A Clustering Based Scenario Generation Method for Stochastic Power System Analysis

Binghui Li, Jie Zhang
The University of Texas at Dallas
Richardson, TX 75080, USA
Email: {binghui.li, jiezhang}@utdallas.edu

***Abstract*—A critical step in stochastic optimization models of power system analysis is to select a set of appropriate scenarios and significant amounts of scenario generation methods exist in the literature. This paper develops a clustering based scenario generation method, which aims to improve the performance of existing scenario generation techniques by grouping a set of correlated wind farms into clusters according to their cross-correlations. Copula based models are utilized to model spatio-temporal correlations and the Gibbs sampling is then used to generate scenarios. Our results show that the generated scenarios based on clustered wind farms outperform existing approaches and can provide a better characterization of wind power uncertainties.**

***Index Terms*—Probabilistic forecast, Cluster analysis, Gibbs sampling, Scenario generation.**

I. Introduction

Stochastic programming represents a potentially promising technique to address uncertainties associated with wind power. However, one challenge in its practice is to select a set of appropriately weighted scenarios to represent the space of uncertainty. Such techniques usually involve fitting forecasted wind power or forecast errors to specific distributions [1] and scenarios are then generated by sampling the derived distributions [2]. Ma et al. [3] characterize forecast errors using empirical distributions and the inverse transformation method is applied to obtain a set of scenarios. Cui et al. [4] develop a generalized Gaussian mixture model to fit forecast errors of aggregated wind power from hundreds of wind farms and the fitted distribution is used to sample scenarios for probabilistic wind ramp forecasting.

In addition, an increasing number of studies have placed emphasis on spatio-temporal correlation in scenario generation. Typically, such correlation is modeled with multivariate joint distributions. For example, Pinson et al. [2] employ multivariate Gaussian distribution to describe correlations between wind power forecasts made at different lead times, and this method has been widely adopted in a significant number of recent studies. Nevertheless, modeling high dimensional multivariate non-Gaussian distributions can be challenging, and a common approach is to use copula [5]. By applying the marginal cumulative distribution functions to stochastic variables, the original variables are transformed from the original space into a common uniform domain, in which correlations among the original variable can be further characterized using copulas. Zhang et al. [6] model spatio-temporal correlations of clustered wind farms using a copula based model and implement a scenario generation method. A similar method is developed by Tang et al. [7] and the Gibbs sampling method is adopted to generate scenarios for a stochastic unit commitment (UC) model.

The above methods represent a considerable number of scenario generation approaches which can be used in stochastic UC and economic dispatch (ED) models. However, real-world power systems usually involve hundreds or even thousands of wind farms, and it is usually computationally prohibitive to generate scenarios for such a large number of wind farms. In this paper, we present an improved scenario generation method based on cluster analysis, where wind farms are grouped into clusters by correlation and each cluster has a reduced number of wind farms, which can be sampled using existing approaches. Our approach represents an improved technique based on existing methods and can be applied to real-world power systems with high wind penetration. Cluster analysis has been found in wind power forecast. For example, Dong et al. [8] applied a k-means method to group historical daily profiles of wind power into clusters and base their wind power prediction on the most similar cluster.

The remaining of this paper is organized as follows: Section II details the developed clustering based scenario generation method, including the cluster analysis and the sampling method. Section III presents the clustered wind farms, the generated scenarios, and a comparison with one existing study. Last, Section IV concludes this paper.

II. Clustering based scenario generation

A. Cluster analysis

Cluster analysis groups data objects based only on information provided by the data itself, where data objects in one cluster often share similar characteristics that are different from other clusters. It is sometimes referred to as unsupervised classification since no data labeling information is given, as opposed to supervised classification where class specific data labels are typically provided beforehand.

The similarity or distance between data objects are usually expressed in terms of mathematical functions called proximity measures, which can be ℓ_p norms, or the cohesion of the cluster, i.e., the cosine of the included angle between points. In this study, we use Pearson's correlation coefficient $r(x_i, x_j)$ as the proximity measure in this analysis since the wind farms

are clustered based on correlation, i.e., wind farms in a cluster should be more closely correlated to each other than they are to members of a different cluster.

A variety set of methods exist in cluster analysis. In this study, two widely used clustering techniques are adopted: k-means and k-medoids. Both are prototype-based, partitional methods that divide data objects into multiple non-overlapping clusters defined by prototypes. The only difference is the prototype in the k-means method is a centroid (c_k), which is usually given by the mean of all data objects in that cluster, whereas the prototype of the k-medoids method is selected as the most representative data object in that cluster and referred to as the medoid (m_k). While the medoid of a cluster is always an actual data object, the centroid almost never falls on an actual one.

Using the above notations, the objective function of the k-means method becomes:

$$\min \sum_{k=1}^{K} \sum_{x \in C_k} dist(c_k, x) \tag{1}$$

where, $dist()$ represents the selected proximity measure and the k-medoids method follows similar forms except c_k is replaced with m_k. Note that since higher r represents stronger correlation, hence higher similarity, the objective function minimizes $1 - r$ instead:

$$\min \sum_{k=1}^{K} \sum_{x \in C_k} [1 - r(c_k, x)] \tag{2}$$

To solve the optimization problem, one can exhaustively enumerate all possible ways of grouping objects into K clusters. However, this is often computationally prohibitive and heuristic iteration-based algorithms are typically used. Such heuristic algorithms often require initial conditions (i.e., centroids and medoids) as input and iteratively converge to a solution where the objective value no longer decreases. Although the heuristic algorithms can converge within an acceptable amount of time, they are sensitive to initial conditions and often converge to a local minimum rather than a global minimum. Therefore, multiple runs with randomly selected initial conditions are typically required. In this paper, the `kmeans` and `kmedoids` functions in MATLAB are employed and both functions are repeated for 10 times.

The validity of the cluster analysis can be quantitatively evaluated using a diverse set of metrics, many of which are based on cohesion and separation. For prototype-based clusters, cohesion measures the sum of proximities of members of a cluster to its centroid (or medoid), hence representing the overall similarity of individual clusters. By contrast, separation measures proximities between centroids (or medoids) of a pair of different clusters, hence indicating dissimilarities. While both metrics place emphasis on one aspect, we use the Silhouette coefficient, which accounts for both cohesion and separation, to give a more comprehensive evaluation of cluster validity. The Silhouette coefficient of data object $i \in C_k$ is given by:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{3}$$

where, a_i denotes the average proximity of data object i to all other objects in its cluster, and b_i indicats the smallest average proximity of data object i to all objects in any other clusters. To measure the overall validity, we take the average over all data objects:

$$S = \frac{1}{N} \sum_{i}^{N} s_i \tag{4}$$

The value of the Silhouette coefficient ranges from -1 to 1, with negative values indicating undesirable results. In addition, the optimal number of clusters K is selected based on the Silhouette coefficient.

B. Scenario generation

Given a cluster, the spatial correlations between N wind farms at the same time in that cluster is modeled using a multivariate joint probability distribution, whose cumulative density function (CDF) and probability density function (PDF) can be expressed by the following equations:

$$F_{a_1 \cdots a_N f_1 \cdots f_N} = Prob(X_1^a \le x_1^a, \cdots, X_n^a \le x_N^a, X_1^f \le x_1^f, \cdots, X_n^f \le x_N^f) \tag{5}$$

$$f_{a_1 \cdots a_N f_1 \cdots f_N} = \frac{\partial^{2N} F_{a_1 \cdots a_N f_1 \cdots f_N}}{\partial X_1^a \cdots \partial X_N^a \partial X_1^f \cdots \partial X_N^f} \tag{6}$$

Therefore, scenario generation given deterministically forecasted wind power becomes sampling a multivariate distribution of the actual wind power $(x_1^a, \cdots, x_N^a)$ conditioned on the forecasted wind power $(x_1^f, \cdots, x_N^f)$:

$$F_{a_1 \cdots a_N | f_1 \cdots f_N} = Prob(X_1^a \le x_1^a, \cdots, X_n^a \le x_N^a | X_1^f = x_1^f, \cdots, X_n^f = x_N^f) \tag{7}$$

$$f_{a_1 \cdots a_N | f_1 \cdots f_N} = \frac{\partial^{N} F_{a_1 \cdots a_N | f_1 \cdots f_N}}{\partial X_1^a \cdots \partial X_N^a} \tag{8}$$

In this analysis, we follow the method developed by Tang et al. [7] by modeling the conditional multivariate distribution using marginal distributions of individual variables, and their correlations are described by a copula. In addition, Gibbs sampling is applied to simulate sampling of the multidimensional random variable by sequentially sampling each component.

1) Gibbs sampling: The Gibbs sampling method is a Markov chain Monte Carlo algorithm and the idea is to iteratively sample only one variable or a block of variables at a time from its distribution conditioned on the remaining variables. Therefore, Gibbs sampling converts sampling a multivariate distribution into sampling a set of conditional univariate distributions. A detailed procedure is given in Algorithm 1. Note in this analysis, we use the forecasted wind power as the initial input.

Algorithm 1 Gibbs sampling

procedure GIBBS SAMPLER
Initialize $x_i^{a(0)} \leftarrow x_1^f, \forall i = 1, \cdots, N$
for scenario $\xi = 1, \cdots, \Xi$ **do** Sample:
$X_1^a | x_2^{a(\xi-1)}, \cdots, x_N^{a(\xi-1)}, x_1^f, \cdots, x_N^f$
$X_2^a | x_1^{a(\xi)}, x_3^{a(\xi-1)}, \cdots, x_N^{a(\xi-1)}, x_1^f, \cdots, x_N^f$
...
$X_N^a | x_1^{a(\xi)}, \cdots, x_{N-1}^{a(\xi)}, x_1^f, \cdots, x_N^f$
end for
end procedure

2) The copula approach: According to Sklar's theorem, any multivariate joint distribution can be written in terms of marginal distributions of each component and a copula that describes the dependence structure between the components. Therefore, the joint CDF in (5) can be transformed to

$$F_{a_1 \cdots a_N f_1 \cdots f_N} = C_{2N}(u_1^a, \cdots, u_N^a, u_1^f, \cdots, u_N^f) \quad (9)$$

where the copula $C_{2N} : [0,1]^{2N} \rightarrow [0,1]$ is a continuous function and $u_i^a, u_i^f \in [0,1]$ are cumulative marginal probabilities associated with X_i^a and X_i^f:

$$u_i^a = F_{a_i}(x_i^a) = Prob(X_i^a \leq x_i^a) \quad (10)$$

$$u_i^f = F_{f_i}(x_i^f) = Prob(X_i^f \leq x_i^f) \quad (11)$$

Similarly, the joint PDF can be expressed as

$$f_{a_1 \cdots a_N f_1 \cdots f_N} = c_{2N} \cdot \prod_{i=1}^{N} (f_{a_i} f_{f_i}) \quad (12)$$

where, f_{a_i} and f_{f_i} are marginal PDFs associated with X_i^a and X_i^f, respectively, and $c_{2N} : [0,1]^{2N} \rightarrow \mathbb{R}_+$ is:

$$c_{2N} = \frac{\partial^{2N} C_{2N}}{\partial u_1^a \cdots \partial u_N^a \partial u_1^f \cdots \partial u_N^f} \quad (13)$$

Therefore, by applying Bayes' theorem, the conditional univariate distribution of X_i^a is given by:

$$F_{a_i | a_1 \cdots a_{i-1} a_{i+1} \cdots a_N f_1 \cdots f_N} = \frac{Prob(X_i^a \leq x_i^a, X_j^a = x_j^a, X_k^f = x_k^f)}{Prob(X_j^a = x_j^a, X_k^f = x_k^f)}$$
$$j = 1, \cdots, i-1, i+1, \cdots, N, k = 1, \cdots, N \quad (14)$$

where, the numerator can be written as:

$$Prob(X_i^a \leq x_i^a, X_j^a = x_j^a, X_k^f = x_k^f) = \int_0^{x_i^a} f_{a_1 \cdots a_N f_1 \cdots f_N} dX_i^a = \int_0^{x_i^a} c_{2N} \cdot \prod_{j=1}^{N} (f_{a_j} f_{f_j}) dX_i^a$$
$$= \frac{\partial^{2N-1} C_{2N}}{\partial X_1^a \cdots \partial X_{i-1}^a \partial X_{i+1}^a \cdots \partial X_N^a \partial X_1^f \cdots \partial X_N^f} \cdot \frac{\prod_{j=1}^{N} (f_{a_j} f_{f_j})}{f_{a_i}} \quad (15)$$

and the denominator can be written as the following by applying (12):

$$Prob(X_j^a = x_j^a, X_k^f = x_k^f) = f_{a_1 \cdots a_N \cdots a_{i-1} a_{i+1} \cdots a_N f_1 \cdots f_N} = c_{2N-1} \cdot \frac{\prod_{j=1}^{N} (f_{a_j} f_{f_j})}{f_{a_i}} \quad (16)$$

In summary, the conditional univariate CDF is

$$F_{a_i | a_1 \cdots a_{i-1} a_{i+1} \cdots a_N f_1 \cdots f_N} = \frac{\partial^{2N-1} C_{2N}}{\partial X_1^a \cdots \partial X_{i-1}^a \partial X_{i+1}^a \cdots \partial X_N^a \partial X_1^f \cdots \partial X_N^f} \cdot \frac{1}{c_{2N-1}} \quad (17)$$

Once the univariate CDF is given, the Gibbs sampler defined in Algorithm 1 is used to sequentially produce Ξ samples. Note that the samples obtained at this stage must be mapped into the wind power domain by taking the inverse of the marginal CDFs defined in (10). In this study, the logit-normal distribution is used to fit the marginal distributions, and the copula is selected from Gaussian, student t, and the Archimedean family based on the Bayesian information criterion (BIC).

3) Temporal correlation: Samples produced from the above method represent simultaneous power production from multiple spatially correlated wind farms. However, as suggested by previous studies [2], power produced from one wind farm at different times is typically temporally correlated, and such correlation can be modeled using a multivariate Gaussian distribution:

$$\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_i}) \quad (18)$$

where, $\mathbf{X}_i = (X_{i,t+1}, \cdots, X_{i,t+T})$ is a T-dim vector representing wind power at time 1 to T, $\mathbf{0}$ is a T-dim vector of zeros, and $\boldsymbol{\Sigma}_i$ is a covariance matrix of which each entry is defined by:

$$\text{cov}(X_{i,t+k_1}, X_{i,t+k_2}) = \exp(-\frac{|k_2 - k_1|}{v}) \quad (19)$$

The covariance between wind power at different times implies the temporal correlation decays exponentially and is dictated by the parameter v. Although existing literature suggests different values of v, we select the value of v by fitting $\boldsymbol{\Sigma}_i$ to historical data such that the sum of the squares of differences are minimized.

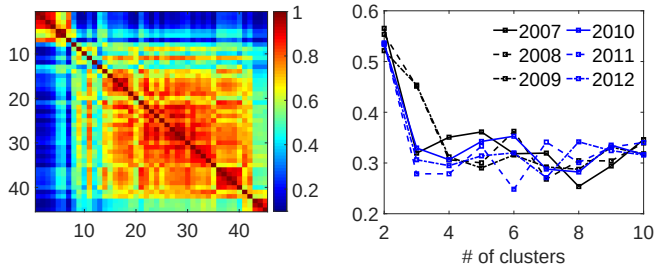
III. RESULTS

In this study, the day-ahead forecasts drawn from NREL's Wind Integration National Dataset (WIND) toolkit [9] are used as model inputs. The WIND toolkit provides synthetic wind turbine output power as well as meteorological data for more than 126,000 sites in the contiguous United States from 2007 to 2012. The temporal granularity is every 5 minutes for the actual power data and hourly for all forecasts. Therefore, hourly average is calculated for the actual data for consistency. The target day in this paper is May 10, 2012 and the model is trained with the data from Jan. 1, 2007 to May 9, 2012. The day-ahead forecasts for May 10, 2012 are then used as the

initial inputs to the Gibbs sampler. We select 45 sites located in Texas and the total capacity is 646 MW.

A. Cluster analysis

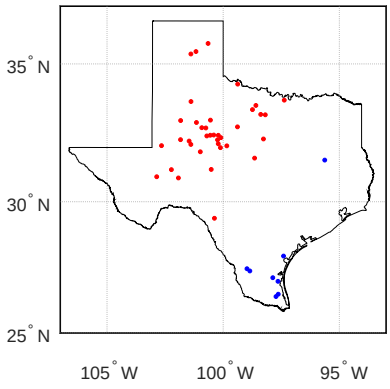
Correlation coefficients between pairs of the 45 wind farms are visualized in Fig. 1a, where warmer color indicates stronger correlation. Since the coefficient is also the proximity measure selected in this analysis, Fig. 1a also represents the proximity matrix. It clearly presents a very strong, block-diagonal pattern and gives an intuitive view of the strength of interdependence between wind farms. Fig. 1b indicates grouping all wind farms into 2 clusters always results in the highest overall Silhouette coefficient, despite yearly variations. Therefore, $K = 2$ is used in the following analysis.



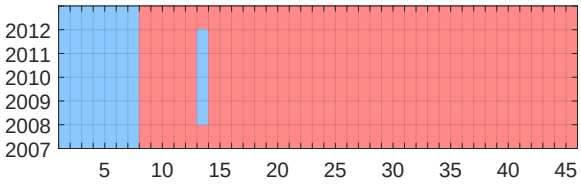
(a) Coefficients of correlation. (b) The Silhouette coefficients.

Fig. 1: The results from the cluster analysis: (a) coefficients of correlation between all wind farms, and (b) The Silhouette coefficients as a function of the number of clusters K.

When $K = 2$, the results from the cluster analysis are presented in Fig. 2. Fig. 2b distinguishes members of each cluster by colors, where the cluster in blue (hereafter "the blue cluster") has 7 members from 2008 to 2011 and 8 members in 2007 and 2012. This is not surprising, since as shown in Fig. 2a, the rightmost wind farm is equally distant from both clusters, while the remaining wind farms are naturally grouped by their geographical distances to each other. The red group mainly scatters over the northwestern Texas and members of the blue group are located in the concentrated southern tip. Since geographical closeness of two wind farms usually indicates stronger correlation in power production, the rightmost wind farm can be categorized into either cluster with equal odds. In this analysis, this wind farm is classified as the blue cluster since Fig. 2b shows higher frequency of it being blue. By taking the mean of correlation coefficients of all pairs in both clusters, the mean of the blue cluster is 0.6965, slightly higher than that of the other cluster (hereafter "the red cluster"), 0.6719, indicating slightly stronger intra-cluster correlation. In addition, both clusters exhibit higher means than the mean of all wind farms together, which is 0.5587, further validating the soundness of our results.



(a) Locations of wind farms.



(b) The cluster members from 2007 to 2012.

Fig. 2: The results from the cluster analysis when $K = 2$: (a) locations of clustered wind farms, and (b) the cluster members from 2007 to 2012 where each row represents one year and each grid represents one wind farm. Clusters are distinguished by colors in both figures.

B. Generated scenarios

By splinting all wind farms into two clusters, the scenario generation method defined in Section II-B is applied to both clusters. The blue cluster has 8 wind farms and the total capacity is 126 MW and the red cluster has 37 wind farms and the total capacity is 520 MW. Since Gibbs sampling requires a large number of iterations before the samples reach stationary distribution, 3,000 samples are generated for each cluster and the first 2,900 samples are discarded (known as the "burn-in" process) and only the last 100 samples are used in further analysis. The implementation is coded in MATLAB and executed on a high performance computing platform to take advantage of its parallel computing capability. A compute node equipped with a 20-core Xeon E5-2698 CPU and 256 GB memory is used, where the model training and scenario generation are both parallelized. It takes 40 minutes to train the model and another 15 minutes to generate 3,000 scenarios for 37 wind farms of the red cluster.

Fig. 3 shows the generated scenarios, where only the total power productions from both clusters are displayed due to page limit. The upper and lower bounds of the $(1-\alpha)$ confidence intervals of the total power productions are calculated by summing up the $\alpha/2$ and $(1-\alpha/2)$ percentiles of empirical distributions of the generated scenarios for each wind farm. It shows that the majority of the generated scenarios lie within the 50% confidence intervals. Surprisingly, the majority of the day-ahead forecasts are not covered by the 10% confidence

interval, as shown in Fig. 3b, where the day-ahead forecasted aggregated output of the red cluster is approximately 20 MW higher than the upper bound of the 10% confidence interval on average.

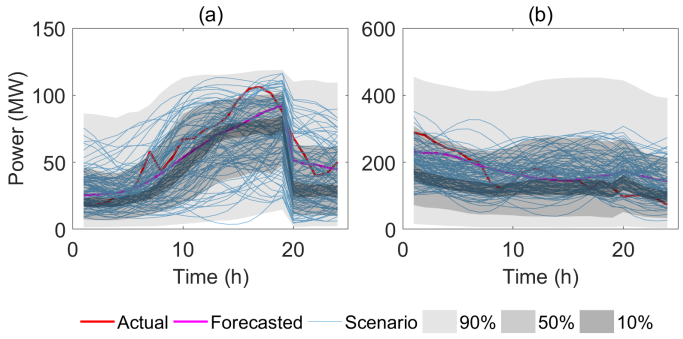


Fig. 3: Aggregated actual power, day-ahead forecasted power, and generated scenarios of (a) the blue cluster and (b) the red cluster. Note the shaded bands represent intervals at a given confidence level of the generated scenarios.

To evaluate the validity of the generated scenarios, the mean absolute error (MAE) and the root mean square error (RMSE) of the generated scenarios are shown in Table I and compared with a previous study by Wang et al. [10]. Note that the RSME from [10] is obtained by taking the square root of their variance (VAR). We must note here that our study only generate scenarios for 24 hours using models trained with hourly day-ahead forecasts, while their model generate scenarios for 52 weeks using 4-hour ahead, 6-hour ahead, and day-ahead forecasts. While their model only gives scenarios for individual wind farms and does not consider spatial correlation, it shows that our results outperform theirs in terms of both MAE and RSME, implying improved performance. In addition, the results from the blue cluster show better accuracy, possibly due to stronger intra-cluster correlation than the red cluster.

TABLE I: MAE and RSME of the generated scenarios.

Method	MAE	RMSE
The blue cluster	0.2676	0.3453
The red cluster	0.2782	0.3560
Wang et al. [10]	0.3542	0.3713

IV. Conclusion

This paper presents a scenario generation approach based on cluster analysis in conjunction with existing multivariate sampling techniques. Two clustering methods, k-means and k-medoids, are applied to divide 45 wind farms into two clusters. Existing multivariate sampling techniques based on copula and Gibbs sampling are then applied to generate 100 scenarios for each cluster.

As shown in the results from the cluster analysis, the mean correlation coefficients of individual clusters are higher than the mean of all wind farms together, indicating stronger intra-cluster correlations. Our results suggest the generated scenarios for the cluster with higher intra-cluster correlation are better in terms of both MAE and RMSE. In addition, the comparison with a previous study also indicates improved performance from the adoption of cluster analysis. Therefore, the scenario generation can benefit from clustering wind farms into groups with higher interdependence.

The results from this study can be used directly as inputs to stochastic UC and ED models. However, the number of generated scenarios must be reduced to an adequate number before the model can be solved efficiently. Scenario reduction techniques can be found in existing literature [11], [12] and will be applied to extend this study to real-life power grid analysis with large scale wind penetration.

Acknowledgment

This work was supported by the National Renewable Energy Laboratory under Subcontract No. XAT-8-82151-01 (under the U.S. Department of Energy Prime Contract No. DE-AC36-08GO28308).

References

[1] P. Pinson, "Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 61, no. 4, pp. 555–576, 2012.

[2] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl, "From probabilistic forecasts to statistical scenarios of short-term wind power production," *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, vol. 12, no. 1, pp. 51–62, 2009.

[3] X.-Y. Ma, Y.-Z. Sun, and H.-L. Fang, "Scenario generation of wind power based on statistical uncertainty and variability," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 4, pp. 894–904, 2013.

[4] M. Cui, J. Zhang, Q. Wang, V. Krishnan, and B.-M. Hodge, "A data-driven methodology for probabilistic wind power ramp forecasting," *IEEE Transactions on Smart Grid*, 2017.

[5] G. Papaefthymiou and D. Kurowicka, "Using copulas for modeling stochastic dependence in power system uncertainty analysis," *IEEE Transactions on Power Systems*, vol. 24, no. 1, pp. 40–49, 2009.

[6] N. Zhang, C. Kang, Q. Xu, C. Jiang, Z. Chen, and J. Liu, "Modelling and simulating the spatio-temporal correlations of clustered wind power using copula," *Journal of Electrical Engineering & Technology*, vol. 8, no. 6, pp. 1615–1625, 2013.

[7] C. Tang, Y. Wang, J. Xu, Y. Sun, and B. Zhang, "Efficient scenario generation of multiple renewable power plants considering spatial and temporal correlations," *Applied Energy*, vol. 221, pp. 348–357, 2018.

[8] L. Dong, L. Wang, S. F. Khahro, S. Gao, and X. Liao, "Wind power day-ahead prediction with cluster analysis of nwp," *Renewable and Sustainable Energy Reviews*, vol. 60, pp. 1206–1212, 2016.

[9] C. Draxl, A. Clifton, B.-M. Hodge, and J. McCaa, "The wind integration national dataset (wind) toolkit," *Applied Energy*, vol. 151, pp. 355–366, 2015.

[10] Z. Wang, C. Shen, and F. Liu, "A conditional model of wind power forecast errors and its application in scenario generation," *Applied Energy*, vol. 212, pp. 771–785, 2018.

[11] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, "Scenario reduction in stochastic programming," *Mathematical programming*, vol. 95, no. 3, pp. 493–511, 2003.

[12] Y. Dvorkin, Y. Wang, H. Pandzic, and D. Kirschen, "Comparison of scenario reduction techniques for the stochastic unit commitment," in *PES General Meeting— Conference & Exposition, 2014 IEEE*, pp. 1–5, IEEE, 2014.