

A clustering-based scenario generation framework for power market simulation with wind integration

Cite as: J. Renewable Sustainable Energy 12, 036301 (2020); doi: 10.1063/5.0006480

Submitted: 2 March 2020 · Accepted: 12 May 2020 ·

Published Online: 3 June 2020



View Online



Export Citation



CrossMark

Binghui Li,¹ Kwami Sedzro,² Xin Fang,² Bri-Mathias Hodge,^{2,a)} and Jie Zhang^{1,b)}

AFFILIATIONS

¹The University of Texas at Dallas, Richardson, Texas 75080, USA

²National Renewable Energy Laboratory, Golden, Colorado 80401, USA

Note: This paper is part of the Special Collection on Best Practices in Renewable Energy Resourcing and Integration.

^{a)}Also at: University of Colorado Boulder, Boulder, Colorado 80309, USA.

^{b)}Author to whom correspondence should be addressed: 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 numbers 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 sites into clusters according to their cross-correlations. Copula based models are utilized to model spatiotemporal correlations and the Gibbs sampling is then used to generate scenarios for day-ahead markets. Our results show that the generated scenarios based on clustered wind sites outperform existing approaches in terms of reliability and sharpness and can reduce the total computational time for scenario generation and reduction significantly. The clustering-based framework can therefore provide a better support for real-world market simulations with high wind penetration.

Published under license by AIP Publishing. <https://doi.org/10.1063/5.0006480>

NOMENCLATURE

Sets

\mathbf{B}	Buses
$\mathbf{B}^G \subseteq \mathbf{B}$	Buses with generators
$\mathbf{B}^L \subseteq \mathbf{B}$	Load buses
\mathbf{G}	Generators
$\mathbf{G}_b \subseteq \mathbf{G}$	Generators on bus $b \in \mathbf{B}^G$
$\mathbf{G}^T \subseteq \mathbf{G}$	Thermal generators
$\mathbf{G}^W \subseteq \mathbf{G}$	Wind generators
\mathbf{L}	Transmission lines, or branches
\mathbf{T}	Time slices
Ξ	Wind power scenarios

Parameters

C_g	Slope of the cost function of generator $g \in \mathbf{G}^T$, in \$/MWh
C^{AL}	System-wide load curtailment cost, \$/MWh
C^{AR}	System-wide regulating up/down reserve curtailment cost, \$/MWh
C^{AS}	System-wide spinning reserve curtailment cost, \$/MWh

C_g^{SU}, C_g^{SD}	Start-up and shut-down cost of generator $g \in \mathbf{G}^T$
C_g^0	Intercept of the cost function of generator $g \in \mathbf{G}^T$, in \$
$D_{b,t}$	Demand at bus $b \in \mathbf{B}^L$ in time $t \in \mathbf{T}$
K	Number of clusters during scenario generation
$p_{g,0}$	Initial power production of generator g
$P_{g,t}^F$	Forecasted power at time t from non-thermal generator $g \in \mathbf{G} \setminus \mathbf{G}^T$
P_g^{max}, P_g^{min}	Upper and lower bounds of power production from thermal generator $g \in \mathbf{G}^T$
R_g^{SU}, R_g^{SD}	Start-up/shut-down ramp rates of generator g
R_g^U, R_g^D	Ramp up/down rates of generator g
RQ_t^{RD}	System regulation down reserve requirement at time t
RQ_t^{RU}	System regulation up reserve requirement at time t
RQ_t^S	System spinning reserve requirement at time t
$SF_{\ell,b}$	Power shift factor, power flow increment on branch ℓ due to power injected at bus b
T_g^U, T_g^D	The minimum online/offline time of generator g
T_g^{U0}, T_g^{D0}	Number of initial time slices generator g must be online/offline
$u_{i,t}^a$	Marginal CDF of $x_{i,t}^a$

$u_{i,t}^f$	Marginal CDF of $x_{i,t}^f$
$v_{g,0}$	Initial commitment status of generator g
Λ_ℓ	Power limit of branch $\ell \in \mathbf{L}$

Variables

$p_{g,t} \in \mathbb{R}_+$	Power production from generator $g \in \mathbf{G}$ in time $t \in \mathbf{T}$
$\bar{p}_{g,t} \in \mathbb{R}_+$	Maximum power available from generator $g \in \mathbf{G}$ in time $t \in \mathbf{T}$
$r_{g,t}^{RD} \in \mathbb{R}_+$	Regulation down reserve provided by generator $g \in \mathbf{G}^T$ in time t
$r_{g,t}^{RU} \in \mathbb{R}_+$	Regulation up reserve provided by generator $g \in \mathbf{G}^T$ in time t
$r_{g,t}^S \in \mathbb{R}_+$	Spinning reserve provided by generator $g \in \mathbf{G}^T$ in time t
$v_{g,t}$	Commitment status of generator $g \in \mathbf{G}^T$ in time $t \in \mathbf{T}$, binary
$x_{i,t}^a, x_{i,t}^f$	Samples of random variables $X_{i,t}^a$ and $X_{i,t}^f$
$x_{i,t}^{a,\xi}$	Realization of variables $X_{i,t}^a$ in scenario ξ
$X_{i,t}^a, X_{i,t}^f$	Random variables indicating actual and forecasted wind power at site i in time t
$z_{g,t}^P \in \mathbb{R}_+$	Production cost of generator $g \in \mathbf{G}^T$ in time $t \in \mathbf{T}$
$z_{g,t}^{SD} \in \mathbb{R}_+$	Shut-down cost of generator $g \in \mathbf{G}^T$ in time $t \in \mathbf{T}$
$z_{g,t}^{SU} \in \mathbb{R}_+$	Start-up cost of generator $g \in \mathbf{G}^T$ in time $t \in \mathbf{T}$
$z_t^\Delta \in \mathbb{R}_+$	Load and reserve curtailment cost in time $t \in \mathbf{T}$
$\delta_{b,t} \in \mathbb{R}_+$	Load curtailment at bus b in time t
$\delta_t^{RD} \in \mathbb{R}_+$	System-wide regulating down reserve curtailment in time t
$\delta_t^{RU} \in \mathbb{R}_+$	System-wide regulating up reserve curtailment in time t
$\delta_t^S \in \mathbb{R}_+$	System-wide spinning reserve curtailment in time t

I. INTRODUCTION

Stochastic programming represents a potentially promising technique to address uncertainties associated with wind power¹⁻³ and has been studied widely in the recent decades⁴⁻⁶ in a variety of fields.⁷⁻¹⁰ 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,¹¹ where scenarios are then generated by sampling the derived distributions.¹² Such distributions can be either parametric distributions, such as Gaussian⁵ and Generalized Gaussian distributions,¹³ the Beta distribution,^{14,15} the Weibull distribution,¹⁶ and the logit-normal distribution,¹² or non-parametric distributions.¹⁷ Ma *et al.*¹⁸ characterized forecast errors using empirical distributions and the inverse transformation method is applied to obtain a set of scenarios. Cui *et al.*¹³ developed 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.

While sampling methods can be applied to these distributions to generate scenarios for a single wind farm or aggregated power output of multiple wind farms,^{5,11,13,15,17,18} an increasing number of studies have placed emphasis on spatiotemporal correlation in scenario

generation.¹⁹⁻²¹ Typically, such correlation is modeled with multivariate joint distributions. For example, Pinson *et al.*¹² employed a multivariate Gaussian distribution to describe correlations between wind power forecasts made at different lead times, and this method has been widely adopted in numerous studies recently. Other commonly used methods include Cholesky decomposition,²² moment matching techniques,²³ and the generalized dynamic factor model.²⁴ In addition, machine learning-based methods are also employed to generate scenarios in the latest studies. Such methods include artificial neural network (ANN),²⁵ quantile regression convolutional neural network (QRCNN),²⁶ radial basis function neural network (RBFNN),¹⁸ generative adversarial networks (GAN),^{27,28} and support vector machine (SVM).²⁹ Particularly, techniques that are used to generate probabilistic forecasts can be used to produce discrete scenarios by sampling the predictive distributions to avoid reliance on prior distributions.³⁰ Nevertheless, modeling high dimensional multivariate non-Gaussian distributions can be challenging, and a common approach is to use copula.³¹ 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. The copula-based methods have been used extensively in the literature recently.^{12,32-35} For example, Zhang *et al.*¹⁹ modeled spatiotemporal correlations of clustered wind farms using a copula based model and implemented a scenario generation method. A similar method is developed by Tang *et al.*²⁰ and the Gibbs sampling method is adopted to generate scenarios for a stochastic economic dispatch (ED) model.

The above methods represent a considerable number of scenario generation approaches. However, real-world power systems usually involve hundreds or even thousands of wind farms, while sampling random variables of such high dimensions is usually computationally prohibitive, especially when spatial-temporal correlation is explicitly considered. Recently, data mining techniques have been applied to power system studies to process large volumes of data in a diverse set of literature studies.^{36,37} Typical applications include cable layout design, evaluation of transfer capabilities, reliability of networks, and also wind power forecasting.³⁸ In the study by Vallee *et al.*,³⁷ a clustering method is developed to group 94 wind areas into 14 clusters, which are then used to generate samples of wind speed to evaluate system reliability through a non-sequential Monte Carlo simulation.

In this paper, we present a scenario generation framework based on cluster analysis, where wind sites are grouped into clusters by correlation. The framework is aimed to improve the quality of generated scenarios at less time expense by maximizing intra-cluster correlation and minimizing inter-cluster correlation. By clustering wind sites into a number of smaller clusters, existing methods can be employed to generate scenarios for each cluster. The generated scenarios of wind power are then reduced and used as inputs to a stochastic day-ahead unit commitment (DAUC) model, which uses a hedging strategy to clear the day-ahead market to determine commitment statuses of thermal units. Given the commitment statuses of thermal generators, a real-time ED model is then solved to evaluate the system cost-effectiveness and reliability. A synthetic, 2000-bus network built on the footprint of the Electric Reliability Council of Texas (ERCOT)³⁹ is employed to demonstrate the applicability of the proposed method in real-world power systems. Our approach represents an enhanced

framework based on existing methods and can be applied to real-world power systems with high wind penetration.

The main contributions of our work include (i) enhancing existing scenario generation methods by substantially improving the quality of generated scenarios with reduced time consumption, (ii) demonstrating the proposed framework with a real-world sized power system and scrutinizing the trade-off between the number of clusters and a variety of evaluation metrics, and (iii) using the generated scenarios as inputs to a stochastic power market simulation of a synthetic network built on the footprint of the Texas power system and assessing its impact in market operation and scheduling. The proposed framework can be extended by integrating with other scenario generation methods and applied to real-world power systems with high wind penetration.

The remainder of this paper is organized as follows: Sec. II details the developed clustering-based framework. Section III formulates the power market simulation model. Section IV summarizes the data and assumptions. Sections V and VI present the generated scenarios and the power market simulation, respectively. Section VII concludes this paper.

II. CLUSTERING BASED SCENARIO GENERATION

A. Cluster analysis

While numerous studies suggest that including correlations across wind farms can provide better support for existing scenario generation methods,^{19–21} these studies typically include all wind farms in a single group. However, wind farms in a real-world power network usually scatter over a wide area since geographical diversification can effectively decrease fluctuations in aggregate output and improve overall economic viability.^{40,41} In addition, correlations normally decay with distances, which results in a weak correlation between two wind farms if they are remotely located. Therefore, it is questionable whether the generated scenarios can benefit from including weakly correlated wind farms and a more appropriate approach is to group wind farms into clusters based on their cross-correlations.⁴²

Cluster analysis groups data objects based only on information provided by the data itself, where data objects in one cluster often share similar characteristics and are distinct from other clusters. The similarity or distance between data objects is usually expressed in terms of proximity measures, such as ℓ_p norms, or the cohesion of the cluster. In this study, we use Pearson's correlation coefficient $\rho(x_i, x_j)$ as the proximity measure since the wind farms are clustered based on correlation, i.e., wind farms in the same cluster should be more closely correlated with each other, while correlations between two clusters are less strong. The formula for $\rho(x_i, x_j)$ is given by the equation below:

$$\rho(x_i, x_j) = \frac{\text{cov}(x_i, x_j)}{\sigma_{x_i} \sigma_{x_j}}. \quad (1)$$

A variety of methods exist in cluster analysis.⁴³ In this study, we use the k -means method. The k -means method is a widely used prototype-based, partitional method that selects the mean of all data objects in a cluster as the prototype. Let C denotes a set of data objects, which can be grouped into K clusters and let $x \in C_k$ denotes the data objects in the k th cluster, the objective function of the k -means method becomes

$$\min \sum_{k=1}^K \sum_{x \in C_k} \text{dist}(c_k, x), \quad (2)$$

where $\text{dist}()$ represents the selected proximity measure. Note that since higher ρ represents stronger correlation, hence higher similarity, the objective function minimizes $1 - \rho$ instead:

$$\min \sum_{k=1}^K \sum_{x \in C_k} [1 - \rho(c_k, x)]. \quad (3)$$

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) 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 k -means function in MATLAB is employed and repeated ten times using different initial cluster centroids.

B. Scenario generation

After all wind sites are grouped into K clusters, existing scenario generation methods can be applied to each cluster. Given a cluster with N wind sites, the spatial correlations of wind power production across all sites in the cluster can be modeled using a multivariate joint probability distribution, whose cumulative density function (CDF) and probability density function (PDF) are given below:

$$\begin{aligned} & F_{a_1 \dots a_N | f_1 \dots f_N}(x_{1,t}^a, \dots, x_{N,t}^a, x_{1,t}^f, \dots, x_{N,t}^f) \\ &= \text{Prob}(X_{1,t}^a \leq x_{1,t}^a, \dots, X_{N,t}^a \leq x_{N,t}^a, X_{1,t}^f \leq x_{1,t}^f, \dots, X_{N,t}^f \leq x_{N,t}^f), \end{aligned} \quad (4a)$$

$$\begin{aligned} & f_{a_1 \dots a_N | f_1 \dots f_N}(x_{1,t}^a, \dots, x_{N,t}^a, x_{1,t}^f, \dots, x_{N,t}^f) \\ &= \frac{\partial^{2N} F_{a_1 \dots a_N | f_1 \dots f_N}}{\partial X_1^a \dots \partial X_N^a \partial X_1^f \dots \partial X_N^f}, \end{aligned} \quad (4b)$$

where $X_{i,t}^a$ and $X_{i,t}^f$ denote actual and forecasted wind power at site i during time t , respectively. Note that the forecasts can be deterministic forecasts with any lead time and in our analysis, day-ahead forecasts are used. Therefore, scenario generation given wind power forecast becomes sampling the distribution of actual wind power ($X_{1,t}^a, \dots, X_{N,t}^a$) conditioned on the forecasted wind power ($x_{1,t}^f, \dots, x_{N,t}^f$),

$$\begin{aligned} & F_{a_1 \dots a_N | f_1 \dots f_N}(x_{1,t}^a, \dots, x_{N,t}^a) \\ &= \text{Prob}(X_{1,t}^a \leq x_{1,t}^a, \dots, X_{N,t}^a \leq x_{N,t}^a | X_{1,t}^f = x_{1,t}^f, \dots, X_{N,t}^f = x_{N,t}^f), \end{aligned} \quad (5a)$$

$$f_{a_1 \dots a_N | f_1 \dots f_N}(x_{1,t}^a, \dots, x_{N,t}^a) = \frac{\partial^N F_{a_1 \dots a_N | f_1 \dots f_N}}{\partial X_1^a \dots \partial X_N^a}. \quad (5b)$$

Here, we follow Tang *et al.*²⁰ by modeling the multivariate distribution with marginal distributions of individual variables and a copula

to link their interdependence. The spatial correlation is modeled with the copula, while the temporal correlation is modeled with a multivariate Gaussian distribution. Next, Gibbs sampling is adopted to simulate sampling of the multivariate random variable by sequentially sampling each component.

1. Gibbs sampling

The Gibbs sampling method is a Markov chain Monte Carlo algorithm^{44,45} 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. Suppose we have a two-dimensional random variable (X, Y) , the Gibbs sampler generates a sequence (known as a Gibbs sequence) by sampling each component conditioned on the other component:

$$y^{(0)}, x^{(0)}, y^{(1)}, x^{(1)}, y^{(2)}, x^{(2)}, \dots, y^{(k)}, x^{(k)}.$$

The initial value $y^{(0)}$ is specified and the rest of the sequence is obtained by iteratively sampling,

$$\begin{aligned} X^{(j)} &\sim f(X|Y^{(j)} = y^{(j)}), \\ Y^{(j+1)} &\sim f(Y|X^{(j)} = x^{(j)}). \end{aligned}$$

The process of generating a Gibbs sequence is Gibbs sampling. It shows that under certain conditions, the distributions of $X^{(k)}$ and $Y^{(k)}$ converge to $f(X)$ and $f(Y)$ (i.e., their marginal distributions), respectively, as $k \rightarrow \infty$. Without knowing the joint distribution, we can generate a set of samples that follow their marginal distributions. It can also be extended to high dimensional space as well. Therefore, Gibbs sampling converts sampling a multivariate distribution into sequentially sampling a set of univariate distributions, i.e., sampling the multivariate distribution in Eq. (5a) becomes sampling $F_{a_i|a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_N, f_1, \dots, f_N}$ iteratively. A detailed procedure is given in Algorithm 1. Note that, in this analysis, we use the forecasted wind power $x_{i,t}^f$ to initialize $x_{i,t}^a$.

2. Spatial correlation

To obtain the univariate distributions required in Algorithm 1, we start from the conditioned probability distributions from Eq. (4a). By applying Bayes' theorem, the CDF of $X_{i,t}^a$ conditioned on other variables in Algorithm 1 is

Algorithm 1. Gibbs sampling.

procedure GIBBS SAMPLER

Initialize $x_{i,t}^{a,0} \leftarrow x_{i,t}^f, \forall i = 1, \dots, N$

for scenario $\xi \in \Xi$ **do** Sample:

$$X_{1,t}^a | x_{2,t}^{a,\xi-1}, \dots, x_{N,t}^{a,\xi-1}, x_{1,t}^f, \dots, x_{N,t}^f$$

$$X_{2,t}^a | x_{1,t}^f, x_{3,t}^{a,\xi-1}, \dots, x_{N,t}^{a,\xi-1}, x_{1,t}^f, \dots, x_{N,t}^f$$

...

$$X_{N,t}^a | x_{1,t}^f, \dots, x_{N-1,t}^f, x_{1,t}^f, \dots, x_{N,t}^f$$

end for

end procedure

$$\begin{aligned} &F_{a_i|a_1 \dots a_{i-1} a_{i+1} \dots a_N f_1 \dots f_N}(x_{i,t}^a) \\ &= \frac{\text{Prob}(X_{i,t}^a \leq x_{i,t}^a, X_{j \neq i,t}^a = x_{j \neq i,t}^f, X_{1,t}^f = x_{1,t}^f, \dots, X_{N,t}^f = x_{N,t}^f)}{\text{Prob}(X_{j \neq i,t}^a = x_{j \neq i,t}^f, X_{1,t}^f = x_{1,t}^f, \dots, X_{N,t}^f = x_{N,t}^f)}, \\ &j \in \{1, \dots, N\}. \end{aligned} \tag{6}$$

By applying the copula function, the equation becomes

$$\begin{aligned} &F_{a_i|a_1 \dots a_{i-1} a_{i+1} \dots a_N f_1 \dots f_N}(x_{i,t}^a) \\ &= \frac{\partial^{2N-1} C_{2N}}{\partial F_{a_1} \dots \partial F_{a_{i-1}} \partial F_{a_{i+1}} \dots \partial F_{a_N} \partial F_{f_1} \dots \partial F_{f_N}} \Big|_0^{F_{a_i}(x_{i,t}^a)} \\ &= \frac{\partial^{2N-1} C_{2N}}{\partial F_{a_1} \dots \partial F_{a_{i-1}} \partial F_{a_{i+1}} \dots \partial F_{a_N} \partial F_{f_1} \dots \partial F_{f_N}} \Big|_0^1, \end{aligned} \tag{7}$$

which only includes a copula function C_{2N} and marginal distribution functions associated with each variable F_{a_i} . Note that detailed derivation can be found in Appendix.

Once the univariate CDF is determined, Algorithm 1 is implemented sequentially to produce samples. In this study, the logit-normal distribution is used to fit the marginal distributions of wind power,¹¹ and the copula is selected from among Gaussian, student t , and the Archimedean family based on the Bayesian information criterion (BIC).¹³

3. Temporal correlation

Samples produced from the above method represent simultaneous power production from multiple spatially correlated wind farms. However, as suggested by Pinson *et al.*,¹² power production from a wind farm at different times are also temporally correlated, which can be modeled as follows: Given wind farm i and let $u_{i,t}^a = F_{a_i}(x_{i,t}^a) \in (0, 1)$ denote the CDFs associated with its actual power production at $t = 1$ to T , suppose $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,T})$ is a T -dim variable whose entries are given by $Y_{i,t} = \Phi^{-1}(u_{i,t}^a)$, then, \mathbf{Y}_i can be modeled using a multivariate Gaussian distribution

$$\mathbf{Y}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_i),$$

where Φ^{-1} is the inverse of the standard Gaussian CDF function, $\mathbf{0}$ is a T -dim vector of zeros, and Σ_i is a covariance matrix whose entries are given by

$$\Sigma_i(t_1, t_2) = \text{cov}(Y_{i,t_1}, Y_{i,t_2}) = \exp\left(-\frac{|t_2 - t_1|}{\nu}\right). \tag{8}$$

Therefore, Eq. (8) implies that the temporal correlation decays exponentially with time and is dictated by the parameter ν . In this study, the value of ν is selected by fitting Σ_i to historical data using the method of least squares.

C. Scenario reduction

After $|\Xi|$ scenarios are generated for each of the K clusters, the generated scenarios can be directly used as inputs to the stochastic programming models. However, the exhaustive combination of all scenarios from K clusters results in $|\Xi|^K$ scenario, which can be computationally prohibitive as K increases. Therefore, the number of generated scenarios must be reduced to an adequate number before the model

can be solved in a time efficient manner. Numerous scenario reduction techniques exist in the literature,^{46–49} where similar scenarios are aggregated based on particular distance measures. However, most of these reduction techniques do not involve clusters. In addition, the process of scenario generation itself can also be time-consuming, particularly given the great number of scenarios if the scenarios in all clusters are exhaustively combined. Therefore, we adopt the method of forward reduction from Dupacova *et al.*⁴⁹ and apply it in a heuristic way to reduce the number of scenarios from $|\Xi|^K$ to 10.

For convenience, we define a T -dim vector $\mathbf{x}_k^{a,\xi}$ to represents scenario ξ in the k th cluster,

$$\mathbf{x}_k^{a,\xi} = \left(\sum_{i=1}^{N_k} x_{i,1}^{a,\xi}, \dots, \sum_{i=1}^{N_k} x_{i,T}^{a,\xi} \right), \quad \forall \xi \in \Xi, \quad (9)$$

where N_k is the number of wind sites in the k th cluster. Therefore, each scenario in a cluster is represented by the sum of wind power from all wind sites in that cluster. Let's denote the original set of sums as $\mathbf{S}_k = \{\mathbf{x}_k^{a,\xi}, k = 1, \dots, K, \xi \in \Xi\}$ and use \mathbf{S} as the reduced set. Using the above definition, the detailed implementation is given in Algorithm 2.

Note that during the k th iteration, a new cluster \mathbf{S}_k is added and \mathbf{S} is updated by taking the Cartesian product of \mathbf{S} and \mathbf{S}_k , i.e., the element in \mathbf{S} is updated by taking all ordered pairs $(\mathbf{x}^{a,j}, \mathbf{x}_k^{a,\xi})$, where $\mathbf{x}^{a,j} \in \mathbf{S}, j = 1, \dots, 10$ and $\mathbf{x}_k^{a,\xi} \in \mathbf{S}_k, \xi \in \Xi$. Therefore, the dimension of set \mathbf{S} increases by T after each iteration and after all iterations, $\dim \mathbf{S} = K \cdot T$. Upon completion of Algorithm 2, each single element in set \mathbf{S} includes the sums of wind power output from all clusters. Based on the indices of the selected scenarios in \mathbf{S} , site-specific wind power output can then be picked up and used in the power market simulation.

D. Performance evaluation

To evaluate the performance of the generated scenarios, two prediction interval (PI) based metrics are introduced: reliability and sharpness. These metrics are used widely in probabilistic forecast to assess the quality of the PIs.⁵⁰ Probabilistic forecasts can take a variety of forms,^{51,52} such as PDFs, quantiles, and scenarios. While the evaluation metrics are often defined based on PDFs, other forms can be converted into PDFs as needed. Since the generated scenarios can be viewed as realizations of probabilistic wind power forecasts, the PIs are estimated based on the generated scenarios and the PI-based metrics are then adopted to evaluate the quality. In addition, the continuous

Algorithm 2. Heuristic scenario reduction.

procedure SCENARIO REDUCTION

Initialize $\mathbf{S} \leftarrow$ empty set \emptyset

for $k = 1, \dots, K$ **do**

Add scenarios from cluster $k: \mathbf{S} \leftarrow \mathbf{S} \times \mathbf{S}_k$

Reduce \mathbf{S} into 10 scenarios using forward reduction

end for

Return the final scenarios $\mathbf{S} = \{\mathbf{x}^{a,j}, j = 1, \dots, 10\}$

end procedure

ranked probability score (CRPS) is also employed to measure the quality of our scenarios.

1. Prediction interval

Given the forecasted mean x and variance σ^2 of a random variable X probabilistically forecasted random variable x and its variance σ^2 , a PI with $100(1 - \alpha)\%$ confidence level can be given by $[L^\alpha, U^\alpha]$,

$$L^\alpha = x - z_{1-\alpha/2}\sigma, \quad (10)$$

$$U^\alpha = x + z_{1-\alpha/2}\sigma, \quad (11)$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard Gaussian distribution. Note that the random variable X here can be the wind power production at site i in time t , i.e., $X_{i,t}^a$, while x and σ^2 in Eqs. (10) and (11) are the estimated first and second moments from the generated scenarios. Since the PI of normalized wind power is bounded in $[0, 1]$, anything out of the interval will be floored at 0 and capped at 1, i.e., the resulting distribution is assumed to be censored Gaussian. Note that a previous study⁵⁰ suggests that even if the actual distribution is non-Gaussian, the assumption of a censored Gaussian can still be applied with satisfactory performance.

2. Reliability

Given confidence level $100(1 - \alpha)\%$, PIs can be derived from the generated scenarios for each random variable. The future measured wind power is expected to lie within the PI. The given confidence level is also known as the PI nominal confidence (PINC), and it is expected that the probability of the obtained PIs covering the measured wind power $x_{i,t}^a$ will asymptotically reach the nominal level of confidence. Therefore, the reliability of the PIs is measured by PI coverage probability (PICP),

$$PICP_{i,t}^\alpha = \frac{1}{N \cdot T} \sum_{i=1}^N \sum_{t=1}^T I_{i,t}^\alpha \times 100\%, \quad (12)$$

where the indicator of PICP, $I_{i,t}^\alpha$, is defined by

$$I_{i,t}^\alpha = \begin{cases} 1, & x_{i,t}^a \in [L_x^\alpha, U_x^\alpha] \\ 0, & x_{i,t}^a \notin [L_x^\alpha, U_x^\alpha] \end{cases} \quad (13)$$

Ideally, the PICP should equal the PINC, i.e., the average coverage error (ACE)⁵⁰ should be as close to zero as possible. In our analysis, we use the average absolute coverage error over the entire forecast time horizon T and across all N wind sites as the indicator of reliability,

$$\overline{ACE}^\alpha = \frac{1}{N \cdot T} \sum_{i=1}^N \sum_{t=1}^T |PINC - PICP_{i,t}^\alpha|, \quad (14)$$

where $PINC = 100(1 - \alpha)$.

3. Sharpness

The definition of PICP implies that a high PICP can be realized by widening PIs. Therefore, another metric, sharpness, is introduced to measure the size of PIs. Sharpness is evaluated in the form of

interval score. Given the width of a PI: $\omega_{i,t}^z = U_{i,t}^z - L_{i,t}^z$, the interval score $Sc_{i,t}^z$ is

$$Sc_{i,t}^z = \begin{cases} -2\alpha\omega_{i,t}^z - 4(L_{i,t}^z - x_{i,t}^a), & \text{if } x_{i,t}^a < L_{i,t}^z \\ -2\alpha\omega_{i,t}^z, & \text{if } x_{i,t}^a \in [L_{i,t}^z, U_{i,t}^z] \\ -2\alpha\omega_{i,t}^z - 4(x_{i,t}^a - U_{i,t}^z), & \text{if } x_{i,t}^a > U_{i,t}^z, \end{cases}$$

where $Sc_{i,t}^z$ rewards a narrow PI and gives penalty if the PI fails to cover the realized value $x_{i,t}^a$.

Similar to the definition of ACE, we use an average interval score (AIS) to evaluate the sharpness of the generated scenarios over time horizon T and across all N wind sites,

$$\overline{AIS}^z = \frac{1}{N \cdot T} \sum_{i=1}^N \sum_{t=1}^T Sc_{i,t}^z. \quad (15)$$

4. Overall metric

Clearly, there is a trade-off between reliability and sharpness, since a wider PI may result in a higher coverage rate at the cost of a lower interval score, and *vice versa*. We, therefore, introduce a synthetic evaluation metric (SEM) as the linear combination of ACE and AIS to give a comprehensive evaluation of the PI,

$$\overline{SEM}^z = \lambda_1 \overline{ACE}^z - \lambda_2 \overline{AIS}^z. \quad (16)$$

Note that a lower SEM indicates a better overall performance since a negative AIS is used here. λ_1 and λ_2 are weighting coefficients, and we assume $\lambda_1 = \lambda_2 = 0.5$ to put equal emphasis on both reliability and sharpness. The weighting coefficients can be varied by decision makers to put more weights on either aspect.

5. Score

There are numerous numerical scores proposed to measure the quality of probabilistic forecasts.⁵³ We use the CRPS score, a proper score that is designed to measure both the reliability and sharpness of a probabilistic forecast.⁵⁴ Suppose the predicted CDF of the wind power production in site i at time t is $\hat{F}_{i,t}(x)$ and the observed wind power is $x_{i,t}^a$, the definition of CRPS can be given by the following equation by measuring the difference between the predicted and observed CDFs,

$$CRPS_{i,t} = \int [\hat{F}_{i,t}(x) - \mathbf{1}_{x \geq x_{i,t}^a}]^2 dx, \quad (17)$$

where $\mathbf{1}_{x \geq x_{i,t}^a} : \{x|x \in \mathbb{R}\} \rightarrow \{0, 1\}$ is the indicator function of event $x \geq x_{i,t}^a$. We further take the average over all wind sites and all time intervals to measure the overall performance,

$$CRPS = \frac{1}{N \cdot T} \sum_{i=1}^N \sum_{t=1}^T CRPS_{i,t}. \quad (18)$$

Note that the predicted CDF is required before calculating the CRPS. In the realm of weather prediction, the predicted CDF can be obtained by three methods: parametric method, non-parametric method, and ensemble forecast.⁵¹ By giving a set of discrete scenarios, our study falls in the category of ensemble forecast. Therefore, to estimate the

predicted CDF, the empirical CDF is derived from the generated scenarios by using the following equation:

$$\hat{F}_{i,t}(x) = \mathbb{E}_{\xi \in \Xi} (\mathbf{1}_{x \geq x_{i,t}^{a,\xi}}), \quad (19)$$

where $x_{i,t}^{a,\xi}$ denotes the generated wind power scenarios for wind site i at time t .

III. POWER MARKET SIMULATION

A critical motivation of this study is to apply the enhanced approach of scenario generation to real-world power systems with high renewable penetration. Power markets in real world are typically operated on multiple timescales ranging from days ahead to real time. Typically, as the markets evolve forward, forecasts with shorter lead times are used as they become available. In most real-world power markets, the day-ahead market is often cleared one day before using day-ahead forecasts and the schedules are updated in the real-time market several hours ahead of the actual operation by using more up-to-date forecasts. The imbalance between the forecasts and the actual values is then offset by deploying operating reserves, which is known as the load frequency control (LFC) process. The purpose of the power market simulation is to show the market benefits from the improvement in forecasts due to clustering. However, we believe that including forecasts with multiple lead times would be unnecessarily complex to illustrate the benefits of the methods utilized in this work. Therefore, although forecasts with different lead times are used in real-world markets, we only focus on day-ahead forecasts and assume that the real-time market is cleared based on actual values to neglect the LFC process. In addition, we assume that both markets are operated on a 24-h horizon at an hourly resolution. To simulate the two markets, we develop a day-ahead unit commitment (DAUC) model and a real-time economic dispatch (RTED) model. The DAUC model takes in day-ahead wind power forecasts and determines the commitment statuses of all thermal units. Next, the RTED model is solved based on actual wind power production to determine the final financial settlement. In addition, the DAUC model is solved both deterministically and stochastically. The stochastic run solves for a hedging strategy that explicitly considers future uncertainties associated with wind power, which is represented by scenarios. By contrast, the deterministic run solves each wind scenario independently and all decision variables vary by scenarios. Once the DAUC model is solved, commitment statuses of all thermal units are fixed and the RTED model is solved by using the actual wind output. We then compute the financial settlement only based on the solutions from the RTED model.

A. The deterministic DAUC model

The DAUC model is formulated as a mixed integer program by following Ref. 55 and enhanced by constraints associated with regulating and responsive reserves. In addition, active power flow on transmission lines is calculated using shift factors and bounded by a set of pre-defined line limits. The objective of the model is to minimize total costs, which include actual production costs, fixed costs, and load and curtailment costs. The full DAUC model formulation is given below. Note that decision variables in each scenario (ξ) are independent of other scenarios in the deterministic runs.

System load balance

$$\sum_{g \in \mathbf{G}} p_{g,t}^{\xi} + \sum_{b \in \mathbf{B}^L} \delta_{b,t}^{\xi} = \sum_{b \in \mathbf{B}^L} D_{b,t}, \forall t \in \mathbf{T}. \quad (20)$$

Maximum and minimum power generation

$$P_g^{\min} \cdot v_{g,t} \leq p_{g,t}^{\xi} \leq \bar{p}_{g,t}^{\xi} \leq P_g^{\max} \cdot v_{g,t}, \forall g \in \mathbf{G}^T, \quad (21a)$$

$$p_{g,t}^{\xi} \leq \bar{p}_{g,t}^{\xi}, \forall g \in \mathbf{G}, \quad (21b)$$

$$\bar{p}_{g,t}^{\xi} \leq P_{g,t}^{\xi}, \forall g \in \mathbf{G} \setminus \mathbf{G}^T. \quad (21c)$$

Note that if wind unit $g \in \mathbf{G}^W$ is installed at site i , the maximum available wind power $P_{g,t}^{\xi}$ is given by multiplying the normalized wind power $x_{i,t}^{a,\xi}$ with the rated capacity of unit g .

Maximum available power of thermal generators accounting for ramp rates

$$\begin{aligned} \bar{p}_{g,t}^{\xi} - p_{g,t-1}^{\xi} &\leq R_g^U \cdot v_{g,t-1}^{\xi} + R_g^{SU} (v_{g,t}^{\xi} - v_{g,t-1}^{\xi}) + P_g^{\max} (1 - v_{g,t}^{\xi}), \\ \forall g \in \mathbf{G}^T, t \in \mathbf{T}, \end{aligned} \quad (22a)$$

$$\bar{p}_{g,t}^{\xi} \leq P_g^{\max} \cdot v_{g,t+1}^{\xi} + R_g^{SD} (v_{g,t}^{\xi} - v_{g,t+1}^{\xi}), \forall g \in \mathbf{G}^T, t \in \mathbf{T}. \quad (22b)$$

Power generation during ramp-down or shut-down

$$\begin{aligned} p_{g,t-1}^{\xi} - p_{g,t}^{\xi} &\leq R_g^D \cdot v_{g,t}^{\xi} + R_g^{SD} (v_{g,t-1}^{\xi} - v_{g,t}^{\xi}) + P_g^{\max} (1 - v_{g,t-1}^{\xi}), \\ \forall g \in \mathbf{G}^T, t \in \mathbf{T}. \end{aligned} \quad (23)$$

Active power flow, $\forall \ell \in \mathbf{L}$,

$$-\Lambda_{\ell} \leq \sum_{b \in \mathbf{B}} SF_{\ell,b} \left(\sum_{g \in \mathbf{G}_b} p_{g,t}^{\xi} - \delta_{b,t}^{\xi} - D_{b,t} \right) \leq \Lambda_{\ell}. \quad (24)$$

Up-time of thermal generators, $\forall g \in \mathbf{G}^T$,

$$\sum_{t=1}^{T_g^{U0}} (1 - v_{g,t}^{\xi}) = 0, \quad (25a)$$

$$\begin{aligned} \sum_{\tau=t}^{t+T_g^U} v_{g,\tau}^{\xi} &\geq T_g^U (v_{g,t}^{\xi} - v_{g,t-1}^{\xi}), \\ \forall t \in [T_g^{U0} + 1, T - T_g^U + 1] \end{aligned} \quad (25b)$$

$$\sum_{\tau=t}^T [v_{g,\tau}^{\xi} - (v_{g,t}^{\xi} - v_{g,t-1}^{\xi})] \geq 0, \quad \forall t \in [T - T_g^U + 2, T]. \quad (25c)$$

Down-time of thermal generators, $\forall g \in \mathbf{G}^T$,

$$\sum_{t=1}^{T_g^{D0}} v_{g,t}^{\xi} = 0, \quad (26a)$$

$$\begin{aligned} \sum_{\tau=t}^{t+T_g^D} (1 - v_{g,\tau}^{\xi}) &\geq T_g^D (v_{g,t-1}^{\xi} - v_{g,t}^{\xi}), \\ \forall t \in [T_g^{D0} + 1, T - T_g^D + 1], \end{aligned} \quad (26b)$$

$$\sum_{\tau=t}^T [(1 - v_{g,\tau}^{\xi}) - (v_{g,t-1}^{\xi} - v_{g,t}^{\xi})] \geq 0, \forall t \in [T - T_g^D + 2, T]. \quad (26c)$$

Reserve availability, $\forall g \in \mathbf{G}^T$,

$$r_{g,t}^{S,\xi} + r_{g,t}^{RU,\xi} \leq \bar{p}_{g,t}^{\xi} - p_{g,t}^{\xi}, \quad (27a)$$

$$r_{g,t}^{RD,\xi} \leq p_{g,t}^{\xi} - P_g^{\min} \cdot v_{g,t}^{\xi}. \quad (27b)$$

Spinning reserve balance, $\forall t \in \mathbf{T}$,

$$\sum_{g \in \mathbf{G}^T} r_{g,t}^{S,\xi} + \delta_t^{S,\xi} = RQ_t^S. \quad (28)$$

Regulating up/down reserve balance, $\forall t \in \mathbf{T}$,

$$\sum_{g \in \mathbf{G}^T} r_{g,t}^{RU,\xi} + \delta_t^{RU,\xi} = RQ_t^{RU}, \quad (29a)$$

$$\sum_{g \in \mathbf{G}^T} r_{g,t}^{RD,\xi} + \delta_t^{RD,\xi} = RQ_t^{RD}. \quad (29b)$$

Cost of thermal generators, $\forall g \in \mathbf{G}^T, t \in \mathbf{T}$,

$$z_{g,t}^{P,\xi} = C_g^0 \cdot v_{g,t}^{\xi} + C_g \cdot p_{g,t}^{\xi}, \quad (30a)$$

$$z_{g,t}^{SU,\xi} \geq C_g^{SU} (v_{g,t}^{\xi} - v_{g,t-1}^{\xi}), \quad (30b)$$

$$z_{g,t}^{SD,\xi} \geq C_g^{SD} (v_{g,t-1}^{\xi} - v_{g,t}^{\xi}). \quad (30c)$$

Curtailment cost, $\forall t \in \mathbf{T}$,

$$z_t^{\Lambda,\xi} = C^{\Delta L} \sum_{b \in \mathbf{B}^L} \delta_{b,t}^{\xi} + C^{\Delta R} (\delta_t^{RU,\xi} + \delta_t^{RD,\xi}) + C^{\Delta S} \cdot \delta_t^{S,\xi}. \quad (31)$$

Constraint (20) ensures system-wide load balance. Different from Ref. 55, we include slack variable $\delta_{b,t}^{\xi}$ for each load bus $b \in \mathbf{B}^L$ to reduce the chances of infeasibilities. Equations (21a) and (21b) allow thermal units to operate within their output limits, while the output of renewable units is limited by power forecast in Eq. (21c). In addition, Eqs. (22a) and (22b) limit the maximum output from thermal generators during start-up and shut-down. The output of thermal units is further limited by Eq. (23) during ramp-down or shut-down. In Eq. (24), the active power flow on transmission line ℓ is calculated using shift factors $SF_{\ell,b}$ and is limited by line limit Λ_{ℓ} . The minimum online/offline time of thermal generators are limited by Eqs. (25) and (26). The amount of available reserves ($r_{g,t}^{S,\xi}$ for responsive reserve, $r_{g,t}^{RU,\xi}$ and $r_{g,t}^{RD,\xi}$ for up/down regulating reserves) from thermal units is provided in Eq. (27) and requirements of reserves are met in Eqs. (28) and (29). Similar to the load balance constraint, slack variables are also introduced to account for reserve shortages.

Costs of thermal generators consist of production costs and fixed costs. Typically, production costs of thermal generators are represented by quadratic functions; here for simplicity, they are modeled as linear functions of power production $p_{g,t}^{\xi}$, as shown in Eq. (30a). The fixed costs, as given by Eqs. (30b) and (30c), include start-up and shut-down costs, which are determined solely by commitment statuses. In addition to the costs of thermal units, the objective function also includes curtailment costs, which are used to avoid load shedding and reserve shortages, as given in Eq. (31). Finally, the objective function in the deterministic run is given in Eq. (32) by summing up all costs:

$$\min : z^{\xi} = z_t^{\Lambda,\xi} + \sum_{g \in \mathbf{G}^T, t \in \mathbf{T}} \left(z_{g,t}^{P,\xi} + z_{g,t}^{SU,\xi} + z_{g,t}^{SD,\xi} \right). \quad (32)$$

B. The stochastic DAUC model

To account for uncertainties associated with wind power, a two-stage stochastic DAUC model is developed based on the deterministic DAUC model. We assume that the commitment statuses ($v_{g,t}^{\xi}$) are the only first-stage decision variables and must remain equal across all scenarios, while all other variables are the second-stage decision variables and are thus scenario-dependent. Therefore, a non-anticipativity constraint is introduced in the following equation:

$$v_{g,t}^{\xi} = v_{g,t}^{\xi'}, \forall g \in \mathbf{G}^T, \quad t \in \mathbf{T}, \xi \neq \xi'. \quad (33)$$

The objective function of the stochastic run is given by Eq. (34), where the probability-weighted costs are minimized. The fixed costs are equal across all scenarios, since they are only dependent on the first stage variables. By contrast, the production costs and curtailment costs are functions of the second stage variables, which are scenario dependent; therefore, Eq. (34) minimizes their probability-weighted expectation. Note that probabilities of scenarios (Pr_{ξ}) are determined in the scenario reduction process

$$\min \hat{z} = \sum_{g \in \mathbf{G}^T, t \in \mathbf{T}} (z_{g,t}^{SU} + z_{g,t}^{SD}) + \sum_{\xi} Pr_{\xi} \cdot \left(z_t^{A,\xi} + \sum_{g \in \mathbf{G}^T, t \in \mathbf{T}} z_{g,t}^{P,\xi} \right). \quad (34)$$

C. The RTED model

After the DAUC model is solved, the RTED model is solved by fixing the commitment status of all thermal units and replacing the forecasted wind power with actual values. In addition, the minimum online/offline time constraints in Eqs. (25) and (26) are also removed. Therefore, the RTED model is formulated as a continuous linear program, due to the removal of all binary variables. The full model consists of Eqs. (20)–(24) and (27)–(31), and the objective function is represented by Eq. (32). Note that while each deterministic scenario results in an independent set of commitment status, hence an independent RTED run, the stochastic run returns only one set of commitment status for all scenarios and is followed by one single RTED run.

IV. DATA AND ASSUMPTIONS

To simulate real-world power systems, we use a well-developed, publicly available synthetic network developed by Ref. 39, which is built on the footprint of the Electric Reliability Council of Texas (ERCOT). The 24-h load profiles from the IEEE 118 bus system are mapped into load profiles of the Texas system. Note that the load profile of one load bus in the 118 bus system may be mapped into multiple load buses in the Texas system, due to the latter's significantly larger size. In addition, the load profiles are scaled up or down to match the original load magnitudes of the Texas system.

Generating units in the Texas system can be grouped into six categories based on their fuel types, as shown in in Table I. Most techno-economic parameters of thermal units, such as the maximum and minimum output and ramp rates, are directly drawn from the original dataset. The production cost functions are mapped from the units with similar fuel types and rated capacities in the IEEE 118 bus system. The minimum online and offline time of a thermal units is determined primarily by its prime driver, while the start-up and shut-down costs are also affected by fuel costs. Since the original dataset does not give

TABLE I. Number and capacity of units by fuel types in the real-world Texas power system⁵¹ and the synthetic power system from Ref. 39. Note that the data from EIA reflect the whole Texas power system, while ERCOT covers over 90% of Texas load, hence the discrepancy.

Fuel type	Synthetic ERCOT system			Real-world Texas system	
	No. of units	Capacity (MW)	Share	Capacity (MW)	Share
Coal	39	14 502	15%	23 589	19%
NG	367	63 810	66%	69 386	56%
Nuclear	4	5139	5%	4960	4%
Wind	87	9587	10%	22 583	18%
Solar PV	22	651	1%	1240	1%
Hydro	25	2603	3%	670	1%
Biomass	0	0	0%	425	0%
Other	0	0	0%	659	1%
Total	546	96 292	100%	123 512	100%

the types of prime drivers, the prime drivers are determined by their capacities following Ref. 56. Fuel costs are drawn from US EIA's Monthly Energy Review.⁵⁷ Note that we assume that all nuclear units are constantly online. In addition, wind and solar units have zero production and start-up/shut-down costs.

For simplicity, we assume that outputs from all solar and hydro units are zero. Neglecting solar and hydro power may affect the commitment statuses of thermal units; however, we do not think it introduces a systematic error given their small capacity shares. As can be seen from Table I, both hydro power and solar PV only contribute marginally to the total capacity in both the real-world and the synthetic systems. In addition, it allows us to focus on the uncertainties associated with wind power.

We assume that load curtailment cost is \$9000/MWh, according to the value of load loss (VOLL) used by ERCOT.⁵⁸ We use \$2000/MWh, the low system-wide offer cap (LCAP) provided by the Texas Public Utility Commission,⁵⁹ as the penalty price for responsive reserve shortage and \$5500/MWh, the average of LCAP and VOLL, as the penalty price for the regulating up/down reserve shortages. The penalty prices are selected such that responsive reserves are curtailed ahead of regulating reserves, and load will be the last to be curtailed.⁶⁰ Historical data from ERCOT indicate that the regulating up/down reserves account for 1%–2% of system-wide load, while the responsive reserves range from 5% to 11%. Therefore, we assume 10% for responsive reserve margin and 2% for both up and down regulating reserve margins.

The data drawn from NREL's Wind Integration National Dataset (WIND) toolkit⁶² are used as input to the scenario generation model. 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 data from 2007 to 2011 are used as input to the cluster analysis. The WIND toolkit simulates wind power production every 5 min and applies an NWP model to provide hourly power forecasts at multiple lead times. In our study, the day-ahead forecast is selected to generate scenarios for the day-ahead market. Since the temporal granularity is every 5 min for the actual power

data and hourly for all forecasts, we calculate the hourly average of the actual wind power in the training dataset to match the time resolution of the day-ahead forecast. To capture seasonal variations of load profiles in the power market simulation, one day is selected from each season in 2012: February 10, 2012 (winter), May 10, 2012 (spring), August 10, 2012 (summer), and November 10, 2012 (fall). The Gibbs sampler is trained using actual wind power and wind power forecast from January 1, 2007 to the day before each target day. Next, scenarios are generated by the Gibbs sampler based on the day-ahead forecasts for the target day. The 87 wind units in the Texas system are mapped from 45 sites (shown in Fig. 1) in the WIND toolkit based on their geographical coordinates. Therefore, all wind units in one site have identical normalized wind power, although they may have different rated capacities. Figure 2 shows a heat map of the cross-site correlation coefficients across all 45 wind sites in this study, indicating strong correlations.

V. RESULTS: CLUSTERING-BASED SCENARIO GENERATION

In this section, the clustering based framework developed in Sec. II is applied to the 45 wind sites from the Texas system over the four target days. The wind sites are first divided into a given number (K) of clusters and the Gibbs sampler is applied to each cluster independently. Since Gibbs sampling requires a large number of iterations before the samples reach a stationary distribution, 9000 samples are generated for each cluster and the first 8000 samples are discarded (known as the “burn-in” process) and only the last 1000 samples are kept for market simulation. The performance metrics are then derived from the generated scenarios and scrutinized. The goal is to examine the relation between the number of clusters and the quality of generated scenarios. The optimal number of clusters is jointly determined by the performance metrics and time consumption. In this study, the number of clusters K ranges from 1 to 20. Note that since when $K = 1$, all wind sites are put into one single group, it is equivalent to the case where cluster analysis is not applied. Therefore, the case when $K = 1$ represents our baseline case. The implementation is coded in

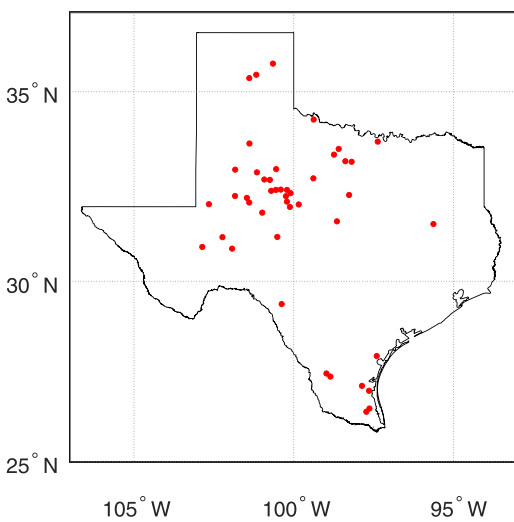


FIG. 1. Locations of all wind sites in the synthetic ERCOT power grid.

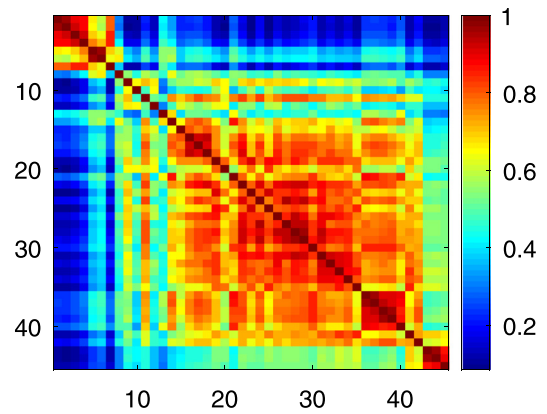


FIG. 2. Correlation coefficients across all 45 wind sites in the study.

MATLAB and executed on a high performance computing platform to exploit 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.

Figures 3 and 4 show the results from an ex-post validation analysis by scrutinizing the autocorrelation coefficients and cross correlation coefficients. The autocorrelation coefficients of historical wind power from site i are given by the following equation:

$$\rho_i(\Delta t) = \frac{\text{cov}(x_{i,t}^a, x_{i,t+\Delta t}^a)}{\sigma(x_{i,t}^a) \cdot \sigma(x_{i,t+\Delta t}^a)}, \quad (35)$$

where Δt is time lag and $x_{i,t}^a$ and $x_{i,t+\Delta t}^a$ are taken from the training dataset. Since the generated scenario $x_{i,t}^{a,\xi}$ can be viewed as an estimator of $x_{i,t}^a$, the estimated autocorrelation coefficient $\hat{\rho}_i(\Delta t)$ can be given by the above equation when $x_{i,t}^a$ is substituted by $x_{i,t}^{a,\xi}$. The autocorrelation is caused by temporal correlations in the wind power time series. By comparing $\rho_i(\Delta t)$ and $\hat{\rho}_i(\Delta t)$ from all 45 sites, Fig. 3 indicates good agreement between the model input and output. The declining trends indicate weaker temporal correlations as time lag increases, as modeled in Eq. (8).

As presented in Fig. 2, the wind sites are strongly correlated. As spatial correlation is modeled by copula functions, the resulting scenarios should present strong correlations across sites. We examine the linear correlation coefficients using the generated scenarios and compare it with the original coefficients from historical data. The correlation coefficient is given by

$$\rho_{ij} = \frac{\text{cov}(x_{i,t}^a, x_{j,t}^a)}{\sigma(x_{i,t}^a)\sigma(x_{j,t}^a)}. \quad (36)$$

Similar to the autocorrelation coefficients, the estimated cross correlation coefficients $\hat{\rho}_{ij}$ are given when $x_{i,t}^a$ are substituted by $x_{i,t}^{a,\xi}$. Due to the large number of sites, we calculate the relative differences between ρ_{ij} and $\hat{\rho}_{ij}$,

$$\frac{|\rho_{ij} - \hat{\rho}_{ij}|}{\rho_{ij}}. \quad (37)$$

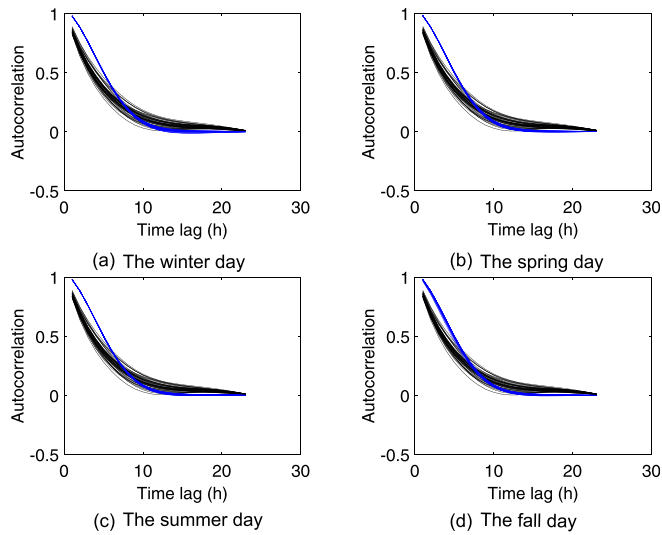


FIG. 3. Autocorrelation coefficients as a function of time lag. Blue curves represent autocorrelation coefficients from the generated scenarios, while black curves represent the coefficients from the training dataset.

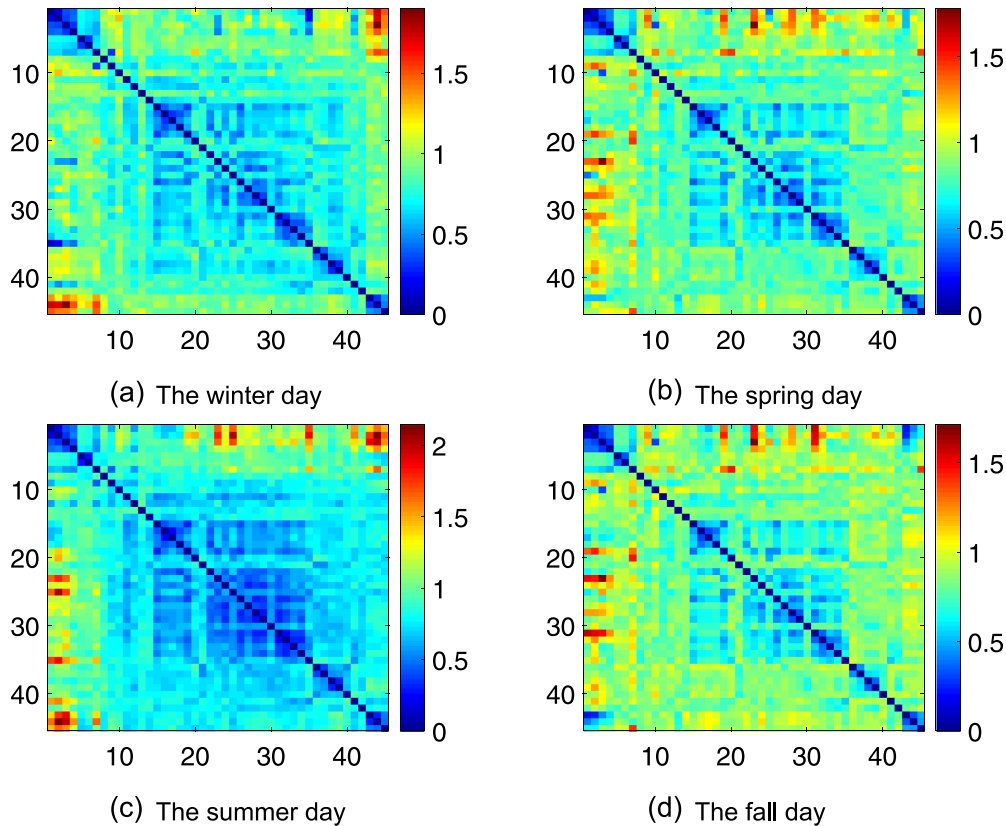


FIG. 4. Relative differences between the modeled cross-variable correlation coefficients and the original correlation coefficients in Fig. 2.

Figure 4 visualizes the relative differences as heat maps, where each pixel represents one pair of sites. By comparing it with Fig. 2, it shows approximation errors using the copula functions. However, it also indicates that while greater errors are presented when the spatial correlations are weak, most errors for pairs with strong spatial correlations are not significant, indicating good agreement between the model input and output.

The results from the winter day (February 10, 2012) are shown in Fig. 5. As illustrated in Fig. 5(a), the diagonal indicates the ideal results where PICP equals PINC; therefore, greater distance between the PICP and the diagonal implies lower reliability. It shows that the reliability is affected by both the number of clusters and PINC. In most cases, the distance between PICP and the diagonal is greater when PINC = 50%. This inspires us to examine the average distance between the PICP and the diagonal over the entire PINC range, i.e., ACE, which measures the overall reliability by taking the mean. Figure 5(c) shows the ACE as a function of K . As can be seen, the ACE drops from around 15% to less than 1% when K increases from 2 to 9, and rebounds to 4% when K increases to 20. The considerable drop of ACE before $K = 9$ implies improved reliability due to clustering, which can be further attributed to the exclusion of sites that are weakly correlated. By contrast, the rise of ACE beyond $K = 9$ indicates less reliable scenarios with further refined clusters, probably due to the exclusion of those sites that present strong correlations.

Similar results are presented by interval scores. Figure 5(b) exhibits monotonically increasing interval scores as PINC increases. In addition, the interval scores are also largely affected by K . Similar to

ACE, AIS gives a comprehensive evaluation to the interval scores of a given K by taking the mean over the entire PINC range. As presented in Fig. 5(d), the AIS reaches its peak at $K=6$ and declines thereafter. Since higher interval scores indicate better performance in terms of sharpness, the results suggest $K=6$ for the optimal sharpness.

As displayed in Fig. 5, both metrics suggest that a medium number of clusters results in the optimal performance of the generated scenarios. However, the two metrics do not agree on the optimal number of clusters: the highest reliability is present when $K=9$ while $K=6$ maximizes sharpness. Therefore, SEM is introduced and used as a comprehensive evaluation metric by factoring in both reliability and sharpness. Figure 6 shows the SEM values and the CRPS scores as a function of K in all four representative days. As can be seen, the winter, spring, and fall days exhibit very similar profiles, where the SEM is minimized when a medium number of clusters is selected. For example, the SEM is minimized when $K=10, 12,$ and 10 in the winter, spring, and fall day, respectively. Although the summer day presents a slightly different profile, where the SEM is minimized at $K=15$ and only rebounds slightly when K increases to 20 , the minimum SEM still indicates the optimal overall performance of the generated scenarios when a medium number of clusters is selected. Interestingly, the CRPS scores present different profiles from the SEM: the scores remain low before $K=10$ and increase when $K > 10$. Therefore, both the SEM and CRPS scores suggest against a high number of clusters, which can

reduce the quality of the generated scenarios due to failure to account for correlations across wind farms. In addition, the CRPS scores also suggest benefits from clustering: the CRPS scores are at their lowest when K ranges from 2 to 4 in all seasons except for winter.

The generated scenarios can be directly used as inputs to the stochastic DAUC models. However, the number of generated scenarios must be reduced to an adequate number before the model can be solved in a time efficient manner. We follow the procedures in Algorithm 2 to reduce the generated scenarios. Figure 7 shows the time consumption as a function of K . The total time consumption is composed of scenario generation time and scenario reduction time. As shown in all four representative days, the total time consumption sees a trade-off between scenario generation and scenario reduction. The time consumed by scenario generation contributes to over 90% of the total time consumption when $K < 3$, while scenario reduction consumes over 65% of the total time when K increases to 20. In addition, as K increases, the time consumption of scenario generation presents a decreasing trend, while scenario reduction presents an increasing trend. For example, the average time consumption of scenario generation across all four days decreases from 140 min to 25 min when K increases from 1 to 20. By contrast, the average time consumption of scenario reduction increases from less than 1 min to more than 50 min over the same range of K . It is not surprising to observe the trade-off, since as K increases, each cluster is more likely to see a smaller number

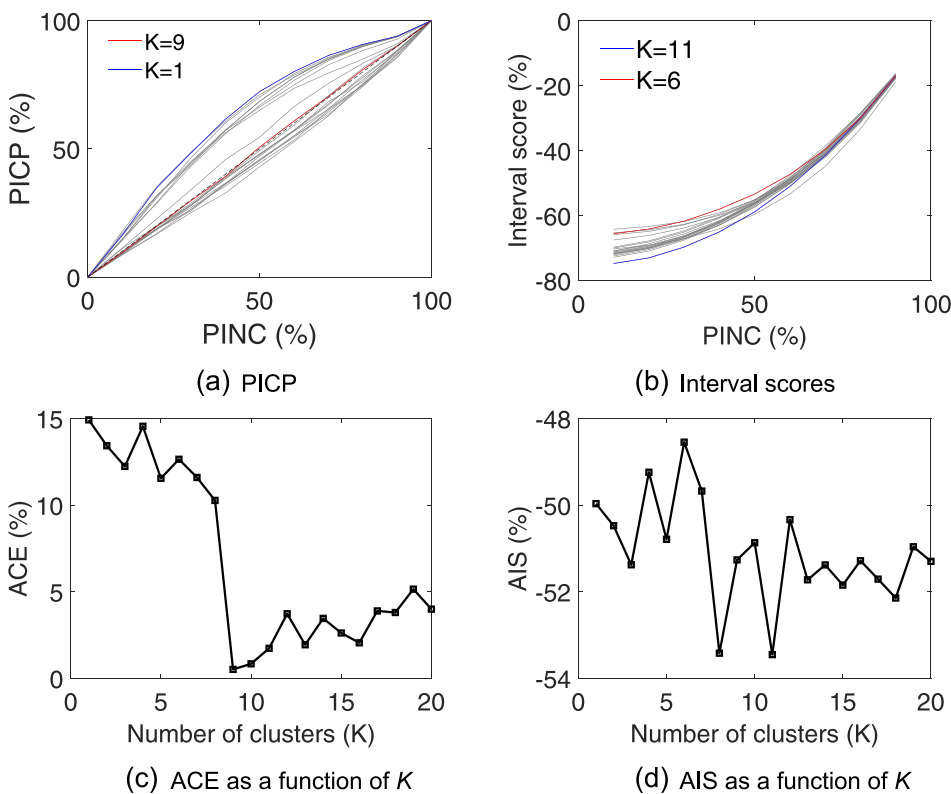


FIG. 5. Performance evaluation of the generated scenarios for the winter day (February 10, 2012): (a) PICP as a function of PINC. The diagonal (dashed line) represents the ideal result, where $PICP = PINC$. The red and blue curves represent the closest and farthest PICP profiles to the diagonal. (b) Interval scores as a function of PINC. The red and blue curves represent the profiles with the highest and lowest mean interval scores, respectively. (c) ACE and (d) AIS as a function of K .

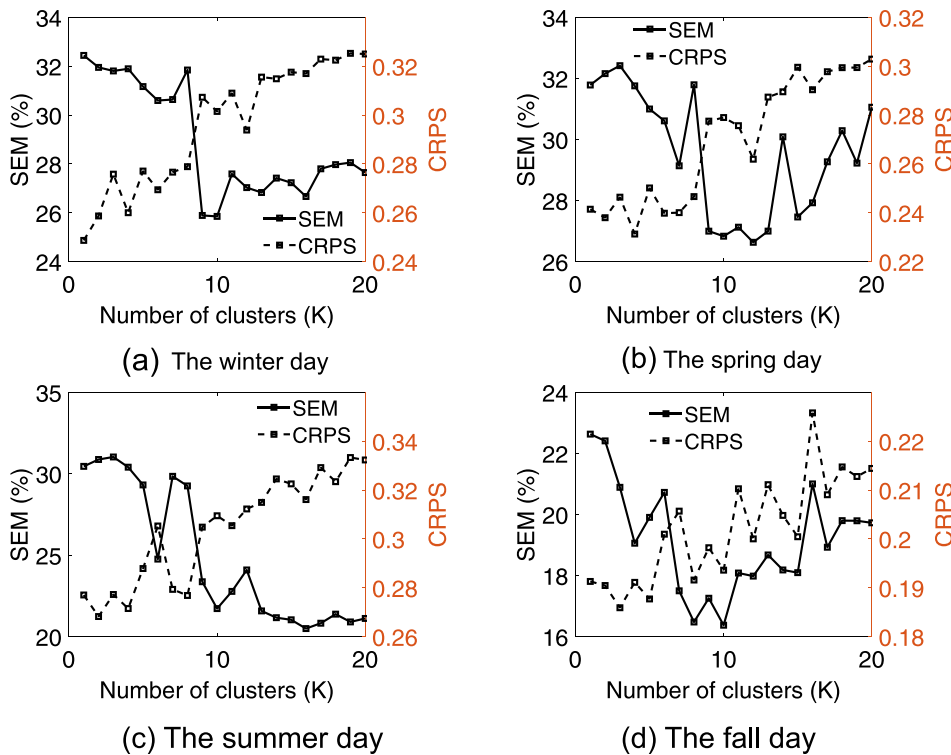


FIG. 6. SEM and CRPS as a function of the number of clusters (K) for all four cases in our study. (a) February 10, 2012, (b) May 10, 2012, (c) August 10, 2012, and (d) November 10, 2012.

of wind farms, which reduces the scenario generation time. However, the number of iterations in scenario reduction is directly related to the number of clusters, and an increased K leads to longer scenario reduction time.

The trade-off of time consumption between scenario generation and reduction again suggests that a medium number of clusters should be selected. In addition, a medium number of clusters can also significantly reduce the total time consumption. As shown in Fig. 7, the total time consumption tends to be minimized when $5 \leq K \leq 10$.

In addition, Table II compares the baseline ($K=1$) with the clustered cases ($K>1$). The comparison indicates that the clustering-based framework can significantly improve the generated scenarios in terms of SEM and time consumption. As can be seen, the SEM in the clustered cases ($K=2-20$) can be reduced by up to 33%, while the total time consumption can be reduced by over 60%. In addition, the CRPS score can also benefit marginally from clustering. Therefore, the clustering-based framework can significantly enhance the quality of the generated scenarios and reduce the total time consumption.

VI. RESULTS: POWER MARKET SIMULATION

As mentioned in Sec. III, the DAUC model will be solved deterministically and stochastically. Each deterministic run solves only one wind scenario, while the stochastic run solves all scenarios simultaneously. The 1000 wind scenarios generated in Sec. V with the minimum SEM values are reduced to ten scenarios and are represented with “1” to “10.” Note that the stochastic run with scenarios generated from the clustered case is denoted by “SC” and we use “SU” to denote

the stochastic run with scenarios generated from the baseline, i.e., the unclustered method. In the clustered cases, the numbers of clusters are selected based on the minimum SEM. In addition, another run (denoted by “A”) takes the actual wind power as inputs to the DAUC model as if we have perfect foresight into the future. In summary, the DAUC model is solved 13 times, including 11 deterministic solves and two stochastic solves. Next, the RTED model is solved by using the actual wind power and committing all thermal units, where the commitment statuses are given by the DAUC model. In this study, Pyomo Stochastic Programming (PySP),⁶³ an open-source Python library, is employed to implement the stochastic programming models and Gurobi is used to solve them. On the same compute node as used in Sec. V, it takes 20–25 min to solve a deterministic run, around 8 h to solve a stochastic run, and less than 5 min to solve a RTED model.

Results of the power market simulations are presented in Fig. 8 and Table III. As shown in Fig. 8, scenario A is always the least expensive in all seasons due to perfect foresight of wind power. Meanwhile, the more expensive total costs in other scenarios can be largely attributed to uncertainties associated with the forecast. In addition, great variations of total costs are presented in the deterministic scenarios. For example, in the fall day, the total cost from scenario 2 is 66% more expensive than that of scenario A, while scenario 3 is only 7% more expensive. Overall, the increased costs in the deterministic runs range from 2% to 103% across all seasons and scenarios, implying great uncertainties when the day-ahead market is cleared based on the deterministic wind power forecast. In contrast, the total costs of scenario SC only increase by 4%, 2%, and 4% in the spring, summer, and fall days, respectively. Note that although scenario SC results in a 50%

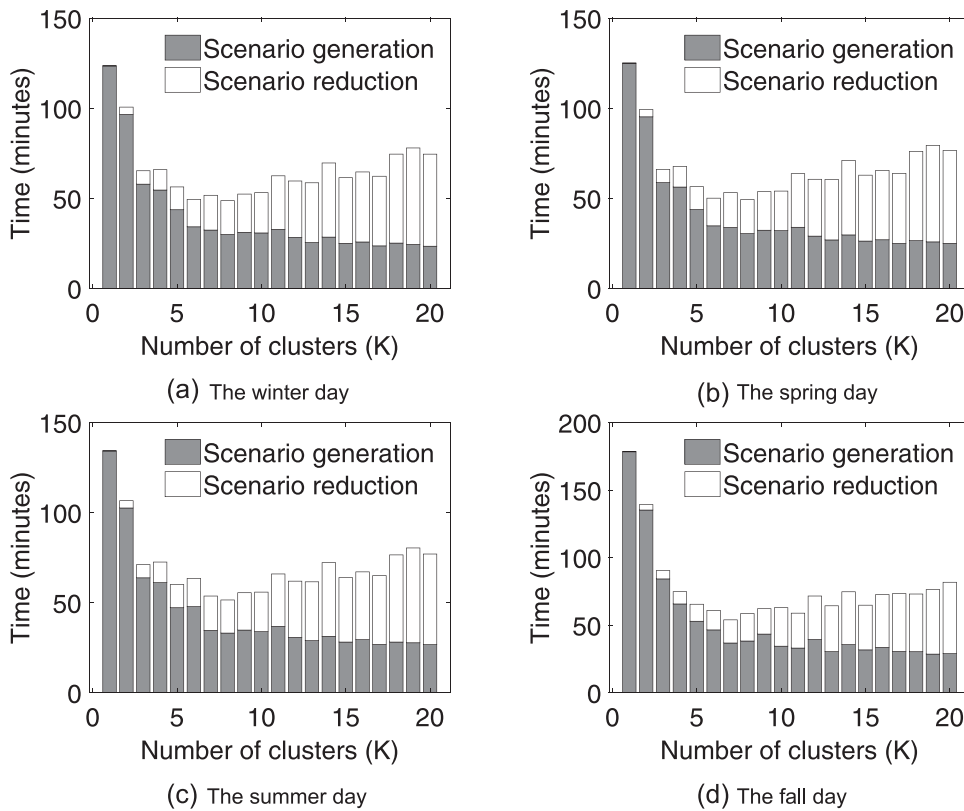


FIG. 7. Time consumption as a function of the number of clusters (K) for all four cases in our study.

more expensive total cost in the winter day, it is still lower than most of the deterministic runs in the same day, which range from 49% to 103%. The comparison between scenarios SC and SU indicates that the total costs are increased when the clustering-based framework is applied. However, the increases are only marginal in most cases: the total costs increase by 1%–3% in seasons 2, 3, and 4.

TABLE II. Comparison between the baseline ($K=1$) and the clustered cases ($K > 1$). Note the minimum SEM, CRPS, and time consumptions in the clustered cases are listed. The optimal K only applies to the clustered cases, i.e., $K > 1$.

		Winter	Spring	Summer	Fall
SEM	Optimal K	10	12	16	10
	Baseline	32%	32%	30%	23%
	Clustered	26%	27%	20%	16%
	Improvement	20%	16%	33%	28%
CRPS	Optimal K	2	4	2	3
	Baseline	0.249	0.242	0.277	0.191
	Clustered	0.259	0.231	0.268	0.186
	Improvement	−4%	5%	3%	3%
Total time (min)	Optimal K	8	6	8	7
	Baseline	123	125	134	178
	Clustered	49	49	51	54
	Improvement	61%	61%	62%	70%

A closer examination of the cost breakdown indicates that the differences in total costs are primarily driven by variations in reserve shortages. As shown in Fig. 8, production costs are almost equal across all scenarios in the same day, while costs of reserve shortages vary significantly. Note that regulating reserve requirements are well met in all scenarios and shortages of responsive reserves are shown in Table III. Similar to the results of total costs, scenario A always results in the least amount of shortages and scenario SC results in less shortages than most deterministic scenarios. Therefore, the stochastic program represents a better strategy for market scheduling than the deterministic approach in terms of reliability and economy.

VII. CONCLUSION AND DISCUSSION

This paper presents an enhanced scenario generation framework by integrating existing approaches with cluster analysis. As demonstrated in the results, the generated scenarios can benefit significantly from the clustering based framework in terms of reliability, sharpness, and time consumption. While only one scenario generation approach is investigated in this study, the framework can also be extended to many other scenario generation techniques. Therefore, this framework can provide proper support to real-world power system analysis with large-scale wind penetration.

Several insights can be drawn from our study. First, as shown in Sec. V, the SEM of the generated scenarios is optimal when a medium number of clusters is used, due to trade-off between intra-cluster and inter-cluster correlations. In addition, the time consumption is also affected by the number of clusters: a small cluster with less members

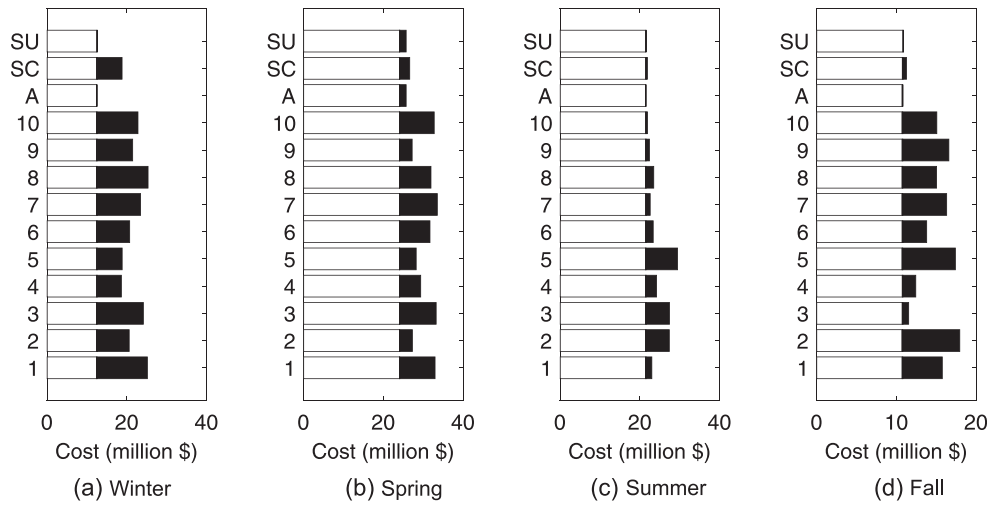


FIG. 8. Cost breakdown of market simulations for all four representative days by scenarios. White represents production costs and black represents reserve shortage costs.

requires less time for scenario generation; however, the increased number of clusters leads to an increased amount of time for scenario generation. Although the CRPS scores suggest a smaller optimal number of clusters, they also indicate that the quality of the generated scenarios clustering can be improved by clustering. The results also imply that it may negatively affect the quality of generated scenarios when all wind farms are grouped into one single cluster regardless of the strength of correlations, which highlights the necessity of the clustering-based framework.

In addition, as indicated in the power market simulation, the hedging strategy given by the stochastic DAUC model presents great advantage over most deterministic runs in terms of both economy and power system reliability. The reserve shortages are effectively reduced without significantly increasing the total costs. Note that even though some deterministic runs may result in lower total costs than the hedging strategy, i.e., scenario 4 in the winter day, it is more likely to result in greater loss by adhering to the deterministic strategy, since all other deterministic runs lead to greater total costs and reserve shortages. By contrast, the stochastic strategy outperforms most deterministic runs in terms of both economy and power system reliability.

Several caveats exist. Figure 8 indicates reserve shortages are present across all scenarios, which can be attributed to multiple factors. First, we assume ancillary services are only provided by thermal units, while other types of qualified resources, such as hydro or even load resources, are excluded. In addition, the economic loss due to reserve shortages may be underestimated as a result of the fixed penalizing

prices, since the dynamic scarcity pricing mechanism in ERCOT allows the prices to rise when a shortage of operating reserves is present.⁵⁸ Finally, ERCOT may procure replacement ancillary service capacity from a supplemental ancillary services market (SASM),⁶⁰ which is not co-optimized with the energy market. Therefore, the reserve shortages do not necessarily reflect resource insufficiency. However, although this study may benefit from including the above factors, the main emphasis is placed on demonstration of the clustering-based scenario generation framework, and modeling of the above factors is beyond our scope.

Besides, we should note that although the stochastic run provides a comparatively better strategy for the power market operation, the time consumption is less acceptable to the operators. In this study, it takes around 8 h to solve its extensive form using state-of-the-art commercial solvers, while it takes 1.5 h to execute day-ahead reliability unit commitment (RUC) in ERCOT.⁶⁰ The slow performance is largely intrinsic to mixed integer programs (MIPs), which is proved to be NP-hard.⁶⁴ Therefore, in the future, additional techniques such as in Refs. 65 and 66 should be employed for a better and more stable time performance.

Another caveat, as shown in Fig. 8, is the more expensive total costs when the clustering-based framework is used to generate scenarios for stochastic runs. However, this does not indicate that the clustering-based framework is outperformed by the baseline in terms of solution quality, since only one number of clusters (K) are examined in each season due to the extremely long computational time of stochastic runs. Ideally, the solution quality should be examined over a

TABLE III. Shortages of responsive reserve in all seasons (MWh).

Season	1	2	3	4	5	6	7	8	9	10	A	SC	SU
Winter	6321	4034	5824	3047	3146	4095	5474	6411	4471	5141	0	3134	0
Spring	4432	1586	4571	2632	2062	3795	4716	3921	1556	4326	803	1253	803
Summer	735	2935	2949	1336	3972	905	523	978	432	191	0	177	0
Fall	2479	3583	365	807	3304	1498	2758	2124	2891	2138	0	223	0

range of cluster numbers. In addition, the optimal K is selected when the SEM is minimized and an underlying hypothesis is that a better SEM of probabilistic scenarios is related to a better solution quality of stochastic optimization. The fact that SEM is weakly correlated with the solution quality implies that further examination of other evaluation metrics and comprehensive studies should be conducted to examine the relationship between the evaluation metrics of probabilistic forecasting and the solution quality of stochastic optimization. Moreover, a real-world decision-maker may need a unique and comprehensive evaluation metric to determine the best set of scenarios, which should account for a variety of factors, such as scenario quality (reliability and sharpness), computational time, and solution quality. In fact, there has been little research into the trade-offs between the quality of stochastic optimization solutions and the quality of probabilistic scenarios and such studies are critical given the discrepancies presented in our study.⁶⁷

Finally, the clustering strategy in Sec. II is based purely on the cross-correlations across wind sites, which only captures the spatial correlation and fails to capture time dependencies. While temporal correlations are accounted for in the scenario generation in the form of autocorrelation, the quality of the generated scenarios could benefit from taking into account more complex dependence structures. Particularly, the spatiotemporal dependencies across wind sites can be complex, and a localized understanding of this structure may provide more targeted information to improve forecasts. For example, previous research suggests that the characterization of spatial correlation across wind sites can be improved by identifying the wind regime that is dominated by a wind direction since sites that are located along the wind direction are more likely to be correlated.⁶⁸ Future studies will explore more clustering methods and criteria, such as distance-based methods, as studies have suggested superior performance when advanced techniques are adopted.^{69,70}

ACKNOWLEDGMENTS

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). This work was authored in part by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Water Power Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

APPENDIX: DERIVATION OF EQUATIONS

Copulas are employed to model the multivariate distributions in the denominator and numerator of Eq. (6). According to Sklar's theorem,⁷¹ any multivariate joint distribution can be written in terms of the marginal distributions of each component and a copula that describes the dependence structure between the components. Therefore, the joint CDF in Eq. (4a) becomes

$$F_{a_1 \dots a_N f_1 \dots f_N}(x_{1,t}^a, \dots, x_{1,N}^a, x_{1,t}^f, \dots, x_{1,N}^f) = C_{2N}(u_{1,t}^a, \dots, u_{N,t}^a, u_{1,t}^f, \dots, u_{N,t}^f), \quad (A1)$$

where the copula $C_{2N} : [0, 1]^{2N} \rightarrow [0, 1]$ is a continuous, real-valued function and $u_{j,t}^a, u_{j,t}^f \in [0, 1]$ are marginal CDFs associated with $X_{j,t}^a$ and $X_{j,t}^f$,

$$u_j^a = F_{a_j}(x_{j,t}^a) = \text{Prob}(X_{j,t}^a \leq x_{j,t}^a), \quad (A2a)$$

$$u_j^f = F_{f_j}(x_{j,t}^f) = \text{Prob}(X_{j,t}^f \leq x_{j,t}^f). \quad (A2b)$$

Similarly, the joint PDF in Eq. (4b) can be expressed as

$$f_{a_1 \dots a_N f_1 \dots f_N} = \frac{\partial^{2N} F_{a_1 \dots a_N f_1 \dots f_N}}{\partial X_{1,t}^a \dots \partial X_{1,N}^a \partial X_{1,t}^f \dots \partial X_{1,N}^f} = c_{2N} \cdot \prod_{j=1}^N (f_{a_j} f_{f_j}), \quad (A3)$$

where $c_{2N} : [0, 1]^{2N} \rightarrow \mathbb{R}_+$ is the $2N$ th order partial derivative of C_{2N} ,

$$c_{2N} = \frac{\partial^{2N} C_{2N}}{\partial F_{a_1} \dots \partial F_{a_N} \partial F_{f_1} \dots \partial F_{f_N}}, \quad (A4)$$

and f_{a_j} and f_{f_j} are marginal PDFs associated with $X_{j,t}^a$ and $X_{j,t}^f$,

$$f_{a_j} = \frac{dF_{a_j}}{dX_{j,t}^a}, \quad (A5a)$$

$$f_{f_j} = \frac{dF_{f_j}}{dX_{j,t}^f}. \quad (A5b)$$

Using the above copula representation, the numerator in (6) can be expressed as follows using the copula function,

$$\begin{aligned} \text{Prob}(X_i^a \leq x_i^a, X_{j \neq i,t}^a = x_{j \neq i,t}^a, X_{1,t}^f = x_{1,t}^f, \dots, X_{N,t}^f = x_{N,t}^f) \\ = \int_0^{x_i^a} f_{a_1 \dots a_N f_1 \dots f_N} dX_i^a = \int_0^{x_i^a} c_{2N} \cdot \prod_{j=1}^N (f_{a_j} f_{f_j}) dX_i^a \\ = \frac{\partial^{2N-1} C_{2N}}{\partial F_{a_1} \dots \partial F_{a_{i-1}} \partial F_{a_{i+1}} \dots \partial F_{a_N} \partial F_{f_1} \dots \partial F_{f_N}} \Big|_0^{F_{a_i}(x_i^a)} \\ \cdot \frac{\prod_{j=1}^N (f_{a_j} f_{f_j})}{f_{a_i}}, \quad j \in \{1, \dots, N\}. \end{aligned} \quad (A6)$$

Similarly, the denominator can be written as the following equation:

$$\begin{aligned} \text{Prob}(X_{j \neq i,t}^a = x_{j \neq i,t}^a, X_{1,t}^f = x_{1,t}^f, \dots, X_{N,t}^f = x_{N,t}^f) \\ = \text{Prob}(X_i^a \leq +\infty, X_{j \neq i,t}^a = x_{j \neq i,t}^a, X_{1,t}^f = x_{1,t}^f, \dots, X_{N,t}^f = x_{N,t}^f) \\ = \frac{\partial^{2N-1} C_{2N}}{\partial F_{a_1} \dots \partial F_{a_{i-1}} \partial F_{a_{i+1}} \dots \partial F_{a_N} \partial F_{f_1} \dots \partial F_{f_N}} \Big|_0^1 \cdot \frac{\prod_{j=1}^N (f_{a_j} f_{f_j})}{f_{a_i}}, \\ j \in \{1, \dots, N\}. \end{aligned} \quad (A7)$$

Finally, the equation in Eq. (6) becomes the form in Eq. (7),

$$F_{a_1|a_1 \cdots a_{i-1} a_{i+1} \cdots a_N f_1 \cdots f_N}(x_{i,t}^a) = \frac{\frac{\partial^{2N-1} C_{2N}}{\partial F_{a_1} \cdots \partial F_{a_{i-1}} \partial F_{a_{i+1}} \cdots \partial F_{a_N} \partial F_{f_1} \cdots \partial F_{f_N}} \Big|_{F_{a_i}(x_{i,t}^a)}}{\frac{\partial^{2N-1} C_{2N}}{\partial F_{a_1} \cdots \partial F_{a_{i-1}} \partial F_{a_{i+1}} \cdots \partial F_{a_N} \partial F_{f_1} \cdots \partial F_{f_N}} \Big|_1}. \quad (\text{A8})$$

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- S. Takriti, J. R. Birge, and E. Long, "A stochastic model for the unit commitment problem," *IEEE Trans. Power Syst.* **11**, 1497–1508 (1996).
- J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming* (Springer Science & Business Media, 2011).
- E. M. Constantinescu, V. M. Zavala, M. Rocklin, S. Lee, and M. Anitescu, "A computational framework for uncertainty quantification and stochastic optimization in unit commitment with wind power generation," *IEEE Trans. Power Syst.* **26**, 431–441 (2011).
- Q. P. Zheng, J. Wang, and A. L. Liu, "Stochastic optimization for unit commitment—A review," *IEEE Trans. Power Syst.* **30**, 1913–1924 (2015).
- J. Wang, A. Botterud, R. Bessa, H. Keko, L. Carvalho, D. Issicaba, J. Sumaili, and V. Miranda, "Wind power forecasting uncertainty and unit commitment," *Appl. Energy* **88**, 4014–4023 (2011).
- J. Wang, M. Shahidehpour, and Z. Li, "Security-constrained unit commitment with volatile wind power generation," *IEEE Trans. Power Syst.* **23**, 1319–1327 (2008).
- F. Bouffard, F. D. Galiana, and A. J. Conejo, "Market-clearing with stochastic security—part I: Formulation," *IEEE Trans. Power Syst.* **20**, 1818–1826 (2005).
- F. Bouffard, F. D. Galiana, and A. J. Conejo, "Market-clearing with stochastic security—part II: Case studies," *IEEE Trans. Power Syst.* **20**, 1827–1835 (2005).
- B. Wang and B. F. Hobbs, "Real-time markets for flexiramp: A stochastic unit commitment-based analysis," *IEEE Trans. Power Syst.* **31**, 846–860 (2016).
- A. Papavasiliou, S. S. Oren, and R. P. O'Neill, "Reserve requirements for wind power integration: A scenario-based stochastic programming framework," *IEEE Trans. Power Syst.* **26**, 2197–2206 (2011).
- P. Pinson, "Very-short-term probabilistic forecasting of wind power with generalized logit-normal distributions," *J. R. Stat. Soc.* **61**, 555–576 (2012).
- 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* **12**, 51–62 (2009).
- M. Cui, J. Zhang, Q. Wang, V. Krishnan, and B.-M. Hodge, "A data-driven methodology for probabilistic wind power ramp forecasting," *IEEE Trans. Smart Grid* **10**, 1326–1338 (2017).
- H. Bludszweit, J. A. Domínguez-Navarro, and A. Llombart, "Statistical analysis of wind power forecast error," *IEEE Trans. Power Syst.* **23**, 983–991 (2008).
- K. Bruninx and E. Delarue, "A statistical description of the error on wind power forecasts for probabilistic reserve sizing," *IEEE Trans. Sustainable Energy* **5**, 995–1002 (2014).
- M. J. Morshed, J. B. Hmida, and A. Fekih, "A probabilistic multi-objective approach for power flow optimization in hybrid wind-PV-PEV systems," *Appl. Energy* **211**, 1136–1149 (2018).
- E. Xydas, M. Qadrdan, C. Marmaras, L. Cipcigan, N. Jenkins, and H. Ameli, "Probabilistic wind power forecasting and its application in the scheduling of gas-fired generators," *Appl. Energy* **192**, 382–394 (2017).
- X.-Y. Ma, Y.-Z. Sun, and H.-L. Fang, "Scenario generation of wind power based on statistical uncertainty and variability," *IEEE Trans. Sustainable Energy* **4**, 894–904 (2013).
- 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," *J. Electr. Eng. Technol.* **8**, 1615–1625 (2013).
- C. Tang, Y. Wang, J. Xu, Y. Sun, and B. Zhang, "Efficient scenario generation of multiple renewable power plants considering spatial and temporal correlations," *Appl. Energy* **221**, 348–357 (2018).
- Z. Wang, C. Shen, and F. Liu, "A conditional model of wind power forecast errors and its application in scenario generation," *Appl. Energy* **212**, 771–785 (2018).
- F. Hafiz, A. R. de Queiroz, and I. Husain, "Coordinated control of PEV and PV-based storage system under generation and load uncertainties," in 2018 IEEE Industry Applications Society Annual Meeting (IAS) (2018), pp. 1–5.
- K. Høyland, M. Kaut, and S. W. Wallace, "A heuristic for moment-matching scenario generation," *Comput. Optim. Appl.* **24**, 169–185 (2003).
- D. Lee and R. Baldick, "Load and wind power scenario generation through the generalized dynamic factor model," *IEEE Trans. Power Syst.* **32**, 400–410 (2017).
- S. I. Vagropoulos, E. G. Kardakos, C. K. Simoglou, A. G. Bakirtzis, and J. P. Catalão, "ANN-based scenario generation methodology for stochastic variables of electric power systems," *Electr. Power Syst. Res.* **134**, 9–18 (2016).
- Y. Xu, L. Shi, and Y. Ni, "Deep-learning-based scenario generation strategy considering correlation between multiple wind farms," *J. Eng.* **2017**, 2207–2210.
- H. Wei, Z. Hongxuan, D. Yu, W. Yiting, D. Ling, and X. Ming, "Short-term optimal operation of hydro-wind-solar hybrid system with improved generative adversarial networks," *Appl. Energy* **250**, 389–403 (2019).
- Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, "Model-free renewable scenario generation using generative adversarial networks," *IEEE Trans. Power Syst.* **33**, 3265–3275 (2018).
- G. Sideratos and N. D. Hatzigiorgiou, "Probabilistic wind power forecasting using radial basis function neural networks," *IEEE Trans. Power Syst.* **27**, 1788–1796 (2012).
- J.-F. Toubeau, J. Bottieau, F. Vallée, and Z. De Grève, "Deep learning-based multivariate probabilistic forecasting for short-term scheduling in power markets," *IEEE Trans. Power Syst.* **34**, 1203–1215 (2019).
- G. Papaefthymiou and D. Kurowicka, "Using copulas for modeling stochastic dependence in power system uncertainty analysis," *IEEE Trans. Power Syst.* **24**, 40–49 (2009).
- J. E. B. Iversen and P. Pinson, "Resgen: Renewable energy scenario generation platform," in 2016 IEEE Power Engineering Society General Meeting (IEEE, 2016).
- F. Golestaneh, H. B. Gooi, and P. Pinson, "Generation and evaluation of space-time trajectories of photovoltaic power," *Appl. Energy* **176**, 80–91 (2016).
- W. Hu, Y. Min, Y. Zhou, and Q. Lu, "Wind power forecasting errors modelling approach considering temporal and spatial dependence," *J. Mod. Power Syst. Clean Energy* **5**, 489–498 (2017).
- D. L. Woodruff, J. Deride, A. Staid, J.-P. Watson, G. Slevogt, and C. Silva-Monroy, "Constructing probabilistic scenarios for wide-area solar power generation," *Sol. Energy* **160**, 153–167 (2018).
- M. Kazerooni, H. Zhu, and T. J. Overbye, "Literature review on the applications of data mining in power systems," in *Power and Energy Conference at Illinois (PECI)*, 2014 (IEEE, 2014), pp. 1–8.
- F. Vallée, G. Brunieau, M. Pirlot, O. Deblecker, and J. Lobry, "Optimal wind clustering methodology for adequacy evaluation in system generation studies using nonsequential Monte Carlo simulation," *IEEE Trans. Power Syst.* **26**, 2173–2184 (2011).
- L. Dong, L. Wang, S. F. Khahro, S. Gao, and X. Liao, "Wind power day-ahead prediction with cluster analysis of NWP," *Renewable Sustainable Energy Rev.* **60**, 1206–1212 (2016).
- A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Trans. Power Syst.* **32**, 3258–3265 (2017).
- B. Drake and K. Hubacek, "What to expect from a greater geographic dispersion of wind farms?—A risk portfolio approach," *Energy Policy* **35**, 3999–4008 (2007).
- F. Roques, C. Hiroux, and M. Sagan, "Optimal wind power deployment in Europe—A portfolio approach," *Energy Policy* **38**, 3245–3256 (2010).
- Y. Ma, T. Runolfsson, and J. Jiang, "Cluster analysis of wind turbines of large wind farm with diffusion distance method," *IET Renewable Power Gener.* **5**, 109–116 (2011).

- ⁴³P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson New International Edition (Pearson Education Limited, 2013).
- ⁴⁴C. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer Science & Business Media, 2013).
- ⁴⁵G. Casella and E. I. George, "Explaining the Gibbs sampler," *Am. Stat.* **46**, 167–174 (1992).
- ⁴⁶Y. Dvorkin, Y. Wang, H. Pandzic, and D. Kirschen, "Comparison of scenario reduction techniques for the stochastic unit commitment," in 2014 IEEE PES General Meeting (2014), pp. 1–5.
- ⁴⁷J. M. Morales, S. Pineda, A. J. Conejo, and M. Carrion, "Scenario reduction for futures market trading in electricity markets," *IEEE Trans. Power Syst.* **24**, 878–888 (2009).
- ⁴⁸D. Xu, Z. Chen, and L. Yang, "Scenario tree generation approaches using K-means and LP moment matching methods," *J. Comput. Appl. Math.* **236**, 4561–4579 (2012).
- ⁴⁹J. Dupačová, N. Gröwe-Kuska, and W. Römisch, "Scenario reduction in stochastic programming," *Math. Program.* **95**, 493–511 (2003).
- ⁵⁰C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, "Probabilistic forecasting of wind power generation using extreme learning machine," *IEEE Trans. Power Syst.* **29**, 1033–1044 (2014).
- ⁵¹P. Lauret, M. David, and P. Pinson, "Verification of solar irradiance probabilistic forecasts," *Sol. Energy* **194**, 254–271 (2019).
- ⁵²H. Quan, A. Khosravi, D. Yang, and D. Srinivasan, "A survey of computational intelligence techniques for wind power uncertainty quantification in smart grids," *IEEE Trans. Neural Networks Learn. Syst.* (in press).
- ⁵³T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
- ⁵⁴H. Hersbach, "Decomposition of the continuous ranked probability score for ensemble prediction systems," *Weather Forecasting* **15**, 559–570 (2000).
- ⁵⁵M. Carrión and J. M. Arroyo, "A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem," *IEEE Trans. Power Syst.* **21**, 1371–1378 (2006).
- ⁵⁶N. Kumar, P. Besuner, S. Lefton, D. Agan, and D. Hilleman, "Power plant cycling costs," Technical Report No. NREL/SR-5500-55433 (National Renewable Energy Laboratory (NREL), Golden, CO, USA, 2012).
- ⁵⁷EIA, *Monthly Energy Review* (U.S. EIA, 2018).
- ⁵⁸R. Surendran, W. Hogan, H. Hui, and C.-N. Yu, *Scarcity Pricing in ERCOT* (FERC Technical Conference, 2016).
- ⁵⁹Public Utility Commission of Texas, "Chapter 25: Rules," *Electric Substantive Rules* (Public Utility Commission of Texas, 2018).
- ⁶⁰ERCOT, *ERCOT Nodal Protocols*(ERCOT, 2019).
- ⁶¹U.S. Energy Information Administration, *State Electricity Profiles 2019* (U.S. Energy Information Administration, 2019).
- ⁶²C. Draxl, A. Clifton, B.-M. Hodge, and J. McCaa, "The wind integration national dataset (wind) toolkit," *Appl. Energy* **151**, 355–366 (2015).
- ⁶³J.-P. Watson, D. L. Woodruff, and W. E. Hart, "PySP: Modeling and solving stochastic programs in Python," *Math. Program. Comput.* **4**, 109–149 (2012).
- ⁶⁴C. Tseng, "On power system generation unit commitment problems," Ph.D. thesis, University of California, Berkeley (1998).
- ⁶⁵J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Comput. Manage. Sci.* **2**, 3–19 (2005).
- ⁶⁶J. R. Birge, "Decomposition and partitioning methods for multistage stochastic linear programs," *Oper. Res.* **33**, 989–1007 (1985).
- ⁶⁷B. Rachunok, A. Staid, J.-P. Watson, D. L. Woodruff, and D. Yang, "Stochastic unit commitment performance considering Monte Carlo wind power scenarios," in *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)* (IEEE, 2018), pp. 1–6.
- ⁶⁸K. Kazor and A. S. Hering, "Assessing the performance of model-based clustering methods in multivariate time series with application to identifying regional wind regimes," *J. Agric., Biol., Environ. Stat.* **20**, 192–217 (2015).
- ⁶⁹E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.* **7**, 358–386 (2005).
- ⁷⁰Y. Fu, *Human Activity Recognition and Prediction* (Springer, 2016).
- ⁷¹A. Sklar, "Random variables, joint distribution functions, and copulas," *Kybernetika* **9**, 449–460 (1973).