5(d) and (e). Because the Emp method only provides forecasting quantiles, which are not in the form of the continuous distribution, the scoring rule of (21) is used instead of CRPS. As shown in Fig.5(d), the skill scores of WKDE decrease slightly as the forecasting horizon increases, and the results of WKDE are significantly better than those of Emp for horizons up to 14 hours. The difference for longer horizons is not obvious. To view the distributions of the skill scores over the test set clearly, two box-plots are presented in Fig.5(e), with the mean and median values of the skill scores provided. More scores close to 0
are achieved by WKDE. As the skill score is an overall metric considering both the reliability and sharpness, the quality of the probabilistic forecasting can be simply evaluated by a single scale value. The results for the other three groups are presented in Table I, indicating that WKDE had better performance than Emp for these groups.

### Table I

| Group | Mean energy score (p.u.) | Median energy score (p.u.) |
|-------|-------------------------|---------------------------|
|       | WKDE                    | Emp                       |
| 1     | 0.5                     | 0.385                     |
| 2     | 0.7                     | 0.370                     |
| 3     | 0.9                     | 0.369                     |
| 4     | 1.1                     | 0.413                     |

As a byproduct of the probabilistic forecasting, the median and the mode of the forecasting distribution of WKDE can be extracted as an improved point forecasting result, denoted as “WKDE (Median)” and “WKDE (Mode)”, respectively. While the median of the empirical quantile forecasting is denoted as “Emp (Median)”. The point forecasting of each wind farms is simply added up as the original forecasting result denoted as “Original”. The performance of the aforementioned point forecasting for Group 2 is shown in Table II. The four performance metrics evaluated are “bias” (mean of errors), “MAE” (mean absolute errors), “RMSE” (root mean square errors), “Corr” (Pearson’s linear correlation coefficients). It is clear that the point forecasting of WKDE are better in performance than the other results.

### Table II

| Point forecasting | bias       | MAE      | RMSE     | Corr    |
|-------------------|------------|----------|----------|---------|
| WKDE (Median)     | 0.0151     | 0.0750   | 0.1089   | 0.8396  |
| WKDE (Mode)       | 0.0004     | 0.0764   | 0.1096   | 0.8379  |
| Emp (Median)      | 0.0179     | 0.0829   | 0.1195   | 0.8028  |
| Original          | -0.0396    | 0.0920   | 0.1258   | 0.8118  |

### B. Performance of Dependence Structure Estimation

Two different copula models—the Gaussian copula model and the proposed R-vine copula model—are estimated to describe the dependence structure of the multi-variate distribution in this section. The log-likelihood and AIC are calculated over the given sample set for these two models, as shown in Table III. According to the results, the R-vine copula model is preferred from the viewpoint of modeling accuracy. Approximately, 66.64% of the pair-copulas are replaced by the independence copulas. This is because the high-dependence pairs are modeled firstly via the sequential method introduced in Section III-D. Thus, the conditional pair variables in the latter trees tend to be independent. In previous studies, truncation [18] and simplification [24] methods are employed to replace the pair-copulas of the trees higher than a specific level with independent copulas and Gaussian copulas, respectively. The high proportion of independence copulas can significantly reduce the computational burden of modeling the R-vine copula. It takes more than four hours to estimate the R-vine copula model without independence tests on a Windows PC with Intel(R) Core(TM) i5-4210U (8 GB of RAM). Whereas it takes only 103 min to estimate the R-vine copula model with the independence tests. The most computationally efficient method is the Gaussian copula, which only takes 5 minutes to finish the calculation under the same condition. For the application of day-ahead planning, the runtime of 103 minutes is reasonable, and with parallel computing, the runtime for estimating R-vine copula can be further reduced. Additionally, the dependence model is an offline model. Once the model is established, enough scenarios can be extracted from it without excessive computation. It is not necessary to update the dependence model daily. In practice, it can be updated every few weeks or months.

### Table III

| Copulas     | Log-likelihood | AIC  |
|-------------|----------------|------|
| Gaussian copula | -1432424      |      |
| R-vine copula      | 718362        | 1432424 |

### C. Performance of Forecasting Scenario

Statistical scenarios generated by different models are compared with each other in this section. In Fig. 6, “Independence” represents the reference model which generates samples independently for each marginal distribution and then connects the samples of the successive time points randomly. Thus, no dependence structure is modeled for the “Independence” case.

![Energy scores of scenarios with 50 members.](image)

Figure 6 shows the energy scores of these models. The number of members for each scenario set is 50. In Fig. 6, Group 1 to Group 4 correspond to the energy score of a single wind farm group, and “Total” represents all the wind farm groups included in the random variable vector \( X \) in (25). Because a lower energy score is preferred, the forecasted scenarios generated by the R-vine copula perform better than those generated by its competitors. The weakness of the reference model “Independence” is
obvious in Fig. 7. Although the marginal distributions of the two methods are the same, the scenarios of the two models are significantly different. Because of the lack of dependence modeling, many abrupt waves and irregular fluctuations are observed in the scenarios of “Independence” which is not in accordance with the actual characteristics of wind power outputs.

To examine the time and spatial dependence of the generated scenarios, autocorrelation functions (ACFs) and cross-correlation functions (CCFs) are calculated, as shown in Fig. 8. Owing to the limited space, only a part of the results for Group 2 are presented. Both the R-vine copula and Gaussian copula perform well with regard to the autocorrelation. In Fig. 8(c), the ACF curve of “Independence” deviates significantly from the actual ACF curve. However, the trend of its ACF curve is still similar to that of the actual one. This is because the marginal distributions of “Independence” are the same as those of the other two models, and the randomly selected samples are concentrated in the range of the corresponding marginal distributions. Therefore, the time dependence of “Independence” is partly preserved without the dependence structure modeled. The difference of the CCF results is significant among these three models, and the R-vine model achieves the best performance in this case. The dependence of the multiple wind farms is more complicated than that of a single one. Thus, a more accurate model is needed to describe the dependence structure.

VI. CONCLUSION

The forecasted wind power scenarios discussed in this paper provide valuable uncertainty information, which can help system operators and electricity market participants solve decision-making problems involving multiple wind farms. The marginal uncertainty information is described by a distance-weighted KDE method, which makes the forecasting distribution more robust and automatically rewards the more relevant samples with higher weights. Dependence structures containing information regarding the spatial and temporal correlations among multiple wind power outputs are described by an advanced R-vine copula model, which is effective for decomposing high-dimensional dependence. Abundant bivariate copula functions are available to make the model more accurate. The effectiveness of the proposed method is verified using actual data for 26 wind farms in East China.

The forecasted scenarios are considered in the framework of stochastic programming and the number of scenarios is constrained by the computing power and the size of the corresponding optimizing problem. The case with few scenarios is not capable of summarizing all of the uncertainty information involved in the stochastic processes of wind power outputs. As computing becomes faster and less expensive, the task of achieving enough scenarios will become easier in the future. The proposed method might not be effective for real-time applications with an updating time less than 1 hour. However, once the complicated dependence structure is modeled, little time is needed to generate as many scenarios as
needed, and the computational burden for day-ahead application is manageable.

There is considerable work to be performed on this topic such as the application of the forecasted scenarios to stochastic programming, the characterization of different conditional dependence models for cases where a long-term dataset is available, and the development of technologies for simplifying the model without a significant loss of accuracy.

REFERENCES

[1] G. Giebel, R. Brownsword, G. Kariniotakis et al., “The state-of-the-art in short-term prediction of wind power: a literature overview,” 2nd Edition. Technical Report, [Online]. Available: http://orbit.dtu.dk/fedora/objects/orbit:83397

[2] P. Pinson and G. Kariniotakis, “Conditional prediction intervals of wind power generation,” IEEE Transactions on Power Systems, vol. 25, no. 4, pp. 1845-1856, Nov. 2010.

[3] J. B. Bremnes, “A comparison of a few statistical models for making quantile wind power forecasts,” Wind Energy, vol. 9, no. 1-2, pp. 3-11, Apr. 2006.

[4] Anastasiades and P. McSharry, “Quantile forecasting of wind power using variability indices,” Energies, vol. 6, no. 2, pp. 662-695, Feb. 2013.

[5] C. Wang, Z. Xu, P. Pinson et al., “Probabilistic forecasting of wind power generation using extreme learning machine,” IEEE Transactions on Power Systems, vol. 29, no. 3, pp. 1033-1044, May 2014.

[6] A. Kavousi-Fard, A. Khosravi, and S. Nahavandi, “A new fuzzy-based combined prediction interval for wind power forecasting,” IEEE Transactions on Power Systems, vol. 31, no. 1, pp. 18-26, Jan. 2016.

[7] R. J. Bessa, V. Miranda, A. Botterud et al., “Time adaptive conditional kernel density estimation for wind power forecasting,” IEEE Transactions on Sustainable Energy, vol. 3, no. 4, pp. 660-669, Oct. 2012.

[8] Y. Zhang, J. Wang, and X. Luo, “Probabilistic wind power forecasting based on logarithmic transformation and boundary kernel,” Energy Conversion and Management, vol. 96, pp. 440-451, May 2015.

[9] P. Pinson and H. Madsen, “Ensemble-based probabilistic forecasting at Horns Rev,” Wind Energy, vol. 12, no. 2, pp.137-155, Mar. 2009.

[10] J. M. Morales, A. J. Conejo, and J. Perez-Ruiz, “Short-term trading for a wind power producer,” IEEE Transactions on Power Systems, vol. 25, no. 1, pp. 554-564, Mar. 2010.

[11] A. Papavasiliou, S. S. Oren, and R. P. O’Neill, “Reserve requirements for wind power integration: a scenario-based stochastic programming framework,” IEEE Transactions on Power Systems, vol. 26, no. 4, pp. 2197-2206, Nov. 2011.

[12] P. A. Ruiz, C. R. Philbrick, E. Zak et al., “Uncertainty management in the unit commitment problem,” IEEE Transactions on Power Systems, vol. 24, no. 2, pp. 642-651, Jun. 2009.

[13] M. Xie, J. Xiong, S. Ke et al., “Two-stage compensation algorithm for dynamic economic dispatch considering copula correlation of multi-wind farms generation,” IEEE Transactions on Sustainable Energy, vol. 8, no. 2, pp. 763-771, Jan. 2017.

[14] A. Callado, P. Escrìba, J. A. García-Moya et al., “Ensemble forecasting,” in Climate Change and Regional/Local Responses, Croatia: InTech, 2013, pp. 3-58.

[15] P. Pierre, M. Henrik, N. H. Aa et al., “From probabilistic forecasts to statistical scenarios of short-term wind power production,” Wind Energy, vol. 12, no. 1, pp. 51-62, Jan. 2010.

[16] J. Tastu, P. Pinson, and H. Madsen, “Space-time trajectories of wind power generation: parameterized precision matrices under a gaussian copula approach,” in Lecture Notes in Statistics, Heidelberg: Springer, 2014, pp. 267-296.

[17] M. Yang, Y. Lin, S. Zhu et al., “Multi-dimensional scenario forecast for generation of multiple wind farms,” Journal of Modern Power Systems and Clean Energy, vol. 3, no. 3, pp. 361-370, May 2015.

[18] H. V. Haghí and S. Lotfifard, “Spatiotemporal modeling of wind generation for optimal energy storage sizing,” IEEE Transactions on Sustainable Energy, vol. 6, no. 1, pp. 113-121, Jul. 2015.

[19] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, USA, Jan. 7-9, 2007, pp. 1027-1035.

[20] L. Durante and C. Sempi, “Copula theory: an introduction,” in Copula Theory and Its Applications, Heidelberg: Springer, 2010, pp. 3-24.

[21] Z. Wang, W. Wang, and B. Wang, “Regional wind power forecasting model with NWP grid data optimized,” Front Energy, vol. 2, no. 11, pp. 175-183, May 2017.

[22] Z. Wang, W. Wang, C. Liu et al., “Probabilistic forecast for multiple wind farms based on regular vine copulas,” IEEE Transactions on Power Systems, vol. 33, no. 1, pp. 578-589, Apr. 2018.

[23] D. Kurowicka and H. Joe, “Dependence modeling: vine copula handbook,” Singapore: World Scientific, 2011, p. 39.

[24] E. C. Brechmann, C. Czado, and K. Aas, “Truncated regular vines in high dimensions with application to financial data,” Canadian Journal of Statistics, vol. 40, no. 1, pp. 68-85, 2012.

[25] J. Dissmann, E. C. Brechmann, C. Czado et al., “Selecting and estimating regular vine copulæ and application to financial returns,” Data Analysis, vol. 59, pp. 52-69, 2013.

[26] T. H. Cormen, C. E. Leiserson, R. L. Rivest et al., “Introduction to algorithms,” 2nd ed, Cambridge: MIT Press and McGraw-Hill, 2001, p. 561-579.

[27] H. Akaike, “Information theory and an extension of the likelihood ratio principle,” in Springer Series in Statistics, Heidelberg: Springer, 1973, p. 281.

[28] C. Genest and A. Favre, “Everything you always wanted to know about copula modeling but were afraid to ask,” Journal of Hydrologic Engineering, vol. 12, no. 4, pp. 347-368, Jul. 2007.

[29] P. Pinson, H. A. Nielsen, I. K. Moller et al., “Non-parametric probabilistic forecasts of wind power: required properties and evaluation,” Wind Energy, vol. 10, no. 6, pp. 497-516, Nov. 2007.

[30] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” Journal of the American Statistical Association, vol. 102, no. 477, pp. 359-378, Mar. 2007.

[31] T. Gneiting, L. I. Stanberry, E. P. Grimit et al., “Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds,” in Proceedings of 8th IEEE International Conference on Bioinformatics and BioEngineering, Athens, Greece, Oct. 2008, pp. 211-223.

Zhao Wang received the B.E. and M.E. degrees from Zhejiang University, Hangzhou, China, in 2011 and 2014, respectively, both in electrical engineering. He received the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2018. He is now an engineer at China Electric Power Research Institute (CEPRI), Beijing, China. His research interests include wind power forecasting, uncertainty estimation, and dependence modelling.

Weisheng Wang received the Ph.D. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1996. He is currently a professor at China Electric Power Research Institute (CEPRI), Beijing, China, with an appointment as the director of Renewable Energy Research Center of CEPRI. He is a senior member of IEEE, member of IEC TC82-WG3, standing member of board of directors of China Renewable Energy Society, etc. His main interests include research and consulting in the field of renewable energy generation and its grid integration.

Chun Liu is a professor at China Electric Power Research Institute (CEPRI), Beijing, China, with an appointment as the vice director of Renewable Energy Research Center of CEPRI. His research interests include power system operations and optimization, renewable energy integration.

Bo Wang is a senior engineer of China Electric Power Research Institute (CEPRI), Beijing, China. His research interests are mainly renewable generation forecasting and its uncertainty estimation.