Applied Energy 293 (2021) 116964

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Machine learning enabled reduced-order scenario generation for stochastic analysis of solar power forecasts

S. Bhavsar[1], R. Pitchumani [*,1], M.A. Ortega-Vazquez[1,2]

Advanced Materials and Technologies Laboratory, Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA 24061, USA

ARTICLE INFO

Keywords:
Renewable energy
Machine learning
Uncertainty management
Probabilistic forecast
Scenario reduction

ABSTRACT

With increased reliance on solar-based energy generation in modern power systems, the problem of managing uncertainty in power system operation becomes crucial. However, in order to properly capture the uncertainty spread of the power forecast time series along with all its statistical properties, a large number of scenarios are normally required to be simulated at significant computational cost. This work presents a novel and efficient method to generate statistically accurate scenarios from probabilistic forecasts and a method based on unsupervised machine learning to reduce the number of scenarios and speed up the computations, while preserving the statistical properties of the original set. Through a systematic parametric study, an optimum clustering-based machine learning method and its associated parameters are derived. This approach yields statistically equivalent characteristics as a full set with a substantially reduced cardinality (from 7000 to 20). The reduced set of scenarios also preserves the temporal correlation, which is imperative in time-series data and complies with the non-parametric distribution of power obtained from a probabilistic forecast at any particular time. Applying the optimal algorithm to the benchmark RTS-GMLC and the actual California ISO yearly solar production data, it is shown that the uncertainty in the estimation of the statistical moments is reduced to less than 2% and 4.5% of the respective daily peak power values.

1. Introduction

As the penetrations of solar generation deepen into power systems [1], it becomes critical to properly capture the increased uncertainty introduced when planning the operation of power systems. Dynamic reserve quantification considering uncertainty, e.g. [2], and stochastic generation scheduling tools, [3,4], are examples of methods used to account for uncertainty in operational planning, which requires the construction of a set of scenarios. Even though the best representation of input uncertainty is via continuous random variables in the form of an analytical probability density function (PDF), it is challenging to use such distributions directly in decision making under uncertainty systems of realistic size [5]. The problem becomes even more complicated when the inputs are correlated temporally and spatially [6–8], such as in the case of time-series data. In these cases, it is practical to work with a deterministic model that can represent different realizations of the stochastic variable, which are simply discrete selections from the PDFs. The classical numerical sampling-based simulation methods such as Monte Carlo (MC) simulation and Latin hypercube sampling (LHS) are based on this approach [9]. In the application of stochastic programming or uncertainty quantification (UQ) it is fundamental to have an accurate set of such realizations that would represent the underlying distribution of random variables while being of a manageable size.

In time-series, the number of input variables is equal to the number of time-indices over the study horizon, where each variable is assumed to have unique marginal non-parametric distribution and to be correlated to the other variables. The probabilistic forecast represents each variable as an independent random number and forecasts its values at different quantiles. Hence, it is expected that different time-indices over the planning horizon may follow different distributions. In other words, the probabilistic forecast does not provide any temporal dependency amongst different time indices. Hence, it is required to interconnect different quantiles of the different variables to form time-series, which yields one potential realization (one scenario) of the stochastic process over the planning horizon. The probabilistic scenarios should preserve the temporal dependencies and respect the non-parametric probabilistic information. Such scenarios can be generated using the concept of the inverse distribution function [10] and the correlation matrix [11,12].

* Corresponding author.
E-mail address: pitchu@vt.edu (R. Pitchumani).
[1] All authors contributed equally to this article.
[2] Electric Power Research Institute, Palo Alto, California.

https://doi.org/10.1016/j.apenergy.2021.116964
Received 31 January 2021; Received in revised form 21 March 2021; Accepted 15 April 2021
Available online 26 April 2021

A similar concept was used in [7] and in [13] to generate a set of wind power production scenarios.

The required number of realizations/scenarios can be determined through stochastic convergence analysis. In most of the cases, the convergence of the first two moments (mean and standard deviation) is sufficient to capture the relevant statistical characteristics [14]. As pointed out in [14], the moment matching of input (scenarios) is necessary to accurately predict the output statistics. The number of scenarios needed for convergence is theoretically infinite and practically large, which makes the approach impractical for scenario-based stochastic analysis and decision-making, such as dynamic reserve estimation and stochastic unit-commitment. A practical alternative is to work with a reduced set of scenarios that retains the essential features of the original distribution [5]. Following this approach, Marwadi and Pitchumani [15] developed a technique called Stochastic Analysis with Minimal Sampling (SAMS) for UQ in physical models. The method works well on practical problems where the nature of the meta-model and the output distribution can be parameterized based on physics considerations. QUICKER is another method similar to SAMS that is reported in the literature for UQ in physical models [16]. It was documented that QUICKER reduced the computation time by 95% compared to LHS. Although SAMS and QUICKER are suited for independent input variables, the applicability of these methods for correlated-input variable problems such as time-series is not established.

Several scenario reduction methods for time-series problems are reported in the literature. Hu and Li [17] proposed clustering-based scenarios reduction, using the similarity function presented in [18] and correlation loss to preserve the correlation between multiple variables in multi-stochastic variable programming. Dupacova et al. [19], on the other hand, demonstrated the advantage of using Fortet–Mourier type probability metrics to reduce the number of scenarios when the distributions of random variables follow standard parametric distributions and are known *a priori*. However, in both of these approaches, the derived scenarios are probabilistic and different from the full set (i.e., the reduced scenarios are a transformed version of a full set of scenarios). Along the lines of selecting a scenario from the full set, Sumaili et al. [20] proposed a method that generates the clusters iteratively based on the areas of high density using particle swarm optimization (PSO) and selects the centroid of each cluster as a candidate in a reduced set. The authors, however, did not discuss the statistical moments of the reduced scenario or the computational time overhead due to the optimization task.

An alternative method is based on the concept of a fast forward selection (FFS) [21] that iteratively selects the scenarios from the original set, based on the shortest distance from the remaining scenarios. The method, however, appears to preserve the first moment well, but not the second moment. To make decisions on the reduced scenario based on the response variable, Wang [22] proposed the Forward Selection Wait and See Cluster (FSWC) method, which considers the impact on the response variable. A similar concept was also presented in [23] by considering the distance metric between two scenarios as a function of the response variable. However, these methods may increase the computational time to obtain a reduced set of scenarios due to the increased burden associated with running the input–output model. Li [24] proposed the Heuristic Search (HS) method to get a reduced set quickly by minimizing the average moment (average of the moment of all variables) mismatch between reduced set and the whole set, that works well in predicting the statistics when all variables have the same marginal distribution.

It is evident from the foregoing discussion that there are three primary limitations of the existing approaches to scenario reduction for stochastic analysis. First, the approaches do not discuss the statistical moments of the reduced scenario or are restricted to predicting the first moment only, whereas the reduced scenarios are truly required to preserve higher order moments such as variance of the original distribution as well. In some cases, the derived scenarios differ in characteristics from the full set. The reported methods, therefore, are not sufficiently accurate for an effective stochastic analysis and optimization of power systems. Second, several of the methods in the literature use computationally intensive approaches, including numerical optimization, or incur increased computational burden associated with running the input–output models, to obtain scenario reduction. The additional computational tedium defeats the primary purpose of seeking reduced scenarios namely, improving computational efficiency in applications to stochastic unit commitment or stochastic economic dispatch. Third, the methods are limited to cases when all variables have the same marginal distribution or when they follow a standard family of parametric distributions, whereas in practice the distributions provided by a probabilistic forecast [25] are often non-parametric that do not obey this assumption.

The present work addresses the knowledge gaps mentioned above and introduces a novel approach to generate accurate reduced order scenario sets directly from the non-parametric probabilistic forecast in a computationally efficient manner, without any prior assumption on the marginal distribution functions. The *novelty* and *specific contributions* are as follows: an efficient method is presented for scenario generation that accurately incorporates the statistics of the non-parametric probabilistic forecast into the generated samples; substantial scenario reduction is achieved by creating a small number of clusters from a large set of scenarios using unsupervised machine learning and then identifying appropriate selections from each cluster to derive the desired equiprobable scenarios; considering the power profile on a typical day in the benchmark RTS-GMLC [26] solar power data, a systematic study is conducted to derive the optimum clustering parameters that lead to the best moment-matched selections; the effectiveness of the reduced scenarios in increasing the accuracy and reducing the uncertainty in the estimation of the statistical moments is demonstrated by considering the probabilistic power forecasts for each hour and each day in a full year of the benchmark RTS-GMLC [26] and the actual California ISO (CAISO) [27] datasets.

The article is organized as follows: the scenario generation method is presented in Section 2, followed by a study of scenario reduction based on unsupervised machine learning in Section 3. Section 4 discusses the results of reduced scenarios on RTS-GMLC and CAISO data.

2. Scenario generation method

As discussed in the previous section, the independent quantiles of power which are obtained from the probabilistic forecast are required to be transformed into temporally correlated scenarios. Fig. 1 depicts the scenario generation method used in this analysis, which is built on the method presented in [7,8] with a few modifications as described in the steps below, each corresponding to those in the four boxes of Fig. 1:

1. The first step consists of generating a matrix of an independent identically distributed sample, $\mathbf{X}$, of size $\tau \times N_s$ from a unit Normal distribution with zero mean and unit standard deviation, such that $\mathbf{X} \sim N(0,1)$, where τ is the number of variables, that depends on the forecast horizon and its resolution (number of time indices over forecast horizon). Since the current study deals with day-ahead forecast with a one-hour resolution, $\tau = 24$ corresponding to the 24 hourly values of power in a day, and N_s is the number of desired scenarios. Conventionally, the random set, $\mathbf{X}$, is generated using a Monte Carlo (MC) sampling method. However, stratified sampling methods such as LHS could improve the convergence characteristics of the moments of the distribution with fewer samples compared to a purely random Monte Carlo sampling. Accordingly, the present approach uses the stratified sampling method, LHS, to generate a random selection of $\mathbf{X}$ instead of the conventional MC sampling. In the stratified sampling approach, the PDF of a Normal distribution

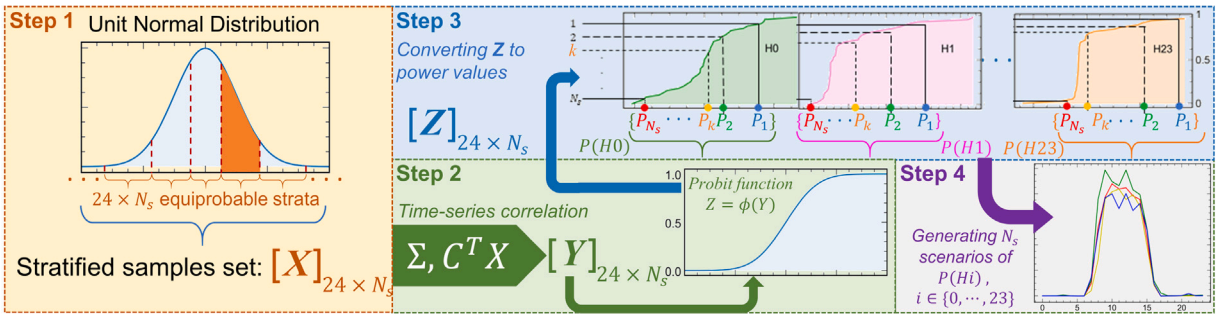

Fig. 1. Schematic illustration of the four steps in the scenario generation method.

is partitioned into $24 \times N_s$ equiprobable strata, and a single selection is then made from each stratum. The selection is then reshuffled and stored in a matrix of size $24 \times N_s$, as shown in Step 1 of Fig. 1.

2. The selected $\mathbf{X}$ needs to be correlated along the 24 hourly values to preserve the temporal dependency of the time series. To achieve the correlation, we first define a covariance matrix, Σ of dimension 24×24, as follows:

$$\Sigma = \begin{bmatrix} \rho & \rho(1-\omega) & \dots & \rho(1-\omega)^{23} \\ \rho(1-\omega) & \rho & \dots & \rho(1-\omega)^{22} \\ \vdots & \vdots & \vdots & \vdots \\ \rho(1-\omega)^{23} & \rho(1-\omega)^{22} & \dots & \rho \end{bmatrix}$$

where the (i, j) element of σ is the correlation between the ith and the jth random variable and is a function of correlation coefficient (ρ) and correlation decay (ω) that are each in the range $[0, 1]$. In the present implementation, ρ was taken to be 0.8, and ω was set to be 0.08. A more systematic selection of ρ and ω could be made recursively based on observed historical trends [7], but these parameter values were found to be adequate for the present study. From a Cholesky decomposition [28], C, of Σ, we can obtain a correlated set $\mathbf{Y}$ corresponding to the uncorrelated set $\mathbf{X}$ as $\mathbf{Y} = C^T\mathbf{X}$, where the superscript T denotes the transpose. The discussed operation yields a matrix $\mathbf{Y}$ of correlated rows, of dimension $24 \times N_s$, as shown in Step 2 of Fig. 1. The resulting matrix $\mathbf{Y}$ is then passed through a probit function to get corresponding quantile values, $\mathbf{Z} = \phi(\mathbf{Y})$, where ϕ is the probit function/CDF of the normal Gaussian distribution, and $\mathbf{Z}$ is a set of correlated random quantile values, of dimension $24 \times N_s$, in the range $[0, 1]$, which is the outcome of Step 2 of Fig. 1.

3. The $\mathbf{Z}$ matrix obtained from Step 2 is of dimension $24 \times N_s$ with correlated rows. There are 24 different CDFs corresponding to the 24 different power values in the forecast horizon. Step 3 of Fig. 1 depicts this with three example CDFs of power at hour-index H0, H2, and H23. Each row of $\mathbf{Z}$, Z_i where $i \in \{0, 2, \dots, 23\}$, is now passed through the corresponding inverse CDF function of the power at ith hour-index to convert each row into its associated power values, $P(Hi)$, a vector of dimension N_s. For illustration, four realizations of the N_s power values at each hour-index are depicted with the different color dots in Step 3 of Fig. 1.

4. Each of the N_s realizations in $P(Hi)$, $i \in \{0, 2, \dots, 23\}$ is linearly interconnected with the corresponding realization at the neighboring hour-indices, $P(Hi-1)$ and $P(Hi+1)$ to get a temporally correlated distinct scenario, for a total scenario set cardinality of N_s, as shown in Step 4 of Fig. 1.

In the above procedure, Step 3 requires a knowledge of CDFs of the power values at each of the τ indices. In this particular application, we need to approximate CDF using discrete quantile points available form a probabilistic forecast. Conventionally, for statistical data with discrete quantiles, a CDF is constructed by piecewise linear interpolation between adjacent points. However, such an approach is known to underestimate the second moment of the distribution [29]. Note that the "moment" here refers to the moment of piecewise linear CDF. In order to improve the accuracy of the second moment estimation, in the present approach, the CDF is subjected to two modifications as discussed in [29]: (1) The first step is to stretch the CDF in the horizontal direction by horizontal adjustment of node points (i.e., quantile points which are used for forming piecewise linear CDF), wherein every node is subjected to different amounts of stretching such that the normalized gap [29] between the adjacent nodes remains the same for a chosen support of modified piecewise linear CDF, and (2) the second step involves the shifting of each node by the same amount in the horizontal direction, determined by solving the quadratic equation corresponding to the variance of the piecewise-linear CDF. The first step ensures a match of the second moment, and the second step ensures a match of the first moment. Such modification is applied to the marginal power distributions at each of the 24-hour indices. However, to accomplish the match of moments in CDFs, the target values of moments need to be defined. The target values of moments are estimated from finite quantiles (40 in the present case) available from the probabilistic forecast using the bootstrap technique [30]. The modified CDF, which has the same first two moments as a target, is then used to obtain a corresponding N_s-dimensional vector of power values, $P(Hi)$, by following Steps 1 through 4, as illustrated in Fig. 1.

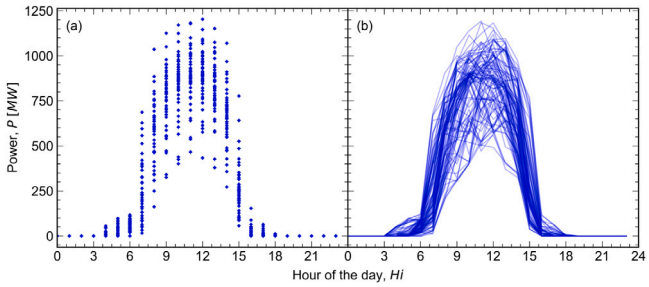

Fig. 2. Example scenarios generated from the probabilistic hourly solar production forecast on a typical day in the RTS-GMLC data set: (a) set of 40 discrete quantile values for each hour index and (b) $N_s = 100$ scenarios generated using the quantile values.

As suggested in Step 4, adjacent hour-index values, $P(Hi-1)$, $P(Hi)$, and $P(Hi+1)$, are linearly connected to produce one realization of the stochastic process—called one power scenario. To get a set of scenarios, one would need to compute different quantiles of power. The quantile values for power are obtained by fitting a quantile regression model [25] on historical data following the boosting approach [31], which can be used to generate a set of scenarios. Fig. 2 depicts such a set of scenarios that are generated by following Step 1 to 4 from discrete quantile values of a probabilistic forecast of hourly solar power production (in MW) on a typical day in the RTS-GMLC data set [26].

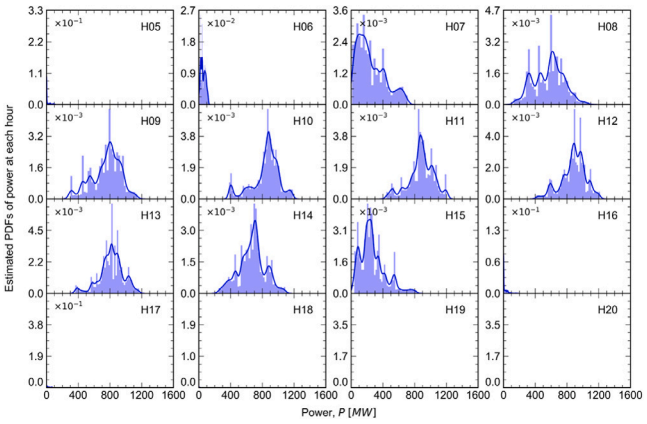

Fig. 3. Probability density functions of power at each hour index during a typical day in the RTS-GMLC data set. Note that the hours of midnight to 4 A.M. and 9 P.M. to 11 P.M. have zero power production and are omitted for brevity.

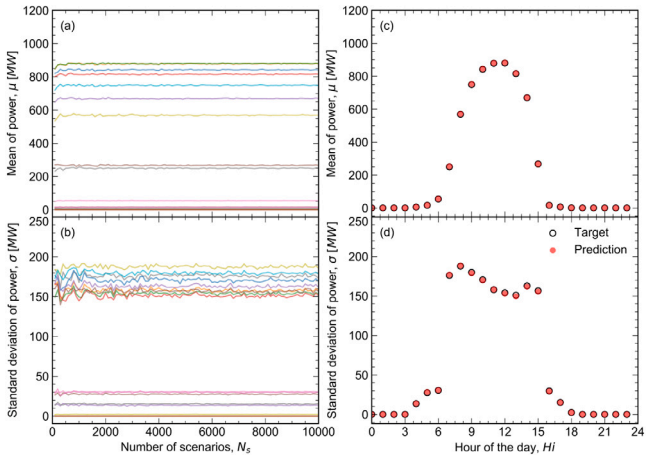

Fig. 4. (a,b) Variation of the mean and standard deviation with the number of scenarios, and (c,d) comparison of the mean and standard deviation based on 7000 samples with those from the probabilistic forecast.

Fig. 2a shows the scattered 40 quantile values of power P (in MW) at hour index Hi (hour of the day) where $i \in \{0, 1, \ldots, 23\}$. By following the method discussed so far, the discrete quantile values are converted into a desired set of scenarios, as shown in Fig. 2b, which presents a superposition of 100 such scenarios that retain the desired temporal correlation ($\rho = 0.8, \omega = 0.08$) and preserve the moments of the distribution at each time index.

The discussion in Fig. 2b is limited to a finite number of scenarios for illustration purposes only. Theoretically, an infinite number of scenarios are needed to get an accurate estimation of the statistical moments (i.e., mean, variance, skewness, and kurtosis) of the underlying distribution. Fig. 3 depicts the nature of the distribution of power at each hour of the day (periods of zero solar power production, e.g., H00 (12:00 midnight) to H04 (4:00 A.M.) and H21 (9:00 P.M.) to H23 (11:00 P.M.) are omitted in Fig. 3). The presented distributions are estimation from a large finite number of scenarios. As shown in Fig. 3, the nature of the distribution is multi-modal and is difficult to be parameterized. In practical considerations, therefore, the challenge is to obtain an accurate moment estimation with as few scenarios as possible, which is presented in the next section.

3. Scenario reduction

As motivation for the need for scenario reduction, we consider the stochastic nature of the statistics of the set of scenarios for each of the 24 hourly marginal CDFs. Fig. 4a and b show the variation in the mean (first moment), μ, (Fig. 4a) and the standard deviation (the second moment), σ, (Fig. 4b) of a set with respect to the number of scenarios (N_s). The different colors in each plot (Fig. 4a and Fig. 4b) correspond to the different hour index. Overall, it is seen from Fig. 4a that the first moment converges rapidly for all 24 variables at about $N_s = 1000$ scenarios. However, the convergence of the second moment to within 0.1% is seen to be relatively slow, requiring $N_s = 7000$ scenarios, Fig. 4b. Fig. 4c and d show the comparison of the first and second moments evaluated using the 7000 scenarios (red dot markers) with the target values (open black circle markers) obtained as the bootstrapped quantile values at each time index as discussed in the previous section. Fig. 4c shows that the estimation of the mean is nearly identical to the target value at every time index, and the estimation of the standard deviation also closely matches the target values for all time indices, as evident in Fig. 4d. Overall, the scenario generation method is seen to be effective in retaining essential statistical characteristics. Based on the accuracy of estimating both the mean and the standard deviation, it reasonable to infer that the results for 7000 scenarios are deterministic since the uncertainty associated with $N_s = 7000$ is quite small as per Fig. 4a and b. Accordingly, 7000 scenarios were considered to yield the "exact" and "deterministic" moments for comparison with the reduced scenario method discussed in the section.

Although 7000 scenarios accurately and deterministically estimate the first two moments, such a set would render any stochastic application intractable. While using fewer scenarios reduces the computational burden, it also introduces bias and uncertainty in the estimation of the moment, as evident in Fig. 4a and b. The bias accounts for the average performance of the scenario reduction method, while uncertainty accounts for the associated degree of certainty in estimating the moment through the scenario reduction method. An appropriate scenario reduction method should estimate the moment with as small bias and uncertainty as possible with a significant reduction of cardinality.

Estimating the second moment is challenging compared to estimating the first moment; however, the methods used to improve the estimation of the second moment would inherently improve the estimation of the first moment as well. The goal of the study is to investigate methods to reduce the number of scenarios (N_s) while simultaneously reducing the bias and uncertainty of the resulting standard deviation (second moment). To this end, the application of an unsupervised machine learning technique called clustering [32] is explored in detail. The general idea behind the method is to make a finite number of clusters of "similar-characteristic" scenarios and then to make an appropriate number of selections, n_i, from each cluster, rather than analyzing the entire cardinality of all clusters. The schematic of the method is illustrated in Fig. 5, where the full set of N_s number of scenarios is divided into N heterogeneous clusters, shown by three representative colors in Fig. 5, from which three scenarios are selected from Cluster 1, two from Cluster 2, and one from Cluster N, for a total of six scenarios in the illustration. In general, therefore, the number of a reduced set of scenarios becomes $\widehat{N}_s = \sum_{i=1}^{N} n_i$ where n_i is the number of selections from an ith cluster.

Considering four different clustering schemes, a systematic parametric study is performed to arrive at the optimum combinations of various associated parameters. Table 1 summarizes the five parameters and their respective levels as follows:

1. *Clustering algorithm*: Clustering allows one to identify and classify interesting patterns and similarities in the underlying data [33], which together with data mining of sequential data such as time-series has received significant attention [34–37]. The set of solar production scenarios can similarly be classified though such an unsupervised method whose parameters can be optimized based on desired target objectives. Four clustering algorithms are chosen for evaluation in this study, as shown in Table 1: a k-Shape

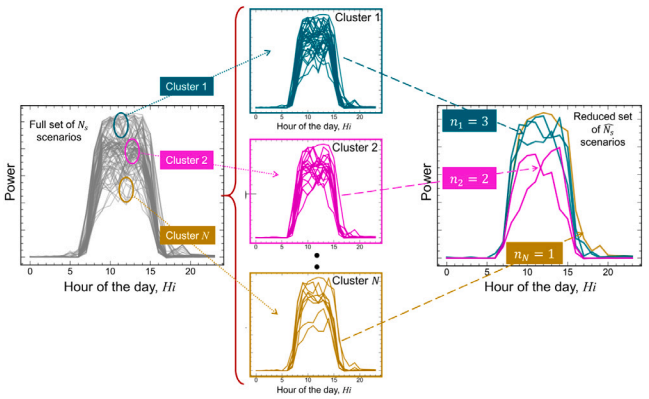

Fig. 5. Schematic illustration of the clustering approach.

clustering approach and three k-means clustering methods. The k-Shape clustering method utilizes cross-correlation as a distance measure that is invariant to scaling and shifting of time-series, which makes it efficient for time-series clustering [37]. The other three are k-means clustering with three different similarity metrics, namely, standardized euclidean (seuclidean), cosine, and correlation [38].

2. *Features*: Clustering is performed either on power-based scenarios or quantile-based scenarios (where the corresponding quantile values replace the power values). A Z-score [39] normalization was applied to the power-based time series before implementing a clustering algorithm to avoid the domination of large numbers. The number of clusters formed is N.
3. *Intra-cluster selection*: After making clusters, the next task is to make a selection from within each of the N clusters. Two mechanisms of selection are considered. The first is a Random Selection (RS), and the second is a Fast Forward Selection (FFS) [5]. Instead of using a single method, the present work explored the hybrid version of both by assigning φ fraction of clusters to RS and the remaining $(1-\varphi)$ fraction to the FFS selection scheme. Five levels of φ are considered in the present parametric study, namely, 0, 0.25, 0.50, 0.75, and 1, such that $\varphi = 0$ represents a fully FFS scheme whereas $\varphi = 1$ denotes purely random selection.
4. *Inter-cluster-selection*: This parameter pertains to selection amongst clusters. In this, two levels are considered: (1) Even selection from each cluster, where the number of selections from a cluster (n) is the same across all clusters (2) Uneven selection from each cluster where the number of selections from each cluster is proportional to the cluster size. However, an *average selection* (n) from each cluster in an uneven selection scheme is defined as $n = \widehat{N}_s/N$. The number of selections from an ith cluster is defined as: $n_i = \left\lfloor C_i \times \widehat{N}_s/N_s \right\rfloor$, where C_i is the size of ith cluster, ($\widehat{N}_s$), is the reduced number of scenarios, N_s (= 7000) is a full set of scenarios, and $\lfloor\cdot\rfloor$ denotes the floor function.
5. *Combinations of* (N,n): This parameter is a combination of (N,n), where N is the number of clusters, and n is an average number of selections from each cluster (for uneven selection, it is $(\widehat{N}_s/N)$. A total of 30 different possible combinations of (N,n) are considered such that $5 \le N \times n \le 120$. For example, for 10 scenarios, the (N,n) combinations consist of (10,1), (5,2), and (2,5).

It is evident from Table 1 that the total number of the different cases spanning the five parameters is: 4 clustering algorithms × 2 features × 5 values of φ × 2 Inter-cluster selection methods × 30 combinations of (N,n) = 2400. Further, each of the 2400 cases was run 50 times to get different replicates of response to observe the associated uncertainty in terms of an inter-quartile range. The results of the parametric study are discussed in the next section.

Table 1
Parameters considered in the clustering algorithm optimization.

Factors	Levels
Clustering algorithm	k-Shape k-means (seuclidean) k-means (cosine) k-means (correlation)
Clustering feature	Power, Quantile
Intra-cluster selection (φ) Inter-cluster selection	0, 0.25, 0.50, 0.75, 1.00 Even selection, Uneven selection
Combinations of (N,n)	30 levels in the range $5 \le N \times n \le 120$

4. Results and discussion

This section is subdivided into two parts. Section 4.1 discusses the method adopted for the parametric study to arrive at the optimum combinations of parameters, while Section 4.2 presents the application of the derived parameters and its performance on the open-source datasets.

4.1. Parametric study

The effectiveness of a clustering algorithm is normally defined based on a clustering loss function, which is the sum of the squares of the distances of each data point to its assigned cluster centroid [32]. In the present study, the goal of the clustering-based scenario reduction is to assess its effectiveness in estimating the statistical moments of the power distributions, particularly the standard deviation, which in turn determines the effectiveness of the method in practical applications such as stochastic unit commitment and economic dispatch. To this end, two related quantities are defined to evaluate the performance of the scenario reduction method: (1) Normalized Bias, $\delta = \frac{\|\mathbb{E}[\hat{\sigma}]-\sigma\|_2}{\|\sigma\|_1}$, where σ is a collection of the target values of the standard deviation of power at each hour index, the ith element of σ is σ_i, where $i \in [0,1,2,\ldots,23]$, $\mathbb{E}[\hat{\sigma}]$ is a collection of the average values of the standard deviation of the 50 replicates of the reduced set of scenarios at each hour, $\|\cdot\|_2$ denotes the L_2 norm of a quantity, and $\|\cdot\|_1$ denotes L_1 norm of a quantity, and (2) Normalized Inter Quantile Range (IQR), $\gamma = \frac{\|iqr\|_1}{\|\sigma\|_1}$, where iqr is the collection of inter-quantile range (i.e., the difference between values at the 75%-ile and the 25%-ile) of power based on the 50 replicates at each hour. The goal of the clustering algorithm is to reduce the number of scenarios such that both the normalized bias, δ, and the normalized inter-quartile range, γ, are minimized simultaneously. To this end, the results of the parametric study are analyzed in terms of the product $\gamma \times \delta$ as the quantity to be minimized.

Figs. 6 and 7 summarize the result of parametric studies, where each subplot has the evaluation metric $\gamma \times \delta$ on the ordinate and pair of (N,n) on the abscissa, where n denotes the equal number of selections each cluster in the case of an even clustering scheme and the *average selection* (n) from each cluster in an uneven selection scheme. With the intent of focusing on the fewest number of scenarios, only those combinations of (N,n) which lead to $\widehat{N}_s(= N \times n)$ of 20, 30, or 40 are considered in the figures, and the different color shades represent the different clustering algorithms. Within each subplot, the different densities of hatches represent a different number of reduced scenarios, ($\widehat{N}_s$). Each row corresponds to one realization of φ, starting from $\varphi = 0$ at the top to $\varphi = 1$ at the bottom. Each column within one color shade represents one of the two features, power or quantile. The two bars within each subplot represent two levels of selection schemes, even or uneven selection. The optimum combinations of the parameters are ones that minimize the $\gamma \times \delta$. As seen from Figs. 6 and 7, overall, k-means

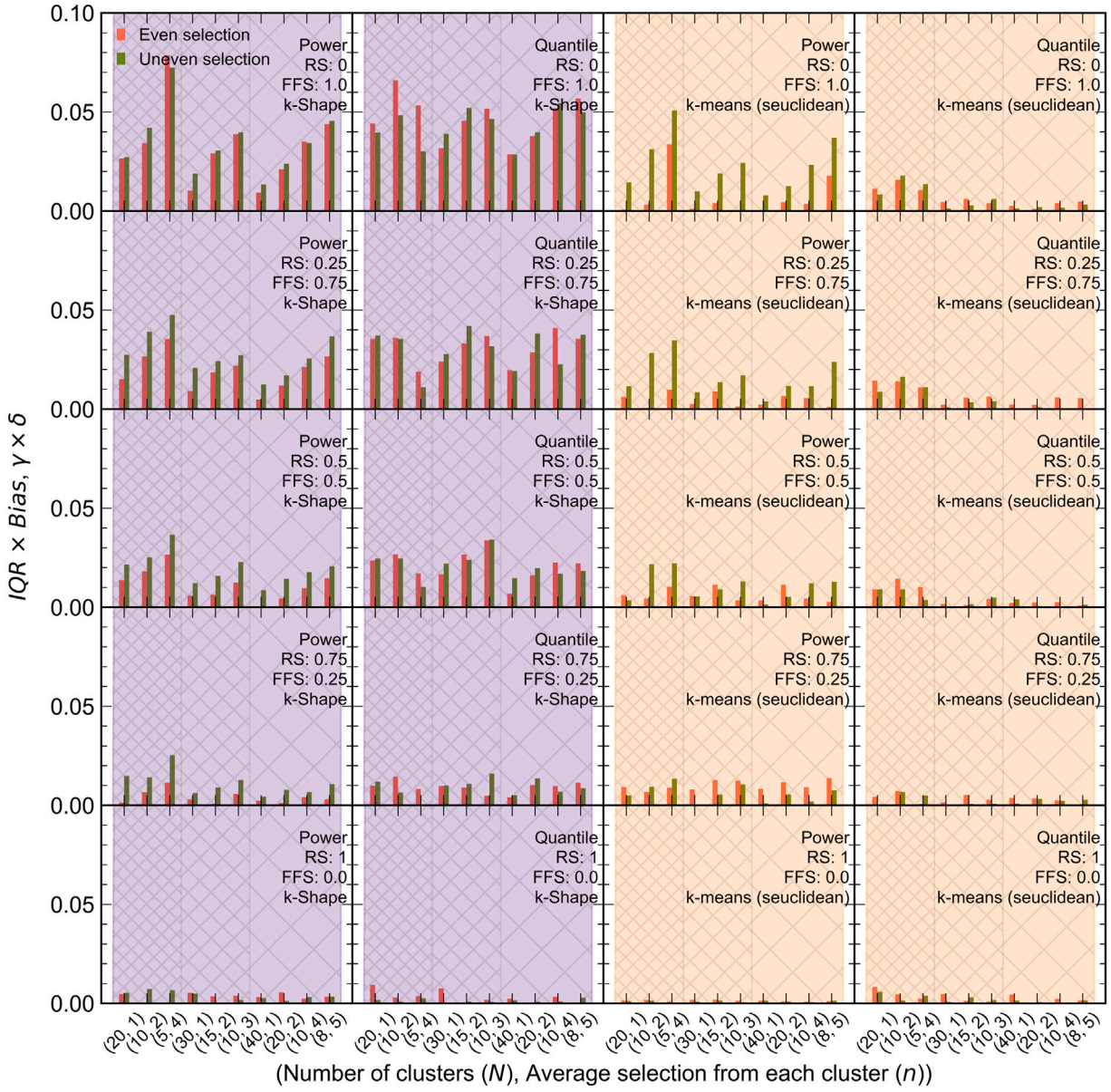

Fig. 6. Results of the parametric study: k-Shape and k-means (seuclidean) clustering algorithms for the different levels of the governing parameters.

clustering worked better than the k-Shape clustering in almost all cases. The optimum level of the other parameters irrespective of the clustering algorithm, for the three values of $\widehat{N}_s$s, are Power (Feature), Random Intra-cluster Selection ($\varphi = 1$), and Uneven Inter-cluster selection. The difference in the response amongst different distance metrics is marginal. Overall, it is also observed that the smaller value of n in an (N, n) pair, for the same $\widehat{N}_s$, leads to better performance in almost all cases.

Fig. 8 shows the variation of $\gamma \times \delta$ for each clustering algorithm, as a metric to be minimized, with the number of scenarios, $\widehat{N}_s$s, for the optimum (N, n) pair corresponding to each $\widehat{N}_s$. As can be seen from Fig. 8, the variation of $\gamma \times \delta$ with $\widehat{N}_s$ forms a characteristic elbow pattern, similar to that observed in any general clustering application [32]. For k-means clustering, the reduction in $\gamma \times \delta$ is seen to be insignificant after $\widehat{N}_s = 20$ suggesting that 20 scenarios may be considered an optimum that maximizes accuracy of estimation of the statistical parameters. Further, there is a little variation amongst the different distance metrics (seuclidean, cosine, and correlation) for k-means clustering. However, k-means with correlation distance metric achieved a slightly lower $\gamma \times \delta$ than others. The variation for k-Shape clustering is seen to minimize $\gamma \times \delta$ at $\widehat{N}_s = 30$, suggesting that this algorithm would require more scenarios than the k-means algorithm. It may be inferred from Fig. 8 that the k-means (with correlation distance metric) and $\widehat{N}_s = 20$ $(N = 20, n = 1)$ is the best clustering algorithm out of the considered search space. Note that in comparison to the 7000 scenarios needed for accurate determination of the mean and the standard deviation of the probabilistic distribution, the use of the optimal clustering strategy yields a substantial reduction of over 99.7% in the number of scenarios to only 20. It is also worth noting that the computation time associated with the optimum clustering method is about 45 s (on a 3.10 GHz Intel Core i9 CPU with 32 GB Memory). In comparison, the computation time for k-Shape clustering is on the order of 300 s.

4.2. Application of derived algorithm

As mentioned previously, the contributions of this work are twofold: (i) a scenario generation through stratified sampling followed by stretching and shifting of the distribution functions (discussed in Section 2), and (ii) a scenario reduction technique that uses the optimal number

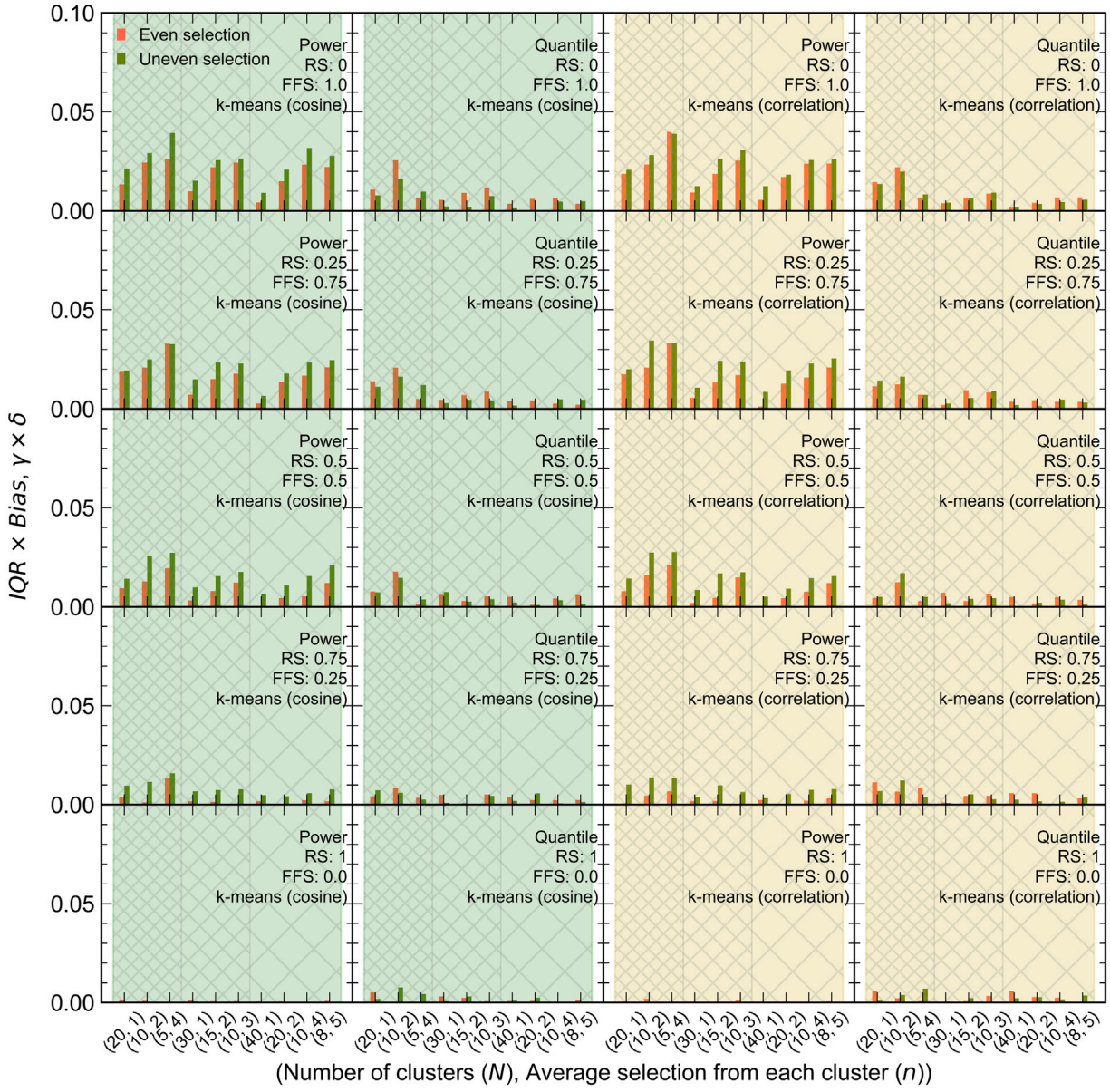

Fig. 7. Results of the parametric study: k-means (cosine) and k-means (correlation) clustering algorithms for the different levels of the governing parameters.

of reduced scenarios (discussed in Section 3). The effectiveness of each of these contributions on reducing the uncertainty of the standard deviation is examined in Fig. 9. Fig. 9a presents box plots of the standard deviation of the power determined from 50 sets of 20 randomly selected scenarios at each hour of the day (without the application of the modifications in the scenario generation method discussed in Section 2) as a function of the hour index. Also plotted for comparison is the target standard deviations shown by the black markers that correspond to a set of 7000 scenarios. Fig. 9b presents box plots of the standard deviation computed using 50 sets of 20 scenarios at each hour of the day, in which the modifications in the scenario generation method discussed in Section 2 are applied. As before, the target values of standard deviation obtained from a set of 7000 scenarios are indicated by the black markers. Fig. 9c combines the scenario generation method and the optimal scenario reduction algorithm and presents the standard deviation computed using 50 sets of 20 optimally clustered scenarios at each hour of the day.

When randomly selecting 20 scenarios, the median values of the standard deviation in the box plots in Fig. 9a are seen to be consistently underpredicted relative to the target value. Furthermore, there is a significantly large associated uncertainty, as observed form the inter-quantile range of the boxplots. With the incorporation of modifications in the scenario generation, the median of the box plots are seen to be closer to the target values in Fig. 9b, resulting in a reduced bias, although the uncertainty in the estimate, as indicated by the height of the box plots, is high. Using the combined approach of improved scenario generation and the reduced set of $\widehat{N_s} = 20$ scenarios obtained through the optimal k-means clustering algorithm, both the bias and the associated uncertainty are seen to be reduced significantly in Fig. 9c. The size of the boxplot in Fig. 9c is so small that it is indistinguishable from the size of the markers in the plot. The finding in Fig. 9 clearly indicates the accuracy of the present method, Fig. 9c, compared to the conventional method of scenario generation in Fig. 9a.

The parametric study to determine the optimum clustering algorithm and its parameters uses one typical day in the RTS-GMLC [26] data. In order to further explore the applicability of the optimal clustering algorithm and its parameters to other data sets, the k-means (correlation) clustering scheme with a reduced set of 20 scenarios is applied to the yearly data from the RTS-GMLC data set [26] and the CAISO yearly data on solar power production for the year December

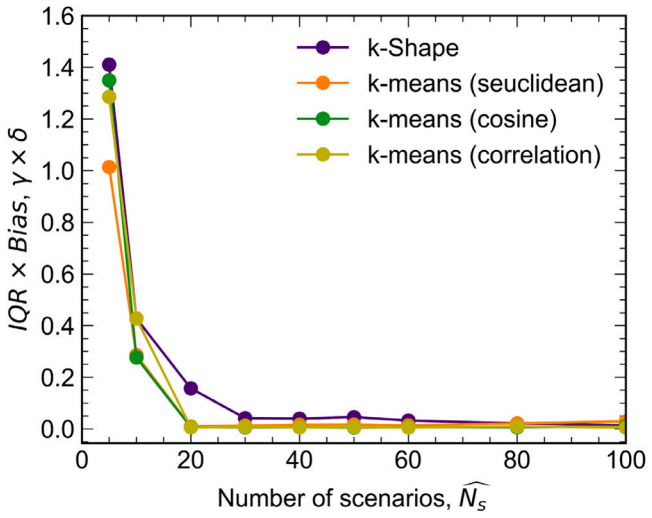

Fig. 8. Variation of IQR × bias of the clustering schemes with the reduced number of scenarios, $\widehat{N_s}$.

2018–December 2019 [27]. For the CAISO data set, the probabilistic forecast on annual solar power production is first obtained via quantile regression [25] through the boosting algorithm [31]. The probabilistic forecast is then converted into a set of scenarios, as shown in Fig. 10. The data in Fig. 10 is ordered by the four seasons—winter, spring, summer, and fall—in the four rows from top to bottom. Further, within each row, five random days from each season are chosen to depict the generated scenarios. The black dashed line in each subplot represents the actual power production on that particular day. The two black dotted lines around the actual power production represent the 25th and 75th quantiles, respectively. The full cardinality of scenarios is considered to be 7000, which is similar to what was considered in RTS-GMLC data. However, only 50 representative scenarios (out of 7000) are shown in Fig. 10 for the sake of clarity. As can be seen from Fig. 10, on average, winter days produce less solar power compared to days in other seasons. Furthermore, the summer days follow a systematic and uniform pattern in solar power production as compared to days in other seasons that exhibit abrupt changes and fluctuations in the profiles.

Fig. 11 shows the estimated hourly standard deviation with $\widehat{N_s} = 20$ for a single randomly chosen day from each season. It is seen that the estimate of the standard deviation is able to match the target with a small amount of uncertainty, and with a significantly reduced bias. It is revealed from the plot that winter and fall days show a relatively larger amount of standard deviation in power production than spring and summer days. Additionally, the standard deviation of power at peak hours is quite similar across the range for spring and summer days. However, there seems to be a non-uniform trend in the standard deviation of power across the spectrum of peak hours for winter and fall days.

Fig. 12 shows the comparison of the estimated standard deviation for $\widehat{N_s} = 20$ scenarios with the target standard deviation obtained for 7000 scenarios, for each hour of the year for the RTS-GMLC data set (Fig. 12a) and the CAISO data set (Fig. 12b). The ordinate of Figs. 12a,b is an expected value of standard deviation obtained as an arithmetic average of all 50 replicates of the 20 sets of scenarios at each hour of the day. A comparison of the value with the target standard deviation indicates the bias in the prediction, as defined in the previous section. The solid lines in the plots indicate the line of exact agreement, and the shaded bands denote the 10% error band with respect to the exact value. It is seen that most of the predictions, 93.4% of the hourly values over the year, fall within the 10% error band for the RTS-GMLC data (Fig. 12a) and in the CAISO data set (Fig. 12b), 93.3% of hourly data over the entire year fall within the 10% error band, suggesting a very small bias in the estimation of the hourly standard deviation throughout the year for both data sets.

To examine the uncertainty in the estimation of the standard deviation, Figs. 12c,d present the histogram of IQR of prediction for each hour of the year for RTS-GMLC and CAISO solar power production data sets, respectively. In Fig. 12c, the distribution of IQR for the RTS-GMLC data set follows a bimodal distribution with peaks at around 2 MW and 10 MW. The bimodal nature of the distribution of IQR may be attributed to the different ranges of solar power production in a single day, which can be distinguished as low production hours, midnight to 6 A.M. and 5 P.M. to 11 P.M., and high production hours, 7 A.M. to 4 P.M. For the CIASO data in Fig. 12d, the histogram of the hourly IQR also follows a bimodal distribution similar to that observed for the RTS-GMLC data set in Fig. 12c. The peaks, however, occur close to 0 GW and 0.2 GW in Fig. 12d, denoting an extremely low level of uncertainty in the estimated standard deviation. In operation, it is desirable that the uncertainty in estimating the standard deviation of power at peak hours be as small as possible. It is seen that the IQR stays well within 20 MW in Fig. 12c, which corresponds to less than 2% uncertainty on a peak power of ~1000 MW. For the CAISO solar power production data set, the uncertainty in the estimated hourly standard deviation, IQR, is below ~0.5 GW in Fig. 12d, which is less than 4.5% of the peak power values in a day.

Following the format of comparison on the hourly standard deviation in Fig. 12, Fig. 13 compares the daily-averaged standard deviation from the 7000 scenarios with that estimated by the 20 scenarios from the optimal k-means clustering scheme, for all days of a year. The daily-averaged standard deviations for the ramp-up (12 A.M. to 1 P.M.) and the ramp-down (1 P.M. to 12 midnight) periods are obtained by taking the L_2 norm of the two sets of the respective hourly values of standard deviation in a single day. The comparison between the L_2 norm of estimated standard deviation for 20 scenarios and the L_2 norm of the target standard deviation is presented in Fig. 13a for the RTS-GMLC data set and in Fig. 13b for the CAISO data. As in Figs. 12a,b, the solid lines represent the line of exact agreement, the red and blue markers represent the estimated standard deviation of the optimal 20 scenarios for the ramp-up and the ramp-down periods, respectively, and the shaded areas denote the 10% error band. It is seen that all the daily averaged standard deviation values, spanning all the ramp-up and the ramp-down periods, from the optimum clustering scenarios fall well within the 10% error band, for both RTS-GMLC (Fig. 13a) and CAISO data (Fig. 13b). Further, the inset plots within Figs. 13a,b present the histogram of percentage error for up and down ramp periods, which is within ±8% and centered around 0% for the RTS-GMLC data (Fig. 13a). The error distribution is seen to follow a Gaussian distribution, which is also a desirable characteristic [40]. The inset plot in Fig. 13b for the CAISO data also confirms that the error is normally distributed around 0 and is within about 8% in almost all cases.

Figs. 13c,d show the histogram of the IQR of the daily-averaged standard deviation for ramp-up (red bars) and ramp-down (blue bars) periods for the RTS-GMLC and CAISO data, respectively. Unlike the hourly IQR in Figs. 12c,d, the IQR for the daily-averaged standard deviation follows a unimodal distribution with a peak at around 9 MW for ramp-down and at approximately 12 MW for ramp-up period for the RTS-GMLC data set in Fig. 13c, which is on average within about 1% of the peak power of ~1000 MW in a day. For the CAISO data, Fig. 13d shows that the histogram of daily IQR is unimodal with a peak of around 0.25 GW for both the ramp-down and the ramp-up periods, with the distribution shifted to slightly higher power values for the ramp-up period. However, the IQR of the daily-averaged standard deviation is all within about 0.45 GW, which corresponds to an uncertainty of less than 4% relative to the daily peak power value of about 11.25 GW.

Figs. 12 and 13 demonstrate the general applicability of the optimal clustering scheme with reduced set of 20 scenarios to a full year of data in the RTS-GMLC and the CAISO data sets. These results offer independent validation that the optimal clustering scheme is capable

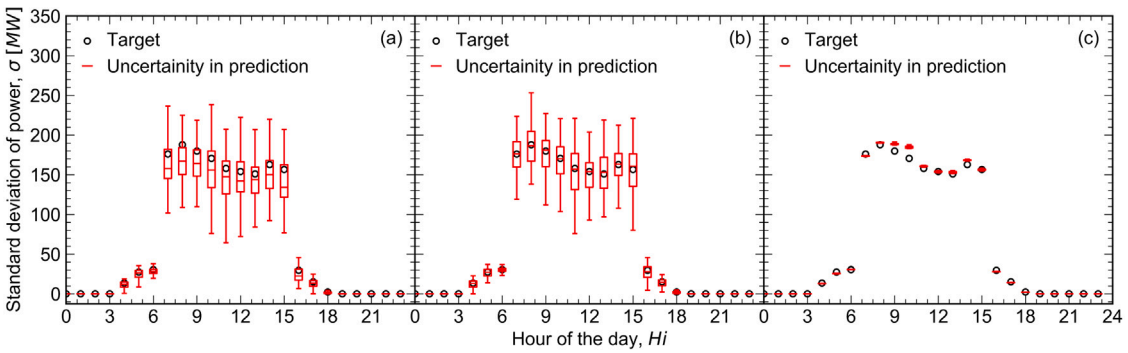

Fig. 9. Standard deviation of the reduced set of 20 scenarios: (a) randomly selecting 20 out of a full set of scenarios using Monte Carlo sampling from an unmodified piecewise linear CDF, (b) randomly selecting out of the full set of scenarios using stratified sampling from a modified piecewise linear CDF, and (c) selection using the optimal k-means (correlation) clustering algorithm.

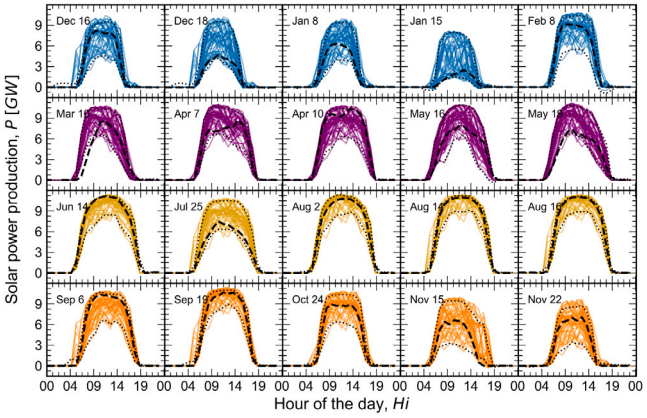

Fig. 10. Representative 50 scenarios of solar power generation in CAISO data for the four seasons: winter (top row), spring (second row from top), summer (third row from top), and fall (bottom row), for randomly selected five days within each season.

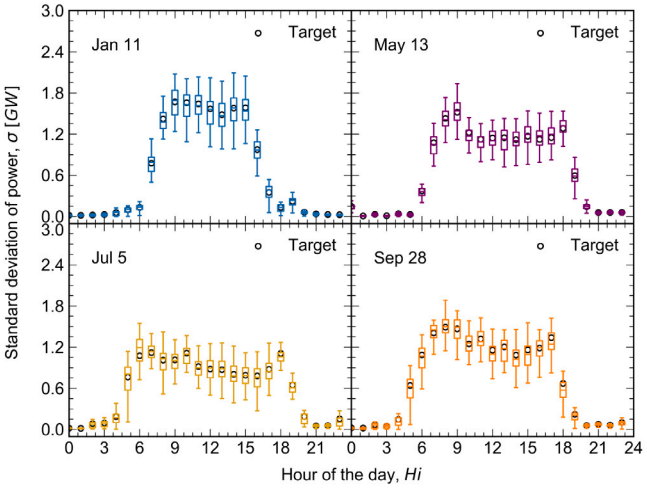

Fig. 11. Estimated hourly standard deviation with $\widehat{N}_s = 20$ for a single randomly chosen day from each season, from the CAISO data set.

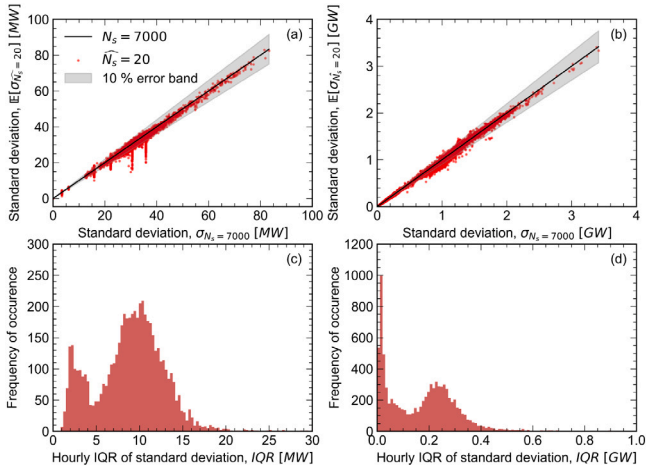

Fig. 12. Comparison of the standard deviation for each hour of a year of (a) RTS-GMLC data and (b) CAISO data obtained from the optimal 20 scenarios with those obtained from the 7000 scenario simulation, and histogram of the IQR of the hourly standard deviation for (c) RTS-GMLC data and (d) CAISO data.

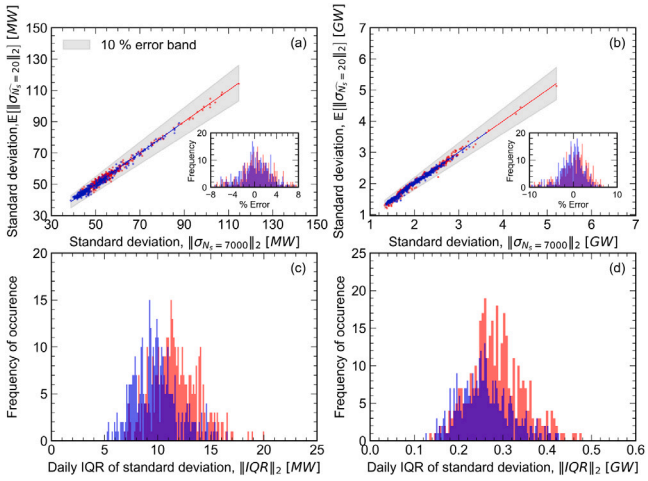

Fig. 13. Comparison of the daily-averaged standard deviation in a year of (a) RTS-GMLC data and (b) CAISO data obtained from the optimal 20 scenarios with those obtained from the 7000 scenario simulation, and histogram of the IQR of the daily-averaged standard deviation for (c) RTS-GMLC data and (d) CAISO data.

of accurately estimating the mean and the standard deviation at the hourly or daily resolution beyond the typical day that formed the basis of the parametric studies used to derive the optimal scheme. Further research, building on the results in this article, might explore the application of the techniques presented for ramp rates as well as reserve estimations. The focus of this article is primarily on the solar production. However, the method presented here can be utilized to generate reduced order scenarios for other renewable energies such as wind and wave energy from their respective probabilistic forecasts.

Other machine learning methods, such as the application of encoder–decoder to encode the essential features in reduced dimensions might be a promising non-linear transformation [41], which may be explored in a future study.

The primary contribution of the present work is a method to accurately derive a reduced set of scenarios of a solar power realization from a non-parametric probabilistic forecast. Unlike previously reported approaches, the presented methodology does not make any assumption on the prior marginal distribution at each time indices. The present method yields accurate statistics from a reduced number of scenarios that, in turn, can be used for accurate and computationally-efficient estimation of uncertainty in further decision variables, such as reserve estimation, cost, dispatch, etc., derived from scenarios [14]. The application of the method to unit commitment and reserve estimation may be considered in a future work.

5. Conclusions

The article presented an improved method for scenario generation and scenario reduction in probabilistic modeling of power forecasting. In the scenario generation aspect, the effect of stratified sampling coupled with the piecewise-linear CDF modification was shown to yield accurate estimation of the second moment compared to a full set of scenarios. On scenario reduction front, the article presented an optimal algorithm and its parameters for uncertainty quantification in the statistics of power profiles with a significantly reduced number of scenarios (20 vs. 7000). The optimum parameters were derived by considering a typical day in the RTS-GMLC data set through an exhaustive parametric study on varying the clustering algorithm and their parameters. The generalized applicability of the algorithm across a broader set of power data sets that were not used in developing the optimum parameters was demonstrated by considering the yearly variation of the power profiles from both RTS-GMLC as well as CAISO data sets. It was shown that the use of clustering reduced the uncertainty in the estimation of the statistical parameters to below 2–4.5% based on the peak daily power.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments and disclaimer

The material is based upon work supported by the U.S. Department of Energy under award number DE-EE0008601. Their support is gratefully acknowledged. The authors also acknowledge Dr. Aidan Tuohy of EPRI for the technical discussions that contributed to the work reported in the article. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

[1] Sun Y, Wachche SV, Mills A, Ma O. 2018 Renewable Energy Grid Integration Data Book.
[2] Vos KD, Stevens N, Devolder O, Papavasiliou A, Hebb B, Matthys-Donnadieu J. Dynamic dimensioning approach for operating reserves: Proof of concept in Belgium. Energy Policy 2019;124:272–85. http://dx.doi.org/10.1016/j.enpol.2018.09.031.
[3] Constantinescu EM, Zavala VM, Rocklin M, Lee S, Anitescu M. A computational framework for uncertainty quantification and stochastic optimization in unit commitment with wind power generation. IEEE Trans Power Syst 2011;26(1):431–41.
[4] Dvorkin Y, Pandžić H, Ortega-Vazquez MA, Kirschen DS. A hybrid stochastic/interval approach to transmission-constrained unit commitment. IEEE Trans Power Syst 2015;30(2):621–31.
[5] Feng Y. Scenario Generation and Reduction for Long-term and Short-term Power System Generation Planning under Uncertainties, https://lib.dr.iastate.edu/etd/14148.
[6] Lange M. Analysis of the uncertainty of wind power predictions, https://oops.uni-oldenburg.de/id/eprint/233.
[7] Pinson P, Madsen H, Nielsen HA, Papaefthymiou G, Klöckl B. From probabilistic forecasts to statistical scenarios of short-term wind power production. Wind Energy 2009;12(1):51–62. http://dx.doi.org/10.1002/we.284, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.284, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/we.284.
[8] Papaefthymiou G, Schavemaker P, van der Sluis L, Kling W, Kurowicka D, Cooke R. Integration of stochastic generation in power systems. Int J Electr Power Energy Syst 2006;28(9):655–67. http://dx.doi.org/10.1016/j.ijepes.2006.03.004.
[9] Papoulis A. Probability and Statistics. USA: Prentice-Hall, Inc.; 1990.
[10] Young GA. Multivariate statistical simulation.. J R Stat Soc Ser A 1988;151(1):229–30. http://dx.doi.org/10.2307/2982203.
[11] Chakraborty A. Generating multivariate correlated samples. Comput Statist 2006;21(1):103–19. http://dx.doi.org/10.1007/s00180-006-0254-y.
[12] Charmpis DC, Panteli PL. A heuristic approach for the generation of multivariate random samples with specified marginal distributions and correlation matrix. Comput Statist 2004;19(2):283–300. http://dx.doi.org/10.1007/BF02892061.
[13] Ortega-Vazquez MA, Kirschen DS. Assessing the impact of wind power generation on operating costs. IEEE Trans Smart Grid 2010;1(3):295–301.
[14] Smith JE. Moment methods for decision analysis. Manage Sci 1993;39(3):340–58. http://dx.doi.org/10.1287/mnsc.39.3.340.
[15] Mawardi A, Pitchumani R. SAMS: Stochastic analysis with minimal sampling—A fast algorithm for analysis and design under uncertainty. J Mech Des 2004;127(4):558–71. http://dx.doi.org/10.1115/1.1866157.
[16] Donato A, Pitchumani R. QUICKER: Quantifying uncertainty in computational knowledge engineering rapidly—A rapid methodology for uncertainty analysis. Powder Technol 2014;265:54–65. http://dx.doi.org/10.1016/j.powtec.2014.01.028.
[17] Hu J, Li H. A new clustering approach for scenario reduction in multi-stochastic variable programming. IEEE Trans Power Syst 2019;34(5):3813–25.
[18] Xie M, Guo J, Zhang H, Chen K. Research on the similarity measurement of high dimensional data. Comput Eng Sci 2010;32(5):92–6.
[19] Dupačová J, Gröwe-Kuska N, Römisch W. Scenario reduction in stochastic programming an approach using probability metrics. Math Program 2003;95(3):493–511.
[20] Sumaili J, Keko H, Miranda V, Botterud A, Wang J. Clustering-based wind power scenario reduction technique, in: 17th Power Systems Computation Conference, Stockholm Sweden, 2011.
[21] Heitsch H, Römisch W. A note on scenario reduction for two-stage stochastic programs. Oper Res Lett 2007;35(6):731–8.
[22] Wang Y. Scenario reduction heuristics for a rolling stochastic programming simulation of bulk energy flows with uncertain fuel costs (Ph.D. thesis), Iowa State University; 2010.
[23] Morales JM, Pineda S, Conejo AJ, Carrion M. Scenario reduction for futures market trading in electricity markets. IEEE Trans Power Syst 2009;24(2):878–88.
[24] Li J, Lan F, Wei H. A scenario optimal reduction method for wind power time series. IEEE Trans Power Syst 2016;31(2):1657–8.
[25] Liu B, Nowotarski J, Hong T, Weron R. Probabilistic load forecasting via quantile regression averaging on sister forecasts. IEEE Trans Smart Grid 2017;8(2):730–7.
[26] Barrows C, Bloom A, Ehlen A, Ikäheimo J, Jorgenson J, Krishnamurthy D, Lau J, McBennett B, O'Connell M, Preston E, Staid A, Stephen G, Watson J. The IEEE reliability test system: A proposed 2019 update. IEEE Trans Power Syst 2020;35(1):119–27.
[27] California ISO - Today's Outlook. URL http://www.caiso.com/TodaysOutlook.
[28] Stewart GW. Matrix Algorithms. Society for Industrial and Applied Mathematics; 1998, http://dx.doi.org/10.1137/1.9781611971408.
[29] Kaczynski W, Leemis L, Loehr N, McQueston J. Nonparametric random variate generation using a piecewise-linear cumulative distribution function. Comm Statist Simulation Comput 2012;41(4):449–68. http://dx.doi.org/10.1080/03610918.2011.606947.
[30] Efron B. Bootstrap methods: Another look at the jackknife. Ann Statist 1979;7(1):1–26.
[31] Chen T, Guestrin C. XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
[32] Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.
[33] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. Intell Inf Syst 2001;17(2–3):107–45.
[34] Bagnall A, Janacek G. Clustering time series from ARMA models with clipped data. In: KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004, p. 49–58.
[35] Fu T-C. A review on time series data mining. Eng Appl Artif Intell 2011;24(1):164–81. http://dx.doi.org/10.1016/j.engappai.2010.09.007.
[36] Papadimitriou S, Sun J, Yu P. Local correlation tracking in time series, in: Proceedings - IEEE International Conference on Data Mining, ICDM 2006, pp. 456–465.

[37] Paparrizos J, Gravano L. K-shape: Efficient and accurate clustering of time series, in: Proc. ACM SIGMOD Int. Conf. Manag. Data 2015, pp. 1855–1870.
[38] Jin R, Goswami A, Agrawal G. Fast and exact out-of-core and distributed k-means clustering. Knowl Inf Syst 2006;10:17–40. http://dx.doi.org/10.1007/s10115-005-0210-0.
[39] Kreyszig E, Kreyszig H, Norminton EJ. Advanced Engineering Mathematics. Tenth edn.. Hoboken, NJ: Wiley; 2011.
[40] James G. An introduction to statistical learning : with applications in R. New York, NY: Springer; 2013.
[41] Madiraju NS, Sadat SM, Fisher D, Karimabadi H. Deep temporal clustering : Fully unsupervised learning of time-domain features. 2018, arXiv:1802.01059.