Statistical metrics for assessing the quality of wind-power scenarios for stochastic unit commitment

Didem Sari, Youngrok Lee, Sarah M. Ryan
Department of Industrial & Manufacturing Systems Engineering, Iowa State University

David L. Woodruff
Graduate School of Management, University of California Davis

June, 2015[1]

ABSTRACT

In power systems with high penetration of wind generation, probabilistic scenarios are generated for use in stochastic formulations of day-ahead unit commitment problems. To minimize the expected cost, the wind power scenarios should accurately represent the stochastic process for available wind power. We employ some statistical evaluation metrics to assess whether the scenario set possesses desirable properties that are expected to lead to a lower cost in stochastic unit commitment. A new mass transportation distance (MTD) rank histogram is developed for assessing the reliability of unequally likely scenarios. Energy scores, rank histograms, and Brier scores are applied to alternative sets of scenarios that are generated by two very different methods. The MTD rank histogram is best able to distinguish between sets of scenarios that are more or less calibrated according to their bias, variability and autocorrelation.

Keywords: Reliability, Energy score, Mass transportation distance, Rank histogram, Brier score.

1. Introduction

The wind energy industry is one of the fastest growing renewable energy industries in the world and global wind power capacity continues to grow rapidly. High penetration of wind power requires more sophistication in operational planning to accommodate variability. One of the most significant short-term planning problems for electrical power generation is unit commitment, in which an optimal on-off schedule is found for each thermal generating unit over a given period of time [1]. Unit commitment problems traditionally have been solved by imposing fixed reserve limits to manage uncertainty in load and small amounts of variable generation. However, in systems with a large amount of wind power, cost savings have been demonstrated by solving stochastic unit commitment (SUC) problems with probabilistic scenarios for the wind power trajectory [2, 3, 4, 5].

Errors in the day-ahead wind power forecast and variability in electricity demand create uncertainty in the forecast for net load, which equals the load less the available wind power. A high level of wind penetration increases that uncertainty. In stochastic unit commitment, implicit

[1] For publication in *Wind Energy*.

reserve levels are identified by finding a unit commitment schedule that minimizes expected costs with respect to a set of probabilistic scenarios. The stochastic optimization approach is based on the concept of recourse. In a two-stage model, a single commitment schedule is determined in the first stage by considering how each unit would be dispatched in each scenario of available wind power in the second, or recourse, stage. Compared to the schedule found with fixed reserve limits, costs are saved by committing more resources on days when uncertainty is high and/or expected wind contribution is low, which avoids having to start up additional generating units in real time, and by committing fewer resources on days when uncertainty is low and/or expected wind contribution is high, thus incurring lower start-up and no-load costs. For the stochastic planning approach to be effective, the scenarios must accurately represent the stochastic process for available wind power given information available when the schedule is generated. The scenario time series of wind power amounts should somehow resemble the corresponding observed time series in attributes such as the levels of wind power available at time points throughout the planning horizon, the correlations among these levels, the presence and severity of ramps, etc., and their probabilities should accurately reflect the frequency of similar occurrences.

Recently, considerable effort has been devoted to developing methods for generating wind power scenarios. Different methods yield sets of scenarios that differ quantitatively and qualitatively, in both obvious and subtle ways. The definitive way to evaluate a scenario generation method is to simulate employing the resulting scenarios in stochastic unit commitment while measuring the costs incurred [6]. However, although computational methods for solving stochastic unit commitment problems have improved significantly [4], a simulation study sufficiently thorough to accurately detect meaningful differences among scenario sets is computationally very demanding. Therefore, in this paper we explore the use of statistical metrics to measure properties of scenarios supposed to be desirable for achieving cost savings in stochastic unit commitment. We apply them to distinguish between two very different scenarios generated by approaches: a statistical approach that combines quantile regression with a Gaussian copula [7] and epi-spline approximation approach. Epi-splines and their applications are discussed in [8] and a similar scenario generation method is used for electricity demand in [5]. Because the latter approach yields scenarios that are not necessarily equally likely, we develop and test a new mass transportation distance (MTD) rank histogram to assess whether scenarios with unequal probabilities have similar temporal patterns as the corresponding observations. We present simulation studies to determine MTD rank histogram features that indicate reliability. The histogram evaluation is combined with comparisons of energy scores [9,10] and event-based Brier scores [9] to compare and contrast multiple attributes of the scenario sets.

This paper examines several verification tools that are used to test wind power scenarios for reliability, sharpness, skill and their ability to capture critical characteristics of stochastic processes. We employ energy scores to inform on the forecast skill of scenarios for individual lead times as done in [9]. The energy score has also been used for probabilistic forecasts of surface wind vectors in [10]. However, when applied to multivariate probabilistic forecasts, it has limited ability to discriminate among sets of forecasts with different levels of autocorrelation [11]. Minimum spanning tree rank histograms are used to check reliability of equally likely scenarios [9] or ensemble forecasts [10,12,13]. Smith suggested the use of MST lengths as a scalar pre-rank function for multidimensional forecasts [14]. The MST rank histogram was further studied by Wilks [12] and Gombos et al. [13]. Because all of the scenario sets generated in [9] were equally likely, MST rank histograms were employed to check the temporal dependence structure. However, one of the scenario generation methods presented in our numerical study produces unequally likely scenarios. To incorporate the probabilities, we employ mass transportation distance [15,16] as a pre-ranking function. The mass transportation distance is motivated by stability analysis for use in scenario reduction for stochastic programming [17]. Finally, event-based verification assesses the ability of wind power scenarios to accurately represent ramp up and ramp down events, which can have a large impact on unit commitment and subsequent dispatch costs. Brier scores [18] are applied as an event-based verification approach as in [9].

There has not been much rigorous evaluation of scenario generation approaches according to their performance in stochastic unit commitment. The study reported in [6] is one exception, where the advantages of using SUC formulations over deterministic ones and the importance of probabilistic wind power scenarios are also emphasized. A small study comparing epi-spline load scenarios with Monte Carlo scenario paths in SUC is reported in [19]. Within scenario generation approaches, different variants and parameter settings can produce different sets of scenarios, and simulating their performance in SUC may be computationally prohibitive. The contribution of this paper is to summarize statistical metrics' capabilities and illustrate their potential use as prescreening tools for either equally or unequally likely scenario sets.

The paper proceeds as follows: The existing statistical metrics for scenario evaluation along with our new MTD rank histogram are explained in detail and some simulation studies on MTD rank histograms are provided in Section 2. The wind power scenario generation methods are described in Section 3 including some variations. In Section 4, we compare the results of the two scenario generation methods according to the metrics using wind power forecast and observational data from a U.S. agency. Finally, we conclude in Section 5 with a brief summary and discussion of research directions.

2. Verification of Scenarios

In this section, some important verification approaches are presented for assessing the quality of scenarios. It is critical to evaluate how well the scenario set reflects the actual wind power output. Some properties of a scenario set are reliability, sharpness, and skill. Reliability refers to the statistical consistency between the probabilistic scenarios and observations [20]. If the relative frequency of occurrence of events assigned a scenario probability tends to be close to the observation, then we accept that scenario set to be reliable or calibrated [21]. Sharpness is the concentration of the scenario distributions. The sharper the scenarios, the less uncertainty they express. What is expected from a forecast is to maximize the sharpness, subject to calibration [20]. Sharpness and calibration are accepted as the components of skill [9]. However, a set of scenarios for stochastic programming has a different purpose than an ensemble of forecasts. A very sharp set of scenarios may not express the uncertainty that decision procedures must consider.

The notation used in this paper is as follows:

$y_d^0 = \{y_{h,d}^0\}$: observed wind power in hour *h*=1,…,*H* on day d =1,…, *D*

$y_d^s = \{y_{h,d}^s\}$: wind power in hour *h*=1,…,*H* on day d =1,…, *D*, in scenario *s* = 1,…, *S*

y_d^{0*} : standardized time trajectory, obtained by scaling the wind power levels, y_d^0 , according to the installed capacity.

y_d^{s*} : standardized time trajectory, obtained by scaling the wind power levels, y_d^s , according to the installed capacity.

$y_d^{s\circ}$: de-biased wind power on day *d* in scenario *s*

z_d^0 : observed wind power trajectory on day *d* after scaling according to Mahalanobis transformation

z_d^s : wind power trajectory on day *d* in scenario *s* after scaling according to Mahalanobis transformation

p_d^s : probability of occurrence of scenario s on day *d*

2.1 Energy Score

The energy score, a multivariate version of continuous rank probability score [7], has been used to measure the skill of scenarios [6,7]. As mentioned above, skill encompasses both calibration and sharpness. Here we explain the energy score in terms of a distance metric. A statistical distance between two probability distributions *F* and *G* can be defined as [22]:

$$D(F,G) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|,$$

where X and X' are independent and identical random vectors having the distribution F , and Y and Y' are independent and identical random vectors having the distribution G . The notation $\mathbb{E}\|.\|$ represents an expectation of the Euclidean norm. Let $F_d(.)$ be the true probability distribution of wind power generation on day *d* and $\hat{F}_d(.)$ be its estimate. An observation of wind power on day *d*, denoted by y_d^{0*} , is the only available sample point from $F_d(.)$, whereas *S* wind power scenarios $\left\{y_d^{1*},\ldots,y_d^{s*}\right\}$ can be seen as approximating or having been sampled from $\hat{F}_d(.)$. Then a distance between $F_d(.)$ and $\hat{F}_d(.)$ is computed by;

$$D\left(F_d,\hat{F}_d\right)=2\sum_{s=1}^{S}p_d^s\left\|y_d^{0*}-y_d^{s*}\right\|-\sum_{s=1}^{S}\sum_{t=1}^{S}p_d^s p_d^t\left\|y_d^{s*}-y_d^{t*}\right\|$$

The energy score is the quantity obtained by dividing $D\left(F_d,\hat{F}_d\right)$ by two:

$$\mathrm{ES}\left(\hat{F}_d,y_d^{0*}\right)=\sum_{s=1}^{S}p_d^s\left\|y_d^{0*}-y_d^{s*}\right\|-\frac{1}{2}\sum_{s=1}^{S}\sum_{t=1}^{S}p_d^s p_d^t\left\|y_d^{s*}-y_d^{t*}\right\|$$

Thus, an energy score can be interpreted as a distance between the true distribution and a scenario distribution of wind power on each day. Therefore, it is a negatively oriented proper score; i.e., lower energy score translates to a higher skill of scenarios [9]. In the case of equally likely scenarios, the formula is simplified as

$$\mathrm{ES}\left(\hat{F}_d,y_d\right)=\frac{1}{S}\sum_{s=1}^{S}\left\|y_d^{0*}-y_d^{s*}\right\|-\frac{1}{2S^2}\sum_{s=1}^{S}\sum_{t=1}^{S}\left\|y_d^{s*}-y_d^{t*}\right\|.$$

A large ES is caused by either the observation being distant from the scenarios or scenarios being too close to each other, or both of these conditions. The ES is not informative with respect to the interdependence structure of the observation or the scenarios [9].

2.2 Distance-based rank histograms

The minimum spanning tree rank histogram was developed to verify the reliability of multidimensional ensemble forecasts. Given a set of *m* points connected by edges, a spanning tree is constructed by selecting *m*-1 edges, such that all points are connected. A minimum spanning tree is a spanning tree with the smallest total edge length (Kruskal, 1956). In the context of evaluating scenarios, we find the MST rank by ordering, from smallest to largest, the lengths of the *S*+1 MSTs that are obtained by only scenario points and by successively substituting the observation for each of the scenario points. The rank histogram plots the frequency of the rank among all of the MST lengths of the MST length that is derived from only scenarios. MST rank histogram construction proceeds as follows [14]:

(a) Standardize the set $\left\{y_d^0,y_d^1,\ldots,y_d^s\right\}$ to obtain a standardized observation y_d^{0*}, and standardized scenarios $y_d^{1*},\ldots,y_d^{s*}$. In the numerical study in Section 4, standardized time trajectories are obtained by scaling the wind power levels according to the installed capacity.

(b) Find the length, l_0, of a MST for the observation from the set $\left\{y_d^{k*} : k \in \{1,...,S\}\right\}$. For each $j = 1,...,S$, compute the MST length, l_j, for scenario *j*, from the set $\left\{y_d^{k*} : k \in \{0,...,S\} \setminus \{j\}\right\}$. In the numerical study when computing the lengths we use the Euclidean (L2) norm; i.e., the distance between y_d^{i*} and y_d^{j*} is $\sqrt{\sum_{h=1}^{H}\left(y_{h,d}^{i*} - y_{h,d}^{j*}\right)^2}$.

(c) Find the MST rank *r*, which is the rank of observation MST length l_0, when $l_0, l_1, \ldots, l_s$ are sorted from smallest to largest. It is an integer between 1 and $S+1$.

For an ideally calibrated ensemble of equally likely scenarios, the probabilities of the observation rank falling into any of the bins are equal. Thus, the resulting MST rank histogram should appear uniform. The lowest MST ranks are seen too often for a biased or under-dispersed ensemble, whereas the highest ranks occur too often for an over-dispersed ensemble [10].

MST rank histograms can be used to assess reliability of equally likely scenarios. For scenarios with different probabilities of occurrence, we have devised the mass transportation distance (MTD) rank histogram for the same purpose. In general, the mass transportation distance between two distributions is the minimum cost of transporting the probability from one distribution to the other, where cost is proportional to the distance between supporting points of the distributions [15, 16]. Although in general the MTD is found by solving a linear program, in our application, it is the minimum cost of transporting the probability from the group to the individual. Assuming an edge exists between each pair of points, the trivial solution to the minimization problem uses the tree composed of the edges between each group member and the individual. Thus, the minimum transportation distance from $\left\{y_d^{k*} : k \in \{1,...,S\}\right\}$ to y_d^{0*} can be computed simply as:

$$\sum_{k=1}^{S}\left\|y_d^{k*} - y_d^{0*}\right\| p_d^k .$$

Note that the MTD is identical to the first term in the energy score formula, which measures reliability. This motivates the use of the MTD as a pre-ranking function for reliability assessment. We compute the rank for the observation distance by successively interchanging each scenario, along with its probability, with the observation. In contrast to the MST length as a pre-ranking function, a relatively small distance from scenarios to observation indicates that the observation falls within the convex hull of scenarios; therefore, we order the MTD values from largest to smallest to determine the rank.

Our studies on MTD rank histogram show that it behaves similarly to the MST rank histogram when applied to equally likely scenarios. Its construction is similar to that of the MST rank histogram. It can be constructed by replacing steps *(b)* and *(c)* with steps *(b)'* and *(c)'* as follows:

(b)' Find the MTD for the observation, l'_0, which is the distance from the set of scenarios $\left\{y_d^{k*} : k \in \{1,...,S\}\right\}$ to the observation y_d^{0*}. Then compute the MTDs for scenarios, l'_j, $j = 1,...,S$, from the set $\left\{y_d^{k*} : k \in \{0,...,S\} \setminus \{j\}\right\}$ to y_d^j. When computing l'_j, assign the probability of scenario y_d^{j*}, which is p_d^j, to the observation y_d^{0*}.

(c)' Find the MTD rank *r*, of the observation MTD l'_0, when $l_0{}', l'_1, \ldots, l'_s$ are ordered from largest to smallest. It is an integer between 1 and $S+1$.

Fig 1 and Fig 2, respectively, illustrate the constructions of both minimum spanning tree and mass transportation distance lengths that could result from an over-dispersed and under-dispersed ensemble.

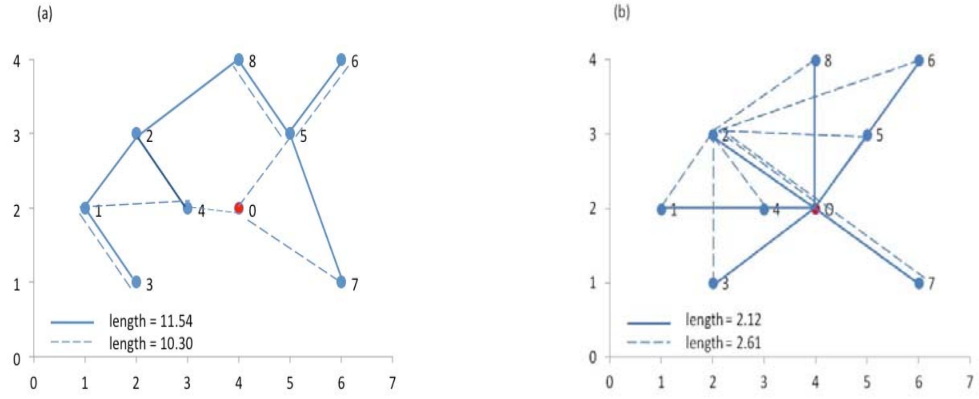


Fig.1 A hypothetical example in 2 dimensions is presented for minimum spanning tree and mass transportation distance. *S* = 8 equally likely scenarios are labeled 1-8 and the corresponding observation is 0. The observation is interior to the scenario points, which causes an over-dispersed ensemble. (a) The solid lines indicate an MST for the scenarios, and the dashed lines indicate an MST that results from the observation being substituted for scenario 2. (b) Similarly, the solid lines indicate the edges used to transport probability from the scenarios to the observation, and the dashed edges are used to transport probability to from all other scenarios plus the observation.

In Fig. 1a, substituting the observation for scenario 2 reduces the MST length. The rank of the solid-line MST depends also on the lengths of the other seven MSTs, which are constructed by replacing each of the seven points by the observation in turn. In Fig. 1a the lengths of six of them are shorter than 11.54, and only the one that results from replacing scenario 4 by the observation is equal to 11.54. Therefore, the rank of the observation's MST length Fig.1a is 8 or 9 out of 9. In Fig. 1b the MTD between scenario 2 and other scenarios along with the observation is 2.61, and the MTD between the scenarios and the observation equals 2.12. Similarly to MST rank, the mass transportation distance rank depends also on the other MTD lengths, which result from replacing each scenario point by the observation in turn. All other MTD lengths are longer than 2.12, which means the rank is 9 out of 9. Thus, the MTD rank agrees with the MST rank in this instance.

We repeat the same process for a biased and/or under-dispersed ensemble in Fig. 2.

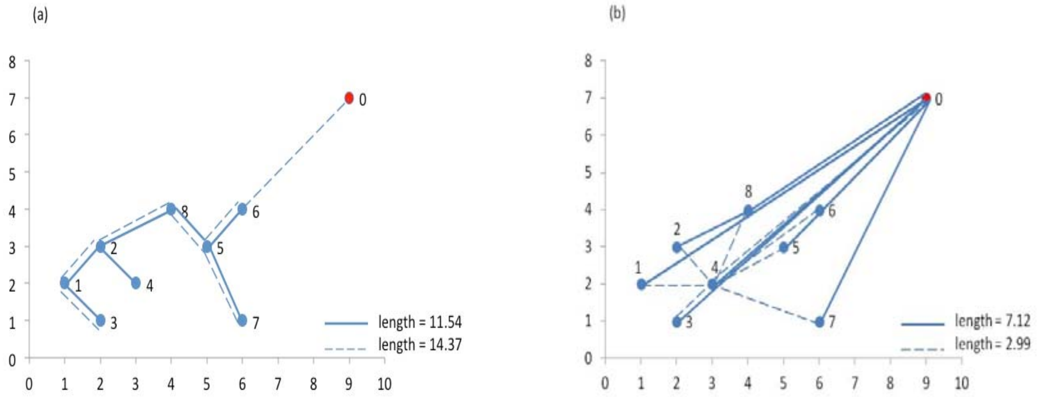


Fig.2 The observation 0, is moved to point (9,7) from point (4,2) to obtain an under-dispersed and/or biased ensemble, whereas the 8 scenarios labeled 1-8, have been kept at their same coordinates. The observation becomes exterior to the convex hull of the scenarios. (a) The solid lines indicate an MST for the scenarios labeled 1-8, and the dashed lines indicate an MST that results from the observation being substituted for scenario 4. (b) Similarly, solid edges are used to transport probability from the scenarios to the observation, and dashed edges are used to transport probability to scenario 4 from all other scenarios plus the observation.

In Fig. 2a, substituting scenario 4 for the observation yields a smaller rank for that MST length. All of the other seven MSTs also have lengths longer than the solid line's length. Therefore, the MST rank of the example shown in Fig. 2a is 1 out of 9. In Fig. 2b the MTD between scenario 4 and other scenarios with the observation, is shorter than the MTD from all scenarios to the observation. The other seven MTDs also have shorter lengths. Because we order the MTDs from largest to smallest, the rank assigned to the observation is 1 out of 9.

In the simulation studies depicted in Figs 3 and 4, respectively, MST and MTD rank histograms are constructed for same sets of observations and hourly scenario values, which are all randomly generated from independent normal distributions. For these equally likely simulated scenarios, both types of histograms show the same patterns as distribution parameters are varied. The horizontal axis identifies the bin and the vertical axis measures the frequencies of the ranks that fall into the corresponding bins. Specifically, the MST and MTD rank histogram both display a downward trend for an under-dispersed ensemble ($\sigma^2_{scen} / \sigma^2_{obs} < 1$) and an upward trend for an over-dispersed ensemble ($\sigma^2_{scen} / \sigma^2_{obs} > 1$), as expected. Flat histograms result when the observation and scenarios are drawn from the same distribution. When the variances are equal, we see a downward trend for larger scenario means, which correspond to bias in the ensemble. Bias over-populates the small ranks similarly as under-dispersion. However, both types of rank histogram appear flat when $\mu_{scen} = 1$ and $\sigma^2_{scen} / \sigma^2_{obs} = 2$. This suggests that high variance in the scenarios can compensate for bias [12].

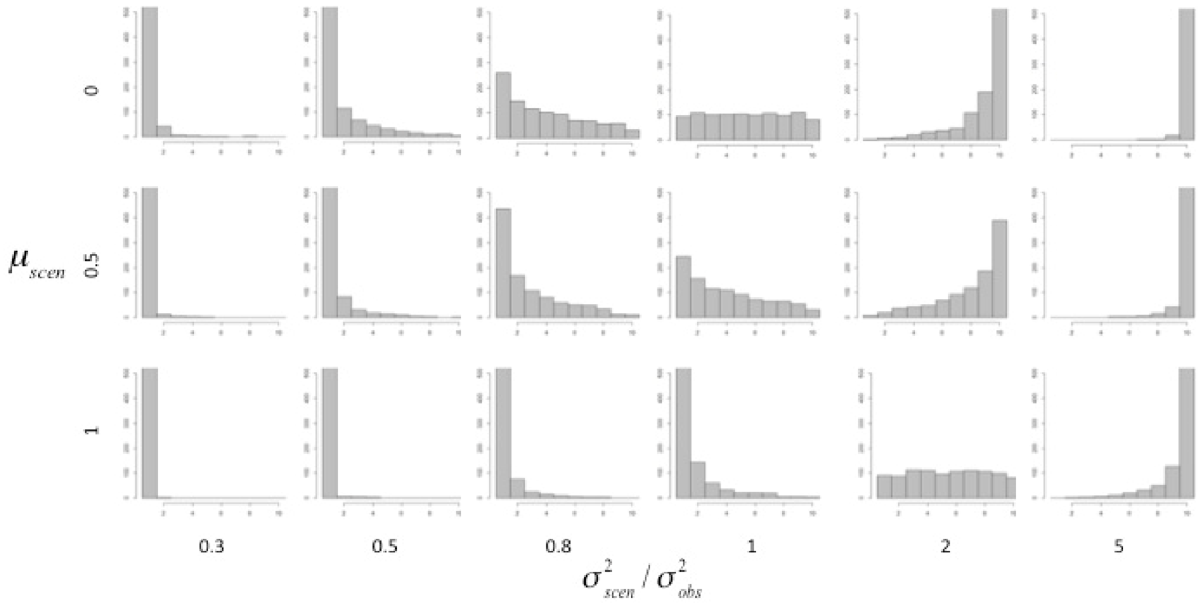


Fig. 3 Minimum spanning tree rank histogram – simulation study. In this simulation, for each panel, 1,000 ensembles each consisting of one observation and 9 equally likely scenarios, which are vectors of length 24, are sampled. The observation is sampled from a standard normal distribution with mean $\mu_{obs} = 0$ and $\sigma^2_{obs} = 1$. The scenarios are sampled from a normal

distribution with mean $\mu_{scen} = 0,\ 0.5, \text{or } 1$. The rows correspond to μ_{scen}, and the columns correspond to the ratio of the scenario variance to observation variance, $\sigma_{scen}^2 / \sigma_{obs}^2$.

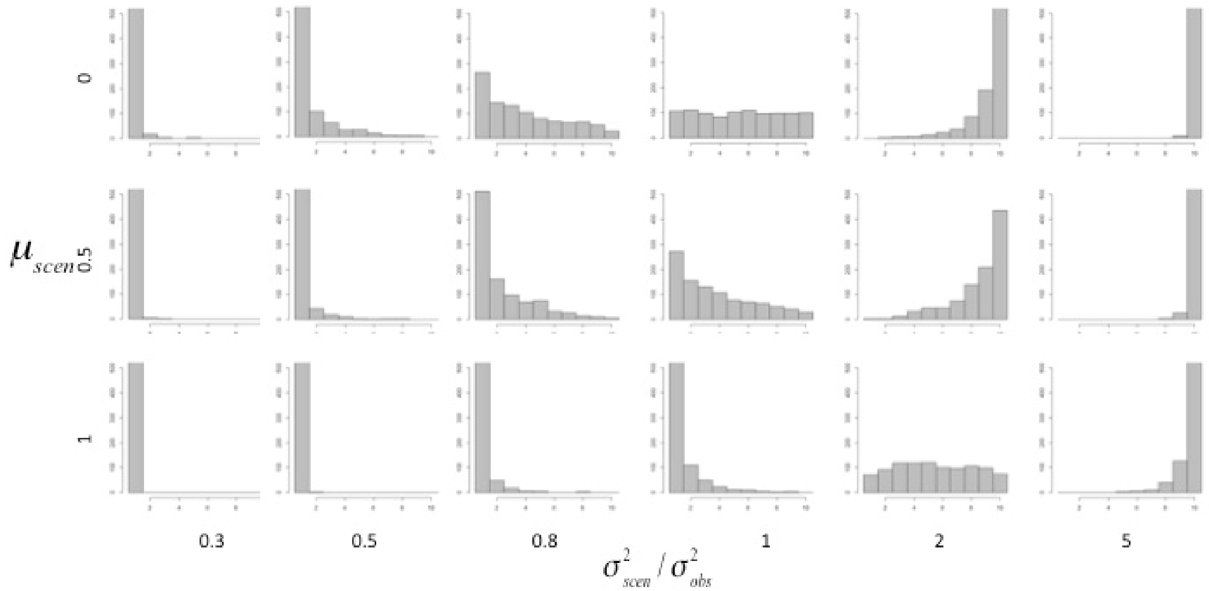


Fig. 4 Mass transportation distance rank histogram – simulation study with the same setup as in Fig. 3.

In Fig. 5 we show the importance of de-biasing for the MTD rank histogram as well as the MST rank histogram. In the left-hand panels, both histograms slope downward because of high bias, even though the scenarios are over-dispersed. To prevent misdiagnosis, Wilks suggested to de-bias the data when constructing MST rank histograms [12]. In the right-hand panels, the data are de-biased according to the following equation:

$$y_{h,d}^{s\circ} = y_{h,d}^{s*} - \frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{S}\sum_{s=1}^{S} y_{h,d}^{s*} - y_{h,d}^{0*}\right), \text{ for } h = 1,\dots,H .$$

The resulting MTD rank histograms appear very similar to the MST rank histograms both before and after de-biasing.

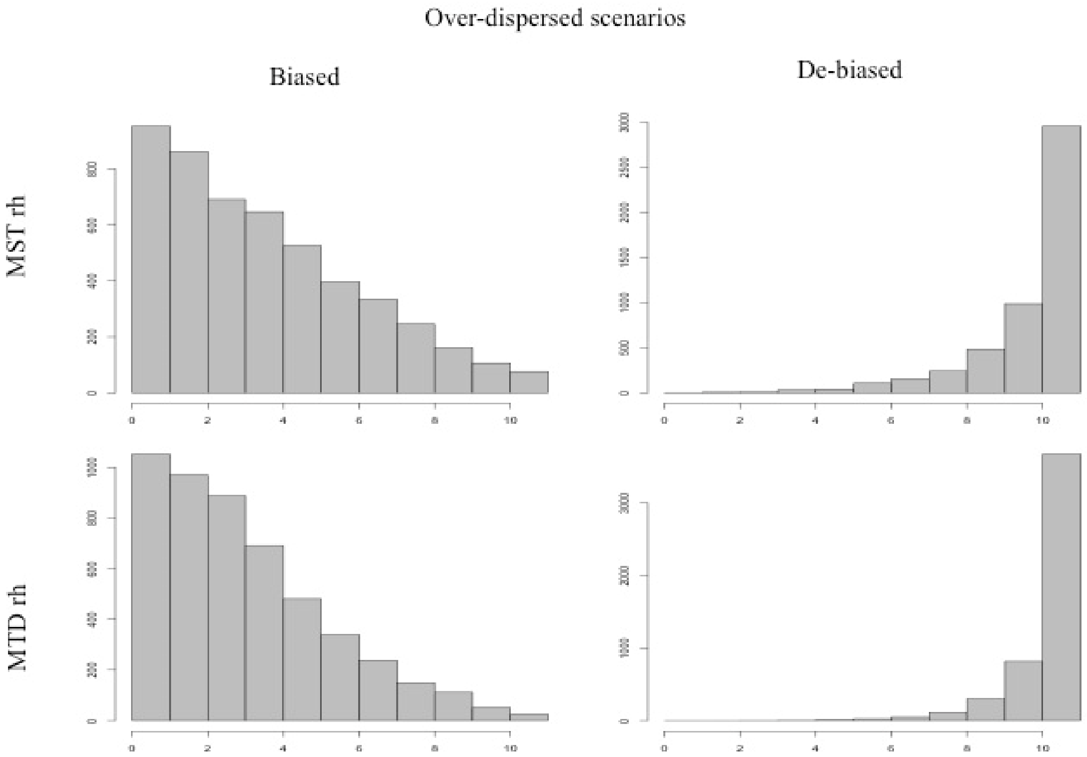


Fig. 5 MTD and MST rank histograms for over-dispersed scenarios with and without bias - For each panel, 5,000 ensembles each consisting of one observation and 10 scenarios, which are vectors of length 8, are sampled. The observation is sampled from a standard normal distribution with mean $\mu_{obs} = 0$ and $\sigma^2_{obs} = 1$. Scenarios are sampled from a normal distribution with mean $\mu_{scen} = 2.5$ and $\sigma^2_{scen} = 5$.

In the context of wind power, we are particularly interested in assessing whether the autocorrelation, as a way of describing temporal smoothness, of scenarios matches that of observations. In [23] the authors investigate the sensitivity of four different multivariate ranking methods, including minimum spanning tree rank histogram, to miscalibration in the dependence structure. They generate their forecasts from an AR(1) process, whereas their observations follow more complex correlation models. Simulation studies presented in Fig. 6 and 7 examine the behaviors of both rank histograms according to autocorrelation. To equalize variances of the marginal distributions the data are scaled according to the Mahalanobis transformation [12].

The Mahalanobis transformation scales the data according to the sample covariance matrix:

$$\mathrm{S}_{scen} = \frac{1}{S}\left[\left(y_d^{0*} - \overline{y}_d^{scen}\right)\left(y_d^{0*} - \overline{y}_d^{scen}\right)^{\mathrm{T}} + \sum_{s=1}^{S}\left(y_d^{s*} - \overline{y}_d^{scen}\right)\left(y_d^{s*} - \overline{y}_d^{scen}\right)^{\mathrm{T}}\right],$$

where

$$\overline{y}_d^{scen} = \frac{1}{S+1}\left(y_d^{0*} + \sum_{s=1}^{S} y_d^{s*}\right)$$

The transformation is a multi-dimensional extension of standardization by subtracting the mean and dividing by the standard deviation:

$$z_d^0 = \mathrm{S}_{scen}^{-1/2}\left(y_d^{0*} - \overline{y}_d^{scen}\right),$$

$$z_d^s = \mathrm{S}_{scen}^{-1/2}\left(y_d^{s*} - \overline{y}_d^{scen}\right)$$

where $\mathrm{S}_{scen}^{-1/2} = D\Lambda^{-1/2}D^{\mathrm{T}}$, D is the matrix whose columns are the eigenvectors of S_{scen}, and $\Lambda^{-1/2}$ is the diagonal matrix containing the reciprocals of the square roots of the corresponding eigenvalues [12].

The MST and the MTD rank histograms behave similarly as the marginal variance and the autocorrelation parameter are varied. For over-dispersed scenarios, as the observation autocorrelation decreases, the histogram becomes flatter; however, an upward trend can still be observed. For under-dispersed scenarios, a downward trend is observed for all levels of autocorrelation levels of the observation but it is less pronounced when the observation autocorrelation is high. If the scenarios and observation have the same autocorrelation and marginal variance, the MST and MTD rank histograms both appear to be flat, as we observe in the middle panels of Figs. 6 and 7. When the marginal variances of scenarios and observation are the same, the difference between autocorrelations will affect the pattern of both rank histograms. For $\rho_{obs} < \rho_{scen}$, a sloping downward trend and for $\rho_{obs} > \rho_{scen}$ a sloping upward trend are observed in Figs. 6 and 7.

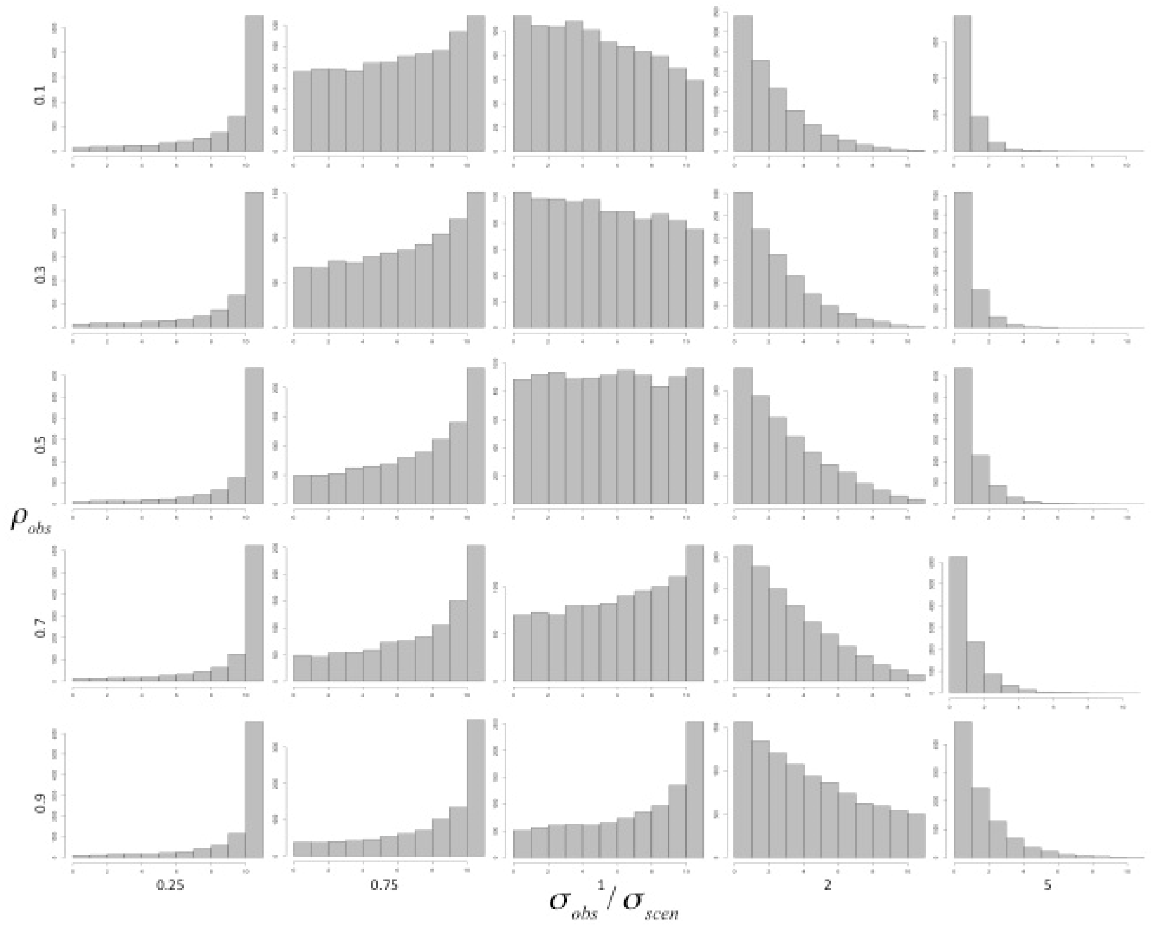


Fig. 6 Minimum spanning tree rank histogram – simulation study for testing scenarios according to their autocorrelations. In this simulation, for each panel, 10,000 ensembles each consisting of one observation and 10 scenarios, which are vectors of length 8, are sampled. The scenarios are sampled from an AR(1) model, defined as $X_k = \rho X_{k-1} + \varepsilon_k$ with coefficient $\rho_{scen} = 0.5$. The standard deviation of the marginal distribution of the scenarios is maintained as $\sigma_{scen} = 1$ by adjusting the standard deviation of ε_k . The rows correspond to the ρ coefficients of the observation which is also sampled from the AR(1) model. The columns correspond to the ratios (observation to scenarios) of the standard deviations of the marginal distributions.

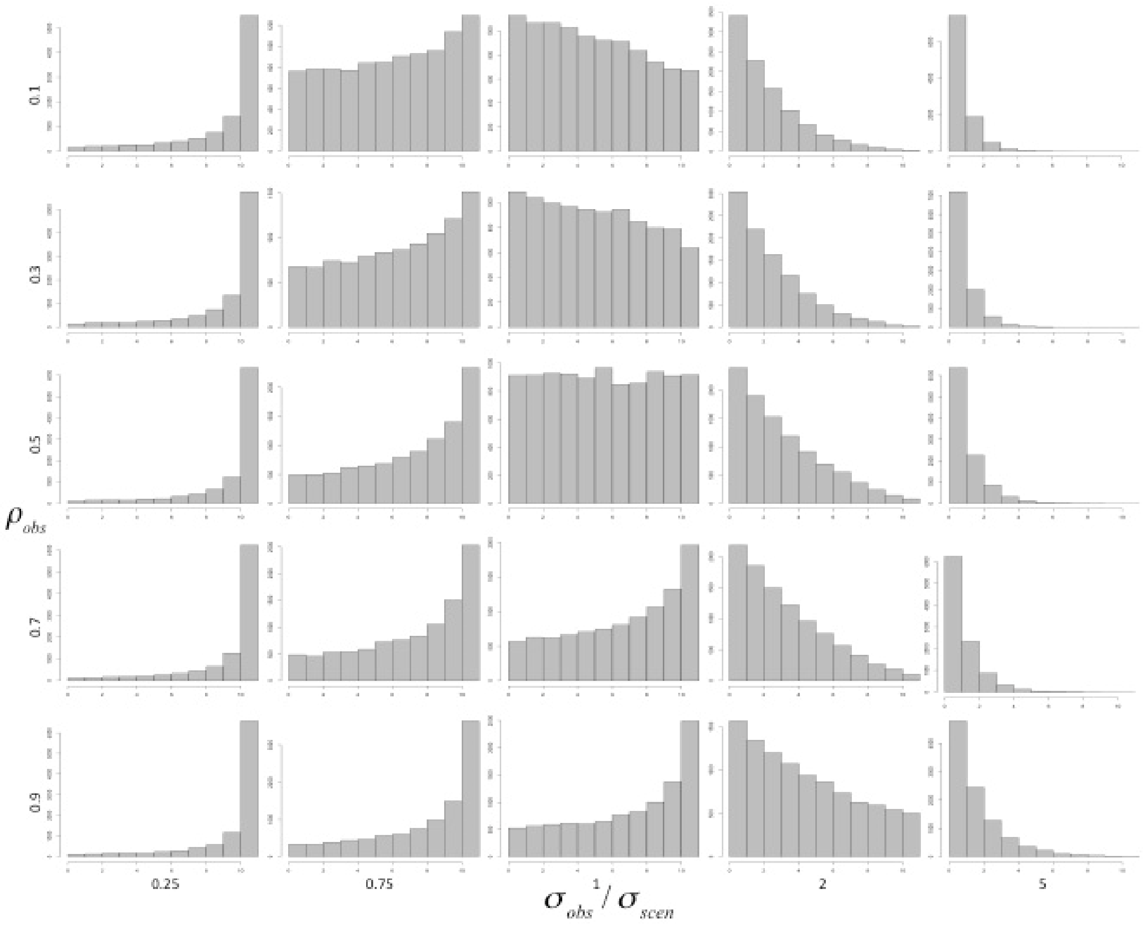


Fig. 7 Mass transportation distance rank histogram – simulation study for testing scenarios according to their autocorrelations with the same setup as in Fig. 6.

Fig. 8 further illustrates the patterns of the MTD rank histogram for the case where variances of the marginal distributions of scenarios and observation are equal. The MST rank histograms are flat when the autocorrelations of observation and scenarios are equal. Above the main diagonal where $\rho_{obs} > \rho_{scen}$, they show an upward-sloping trend, which increases with the difference between the autocorrelation levels. When $\rho_{scen} = 0.1$ and $\rho_{obs} = 0.5$, a U-shaped rank histogram is observed. As both ρ_{obs} and ρ_{scen} are increased, an upward-sloping trend appears. For the case where $\rho_{obs} < \rho_{scen}$, below the diagonal, the rank histograms always slope downward but they are flatter when the difference between autocorrelation coefficients of scenarios and observation is smaller.

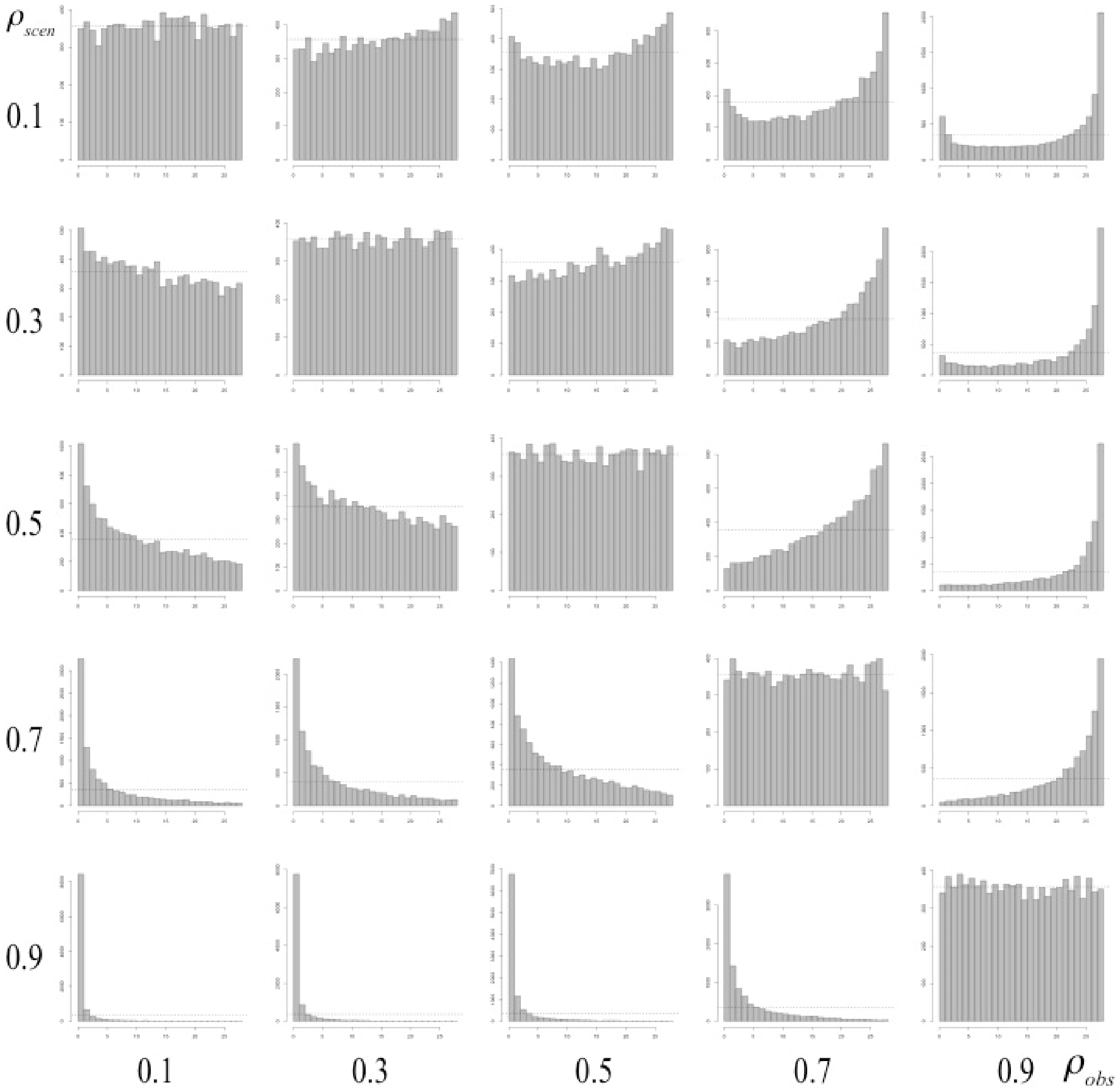


Fig. 8 MTD rank histograms for $\sigma_{obs}/\sigma_{scen}=1$ and various combinations of ρ_{obs} and ρ_{scen}.

If we generate scenarios with heterogeneous autocorrelation levels, we observe a hill-shaped MTD rank histogram as in Fig. 9. This occurs because the presence of both much more and much less smooth scenarios than the observation makes the range of mass transportation distances among scenarios larger. The MTD from the scenarios to the observation will fall in the middle frequently. Overpopulation of the middle ranks results in a hill-shaped MTD rank histogram that is skewed according to the proportions of scenarios with high and low autocorrelation.

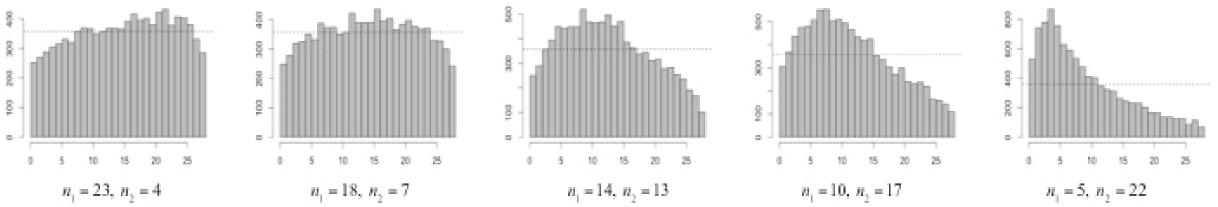


Fig. 9 MTD rank histograms when $\sigma_{obs}/\sigma_{scen} = 1$ and scenarios with both $\rho_{scen} > \rho_{obs}$ and $\rho_{scen} < \rho_{obs}$ are present. n_1 = the number of scenarios that have AR(1) coefficient $\rho = .1$, n_2 = the number of scenarios that have AR(1) coefficient $\rho = .8$. The observation has an AR(1) coefficient $\rho = .5$.

Certain combinations of over-dispersion and weak correlation can result in a deceptively flat histogram. This is a limitation of both MTD and MST rank histograms. For example, Fig. 10 shows relatively flat MTD and MST rank histograms that result from the same setup as in Figs. 6-9 when $\sigma^2_{scen} / \sigma^2_{obs} = 1.5$ and $\rho_{scen} = 0.5,\ \rho_{obs} = 0.1$. The rank histograms could cause the scenarios to be misinterpreted as reliable despite their over-dispersion and higher autocorrelation.

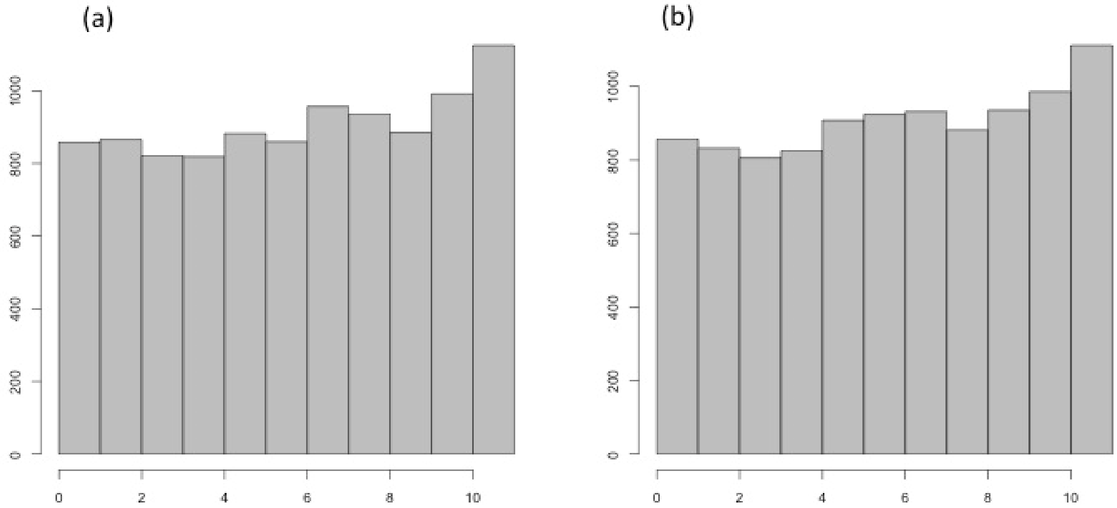


Fig. 10 (a) MTD and (b) MST rank histograms when $\sigma^2_{scen} / \sigma^2_{obs} = 1.5$ and $\rho_{scen} = 0.5,\ \rho_{obs} = 0.1$.

In summary, the shape of the MTD rank histogram closely corresponds to that of the MST rank histogram when applied to equally likely scenarios. It can also be used to diagnose higher, lower, and mixed levels of autocorrelation in the scenarios compared to the observation. To verify its use with unequally likely scenarios, we repeated the same study of the MTD rank histogram as in Fig. 7 with the added step of randomly (without replacement) assigning a probability drawn from the set $\left\{2i/\left(S\left(S+1\right)\right), i = 1,...,S\right\}$ to each of the S scenarios generated. The MTD rank histograms showed the same patterns with varying parameters as in Fig. 7.

2.3 Event-based verification

Event-based verification can be used to explore the scenarios' ability to represent some specific characteristics of stochastic processes as done in [9]. For this verification type, first, it should be determined which stochastic process characteristics are critical to capture. The events can then be defined to detect these critical characteristics.

For instance, a significant gradient event is defined in [9], which is the "maximum absolute variation being greater than a determined threshold in a determined finite duration beginning at a time point". The event parameters are the threshold ξ, and the duration κ. By changing the parameters ξ and κ, different specific events can be defined. Similarly to the significant gradient event, we define ramp up and ramp down events as the "maximum increase and maximum decrease being greater than or equal to ξ, in κ hours beginning at time point *h* respectively". For wind power scenarios, we are particularly interested in ramp down events because an unexpected loss of a significant amount of wind power could trigger the need for expensive peaking generators to be brought into service. In Section 4, we tested wind power scenarios according to both ramp down and ramp up events.

An indicator variable, denoted as $1\{.\}$, takes value 1 if the event occurs or 0 otherwise. Ramp events are defined as follows for a given time series:

$$\text{RampUp}(y_d;h,\kappa,\xi)=1\left\{\exists\ i\in\{0,1,...,\kappa-1\}\quad \text{s.t.}\ \ y_{(h+\kappa),d}-y_{(h+i),d}\geq\xi\right\}$$

$$\text{RampDown}(y_d;h,\kappa,\xi)=1\left\{\exists\ i\in\{0,1,...,\kappa-1\}\quad \text{s.t.}\ \ y_{(h+i),d}-y_{(h+\kappa),d}\geq\xi\right\}$$

Denoting the parameter set as $\theta=(h,\kappa,\xi)$, $\text{RampUp}(y_d^0;\theta)$ and $\text{RampDown}(y_d^0;\theta)$ define the ramp up and ramp down events for observed time series on day *d* beginning at time *h* within a time window of length κ. For the scenarios, the event probabilities can be defined mathematically as:

$$\text{P}_{h,d}[\text{RampUp}(y_d^s;\theta)]=\sum_{s=1}^{S}\text{RampUp}(y_d^s;\theta)p_d^s$$

$$\text{P}_{h,d}[\text{RampDown}(y_d^s;\theta)]=\sum_{s=1}^{S}\text{RampDown}(y_d^s;\theta)p_d^s$$

The probability-weighted average of indicator variables for the scenarios takes a value in the interval [0,1]. The Brier score is a strictly proper score to assess these binary situations, which depend on the occurrence and non-occurrence of the event, as applied in [9]. The Brier score is the sum of squared distances between the observation indicator and scenario average [18]. A daily Brier score can be computed as:

$$\mathrm{Bs}(d)_{daily}=\frac{1}{(H-\kappa)}\sum_{h=1}^{H-\kappa}\left(\mathrm{P}_{h,d}[\mathrm{RampDown}(y_d^s;\theta)]-\mathrm{RampDown}(y_d^0;\theta)\right)^2 \quad \text{for } d=1,\ldots,D$$

In Section 4, we also examine the frequencies of hourly Brier scores:

$$\mathrm{Bs}(h,d)_{hourly}=\left(\mathrm{P}_{h,d}[\mathrm{RampDown}(y_d^s;\theta)]-\mathrm{RampDown}(y_d^0;\theta)\right)^2 \qquad \text{for } h=1,\ldots,H-\kappa, \quad d=1,\ldots,D.$$

Brier scores measure the degree of correspondence between scenarios and observation based on the event occurrence. Brier scores are lower for scenarios that accurately reflect the event's occurrence.

3. Wind power scenario generation methods

We use the methods described above to compare the results of two distinct methods for generating scenarios of short-term wind power generation. Given a forecast time series for amounts of wind power available on the next day, the major challenges in generating scenarios (i.e., alternative time series, each with a probability attached) include modeling the marginal distributions of forecast error at each time point, considering dependencies among these marginal distributions, and building sequences of wind power values that respect the distributions and temporal dependencies. The two general approaches for wind power scenario generation considered in this paper are compared according to these aspects in Table 1. Note that the quantile regression approach yields equally likely *sample points* from the estimated joint distribution while the epi-spline approximation approach results in a collection of time series that, together with their probabilities, *approximate* the stochastic process for wind power. In the following two subsections the approaches are described in more detail to explain some of their variants considered in the numerical study.

Table 1. Overview of scenario generation approaches

Approach	Epi-spline approximation with information [19]	Quantile regression with copula [7]
Marginal distribution for each time point	Epi-spline approximation of log of error density based on historical errors within a forecast cluster	Linear interpolation of quantiles of forecast error estimated by quantile regression
Intertemporal	Conditional distributions of	Gaussian copula applied to marginal

dependence	forecast errors based on categorizations of forecast at certain time points	distributions to approximate joint distribution
Scenario construction	Conditional expected values within segments of conditional forecast distributions	Monte Carlo samples from joint distribution of forecast errors added to given forecast

One-day wind power output scenarios were generated based on day-ahead wind power forecast data. We followed a “leave-one-out” methodology when generating short term wind power scenarios by both methods. For scenarios generated on day *d*-1 for day *d*, the training set consisted of the whole data range except day *d*, whereas the test day was day *d*.

3.1 Wind power scenario generation by quantile regression with Gaussian copula approach

The actual wind power generated at hour *h* on day *d*, $y_{h,d}$, can be observed immediately at the end of hour *h* on day *d*. On day *d*-1 we obtain a vector of day-ahead wind power forecasts (DWPF)

$$\hat{y}_d = \left(\hat{y}_{1,d}, \hat{y}_{2,d}, \dots, \hat{y}_{24,d}\right).$$

Thus, a day-ahead wind power forecast error (DWPFE) can be observed at the end of each hour *h* on day *d*:

$$e_d = \left(e_{1,d}, e_{2,d}, \dots, e_{24,d}\right),$$

where

$$e_{h,d} = y_{h,d} - \hat{y}_{h,d}.$$

In this method, actual wind power output, DWPF, and DWPFE are all assumed to be normalized by wind power capacity and denoted as $y_d^{0*}, \hat{y}_d^*, e_d^*$, respectively, so that $\left(y_{h,d}^{0*}, \hat{y}_{h,d}^*\right) \in \left[0,1\right]^2$ and $e_{h,d}^* \in \left[0-\hat{y}_{h,d}^*, 1-\hat{y}_{h,d}^*\right]$. On day *d*, after DWPF $\hat{y}_d^*$ is obtained, we estimate a distribution of DWPFE

$$F_{h,d}\left(e|\hat{y}_d^*\right) = P\left(e_{h,d}^* \le e \mid \hat{y}_d^*\right)$$

for each hour *h* by linearly interpolating a predicted τ - quantile of $e_{h,d}^*$ for each τ in a pre-defined set of quantiles $\mathbb{T}$ (e.g. $\{.05, .10, \dots, .95\}$). It is assumed that the 0.00 quantile of the predictive forecast is 0 and the distribution below the 0.05 quantile is modeled as a linear interpolation between the 0 and 0.05 quantiles. Similarly, the 1.00 quantile of the predictive forecast is assumed to be 1 and the distribution above the 0.95 quantile is linearly interpolated. These assumptions may lead to extreme scenarios with unrealistically large differences from the forecast. For better results, the predictive distributions should be parameterized with exponential tails, thus reflecting the unlikeliness of extreme events [7]. Each quantile of $e^{*}_{h,d}$ is predicted by using quantile regression models on

$$O_d = \left\{\left(\hat{y}^{*}_{1}, e^{*}_{1}\right), ..., \left(\hat{y}^{*}_{D}, e^{*}_{D}\right)\right\} \setminus \left\{\left(\hat{y}^{*}_{d}, e^{*}_{d}\right)\right\}.$$

The development described above is elaborated in [7]. Here, we introduce two variants on constructing predictor variables for quantile regression models. First, we conduct dimension reduction to improve the reliability of the regression models. The original DWPF is highly inter-correlated 24-dimensional data. We define the following five models of transformed DWPF:

- model 1: A single forecast datum for the particular study period *h*: $\hat{y}^{*}_{h,d}$
- model 2: model 1 + forecasts for an hour before and after. This may outperform model1 if there is inaccurate time prediction called phase error [15].
- model 3: Principal components that take account of the major proportion of variances in the DWPF data matrix. In this study, four components explain over 99% of the total variance-covariance in the training data matrix.
- model 4: model 3 + principal components that take account of the major proportion of variances in local differences within DWPF:

  $$(\hat{y}^{*}_{2,d} - \hat{y}^{*}_{1,d}, \hat{y}^{*}_{3,d} - \hat{y}^{*}_{2,d}, \ldots, \hat{y}^{*}_{24,d} - \hat{y}^{*}_{23,d}).$$

  In this study, five components explain over 90% of total variance-covariance in the local difference data matrix.

- model 5: model 4 on an extended DWPF that includes forecasts for two hours before and after the forecast day:

  $$\hat{y}^{*}_{t} = (\hat{y}^{*}_{(-2),d}, \hat{y}^{*}_{(-1),d}, \ldots, \hat{y}^{*}_{24,d}, \hat{y}^{*}_{(+1),d}, \hat{y}^{*}_{(+2),d}$$

The second variant is to use a spline function to incorporate a possible nonlinear relation between a quantile of forecast error and DWPF. The number of the degrees of freedom (DF) represents the number of basis functions of each regressor, which implies the complexity of nonlinearity between each regressor and the forecast error. We applied natural cubic spline with up to 3 basis functions (DF=1,2,3). By combining five dimension reduction models and three spline functions, we construct 15 different DWPFE distributions for each *h*. A linear interpolation of estimated quantiles may need some exception handling to make sure that $F_{h,d}(.)$ is monotonically increasing and the range of DWPFE is realistic.

After estimating $F_{h,d}(.)$ for each h and d we transform the training forecast error $e^{*}_{h,d}$ into normally distributed random variables

$$\mathrm{z}_{h,d} = \Phi^{-1}\left(\hat{F}_{h,d}\left(e^{*}_{h,d} \mid \hat{y}^{*}_{d}\right)\right),$$

where Φ is the cdf of a standard normal distribution.

Let

$$Z(d) = \begin{bmatrix} z_{1,1} & z_{1,1} & \cdots & z_{24,1} \\ z_{1,2} & z_{2,2} & \cdots & z_{24,2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1,d-1} & z_{2,d-1} & \cdots & z_{24,d-1} \\ z_{1,d+1} & z_{2,d+1} & \cdots & z_{24,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1,D} & z_{2,D} & \cdots & z_{24,D} \end{bmatrix} = \begin{bmatrix} z_{1} \\ z_{2} \\ \vdots \\ z_{d-1} \\ z_{d+1} \\ \vdots \\ z_{D} \end{bmatrix}$$

We generate a scenario of transformed DWPFE by

$$z^{s}_{d} \sim N\left(\mu_0, \hat{\Sigma}_d\right)$$

where μ_0 is a 24-dimensional zero vector and the variance-covariance matrix

$$\hat{\Sigma}_d = \frac{1}{D-1} Z(d)^{\mathrm{T}} Z(d).$$

Next, we generate a DWPFE scenario e^{s*}_{d} as $e^{s*}_{d} = F^{-1}_{h,d}\left(z^{s}_{d}\right)$. Finally a scenario of day-ahead wind power output is computed by adding the scenario of DWPFE to DWPF as

$$y^{s*}_{d} = \hat{y}^{*}_{d} + e^{s*}_{d}.$$

Each scenario is assumed to occur with probability

$$P_s = \frac{1}{S}, \forall s \in \{1, \cdot\cdot\cdot, S\}.$$

3.2 Wind power scenario generation by epi-spline approximation approach

The goal of this functional approximation approach is to sparingly approximate, rather than sample from, the error distributions, while incorporating available information. The three main steps are segmentation, forecast error distribution estimation, and path construction.

The process begins with the DWPF, $\hat{y}_d = \left(\hat{y}_{1,d}, \hat{y}_{2,d}, ..., \hat{y}_{24,d}\right)$, and the observed wind power, $y_d = (y_{1,d}, y_{2,d}, ..., y_{24,d})$, for each day d in the training set. Our goal is to characterize the distributions of the forecast performance so that in the future when given a forecast, we can produce probabilistic statements about possible observations.

Using exponential epi-splines [24], for each hour h an error density function is approximated from the forecast errors $\left\{e_{h,d}, d = 1, ..., D\right\}$, and then numerically integrated to obtain an hourly error cdf.

A key concept is *segmentation* of the data so that data from similar conditions are grouped together. The simplest form of segmentation is to group similar amounts of wind power together, as was suggested in [25,26] who cite [27] as the original work.

In operation, we are given a 24-hour wind forecast and asked to provide a distribution of the forecast for certain hours in that forecast. To do that, we find a distribution for *similar hours* in the historic data, which means we use data from the same data segment as the forecast. We segment wind based on two main attributes: the magnitude of the power forecast and the *derivative pattern*.

The magnitude is taken into account by using only those historic hours with a forecast value within a *window*. The width of the window is controlled by a parameter (typically 0.4) that gives the fraction of the distribution centered at the forecast to include in the window; i.e,, approximately the fraction of the observations to include.

The derivative pattern is a bit more involved. For each hour we compute the derivative one hour before, one hour after, and at the hour. Each derivative is classified as small, substantially negative, or substantially positive with the meaning of "substantially" controlled by a parameter. Hence, for any hour there are nine possible patterns. Because some patterns do not have very much historic data, we cluster the patterns using a metric in the space of error distributions to control the clustering. Patterns with similar error distributions are put in the same cluster, which is done in a pre-processing step along with assignment of each historic hour to a cluster. Then, when an error distribution for some hour in the forecast is requested, the derivative pattern for the forecast is determined and only those hours that are in the same cluster and within the magnitude window are used to compute the error distribution.

Paths are constructed in a fashion based on Rios et al. [19]. One difference is that for wind there is no re-forecasting and segmentation is done as described above, not by error category as in [19]. A few hours are selected by the analyst to be *day part separators* (dps). The probability computations assume that the hours are far enough apart so that correlations in forecast errors between them can be ignored, which has been verified for the studies conducted so far.

At each dps, cutting points on the error distribution are computed and used to compute *skeleton points*, each of which is an expected value conditional on being between a pair of cutting points. The difference in the cutting points is the probability assigned to the skeleton points. A list of skeleton points with one for each dps is called a *skeleton* and, under the assumption that all skeletons are enumerated and that the errors between dps are uncorrelated, the probability attached to a skeleton is simply the product of the probabilities attached to its skeleton points. To connect the points and provide values for the hours between dps the deviation from the forecast is linearly interpolated. This process completes specification of a scenario with serial dependence between the hours based on the forecast dependence and an attached probability.

The number of scenarios generated equals the number of dps multiplied by one less than the number of cutting probabilities. For example, in the numerical study dps consist of hours 0, 12, and 23, to divide the day into two intervals, and 4 probability values including 0 and 1 are used to cut the error distributions into segments. If cutting probabilities (0, 0.1, 0.9, 1) are used then skeleton points computed have probabilities approximately equal to 0.1, 0.8, and 0.1, corresponding to the lower tail, middle, and upper tail of the distribution. (These probabilities are not exact because the conditional distributions are discretized based on the historical data.) Therefore, the probabilities associated with the paths approximately equal $(0.1)^3$ for a scenario inhabiting the tails at all dps, $(0.1)^2(0.8)$ for a path through tails at two dps and the middle of distribution at one dps, $0.1(0.8)^2$ for a path in the tail at one dps and the middle at two dps's, and $(0.8)^3$ for the path that represents the middle of the error distribution at each dps.

4. Example application of the verification approaches

We used day-ahead forecast and observational data from the Bonneville Power Administration in a recent year to generate scenarios by both methods. Scenarios were generated by the quantile regression with Gaussian copula approach according to 5 transformation functions (model 1 through 5) and 3 levels of nonlinearity (DF 1 to 3) for the natural B-spline function as explained in Section 3.1. This resulted in 15 quantile regression models, labeled below as QR(*m*,*n*), for model *m* with *n* DF. The epi-spline approximation method was used with two different sets of cutting points $\left(0, p_1, p_2, 1\right)$ and the resulting scenario sets labeled as EPI(p_1, p_2). For each day, 27 scenarios were generated by each method.

4.1 The BPA dataset

Bonneville Power Administration (BPA) is a federal non-profit agency based in the Pacific Northwest of the U.S. that markets wholesale electric power. BPA works with wind project owners to develop more accurate long-term and short-term wind forecasts. More information on wind power forecasting methodology by BPA can be found from [28]. Our focus is given to data from 2012-10-01 to 2013-09-30, based on private communication with BPA.

Forecast data were obtained from [29], item number 3; and [30], which provides monthly spreadsheets for “BPA wind power forecasting data”. Each month includes three files; maximum, minimum, and average. We used the “AVG BPA Wind Power Forecast” file, which includes the expected average generation over the hour. Forecasts are identified by UTC stamp and extend over 72 hours. Hr01 represents the first hour of the forecast or next hour. We extracted the forecast generated at 11 a.m. Pacific time, which is 18:00 Greenwich mean time (UTC) during daylight savings time and 19:00 UTC during normal time, on day *d*-1, for the 24 hours of day *d* in columns labeled as Hr13…Hr36. When generating scenarios by the quantile regression method according to model 5 explained in Section 3.1, we need 2 hours of extended forecast data. These forecast values, denoted as $\hat{y}_{(-2),d}, \hat{y}_{(-1),d}, \hat{y}_{(+1),d}, \hat{y}_{(+1),d}$, were obtained from columns Hr11, Hr12, Hr37 and Hr38, respectively.

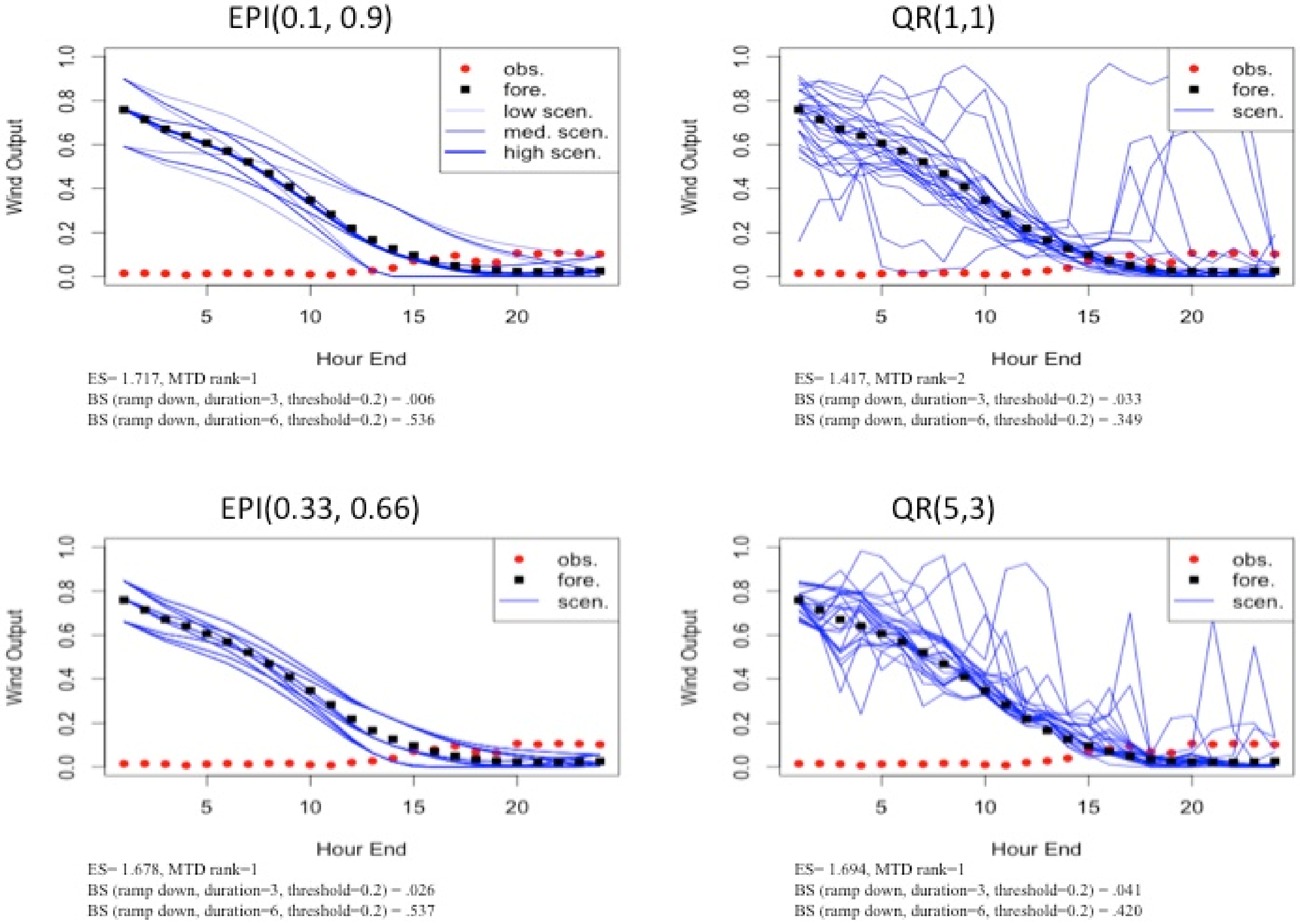
For the observations, we used the total wind generation value in the first 5-minute interval from item number 5 on [29]. When normalizing the observation and forecast data, we used wind generation capacity available from [31].

A few days within this date range were ignored because of missing information or noisy data. The date when daylight savings time began (2013-03-10) was omitted. Moreover, a few days were omitted as abnormal because they were labeled by BPA as abnormal; specifically, we omitted days that had 4 hourly forecasts with wind states greater than or equal to 2 or less than or equal to -2. The wind states can be found from item 12, “Data for BPA balancing reserves deployed and BPA states” of [29]. In addition to these, days with missing information in either forecast or observation were not included in the scenario generation data. In this one-year period there were a total of 22 disregarded days, leaving a leave-one-out training window length of *D*=343 days.

4.2 Verification of BPA scenarios

In Fig. 12-14, we show the observed wind power and the scenarios generated by different variations of the two approaches on selected days along with those days’ Energy scores (ES), MTD ranks and Brier scores. Fig. 12 illustrates the effects of a bad forecast, and Fig. 13 shows the results for a day when the wind output and forecast are both very low. These are the most extreme days in our dataset. We distinguish the scenarios generated by the EPI(0.1, 0.9) according to their approximate probabilities and label them as high, medium, or low (probability). Because the probabilities for scenarios generated with EPI(0.33, 0.66) are very similar, we did not distinguish their probabilities in the plots. As mentioned in Section 3, the quantile regression method generates equally likely scenarios. On 2012-11-08, as shown in Fig. 12, the scenarios are very far from the observation, which results in very high ES for all of the scenario sets. This is evidently because of the bad wind power forecast. Although the first term in the ES is very large, the dispersion of scenarios can reduce the score. For example, the scenarios generated by the QR(1,1) are more dispersed than the other scenarios sets and have a lower ES. The observation is exterior to the scenarios due to under-dispersion and/or high bias, and this condition is detected by the low MTD ranks for all scenario sets. Compared to the other days, Brier scores are relatively high as expected. Conversely, on 2013-03-26 shown in Fig. 13, the scenarios are very close to the observation; thus, the energy scores are very low. Because the first term of the ES is small for all scenario sets, the sharper scenarios generated by EPI(0.33, 0.66) achieve a lower ES. This scenario set is very close to the observation and sharp for that particular day. Brier scores are near or equal to zero for all models. The quantile regression scenarios give higher MTD ranks than the epi-spline scenarios because their wide range causes the observation to lie in their interior. Fig. 14 represents a more typical day. The EPI (0.33, 0.66) scenarios have a low ES but do not contain the observed wind power. The quantile regression scenarios envelop the observation but exhibit much higher volatility than either the forecast or the observed wind power levels.

11-08-2012


EPI(0.1, 0.9)

QR(1,1)

EPI(0.33, 0.66)

QR(5,3)

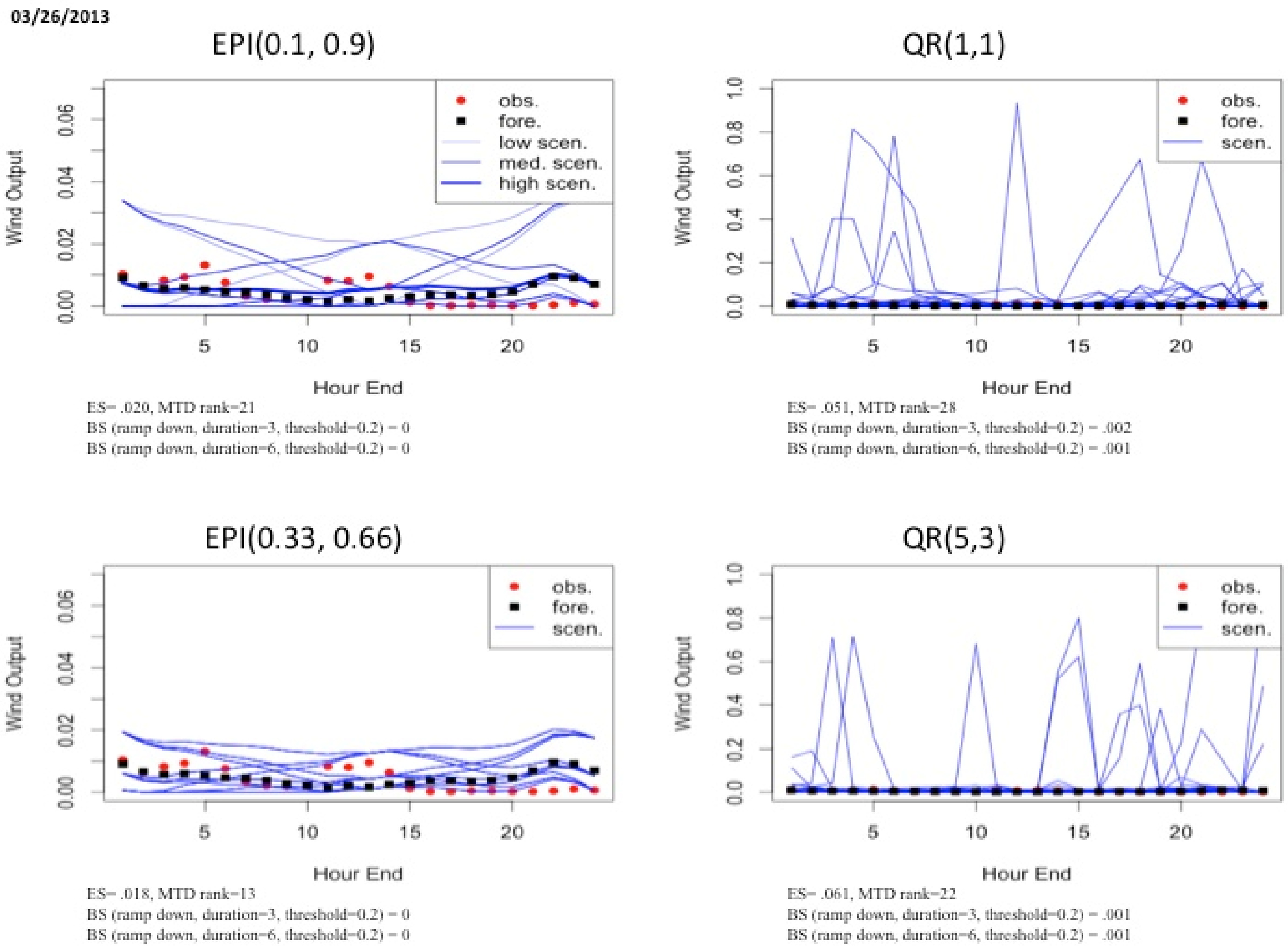
ES= 1.717, MTD rank=1
BS (ramp down, duration=3, threshold=0.2) = .006
BS (ramp down, duration=6, threshold=0.2) = .536

ES= 1.417, MTD rank=2
BS (ramp down, duration=3, threshold=0.2) = .033
BS (ramp down, duration=6, threshold=0.2) = .349

ES= 1.678, MTD rank=1
BS (ramp down, duration=3, threshold=0.2) = .026
BS (ramp down, duration=6, threshold=0.2) = .537

ES= 1.694, MTD rank=1
BS (ramp down, duration=3, threshold=0.2) = .041
BS (ramp down, duration=6, threshold=0.2) = .420

Fig. 11 Wind power scenarios generated for day 2012-11-08


03/26/2013

EPI(0.1, 0.9)

ES= .020, MTD rank=21

BS (ramp down, duration=3, threshold=0.2) = 0

BS (ramp down, duration=6, threshold=0.2) = 0

QR(1,1)

ES= .051, MTD rank=28

BS (ramp down, duration=3, threshold=0.2) = .002

BS (ramp down, duration=6, threshold=0.2) = .001

EPI(0.33, 0.66)

ES= .018, MTD rank=13

BS (ramp down, duration=3, threshold=0.2) = 0

BS (ramp down, duration=6, threshold=0.2) = 0

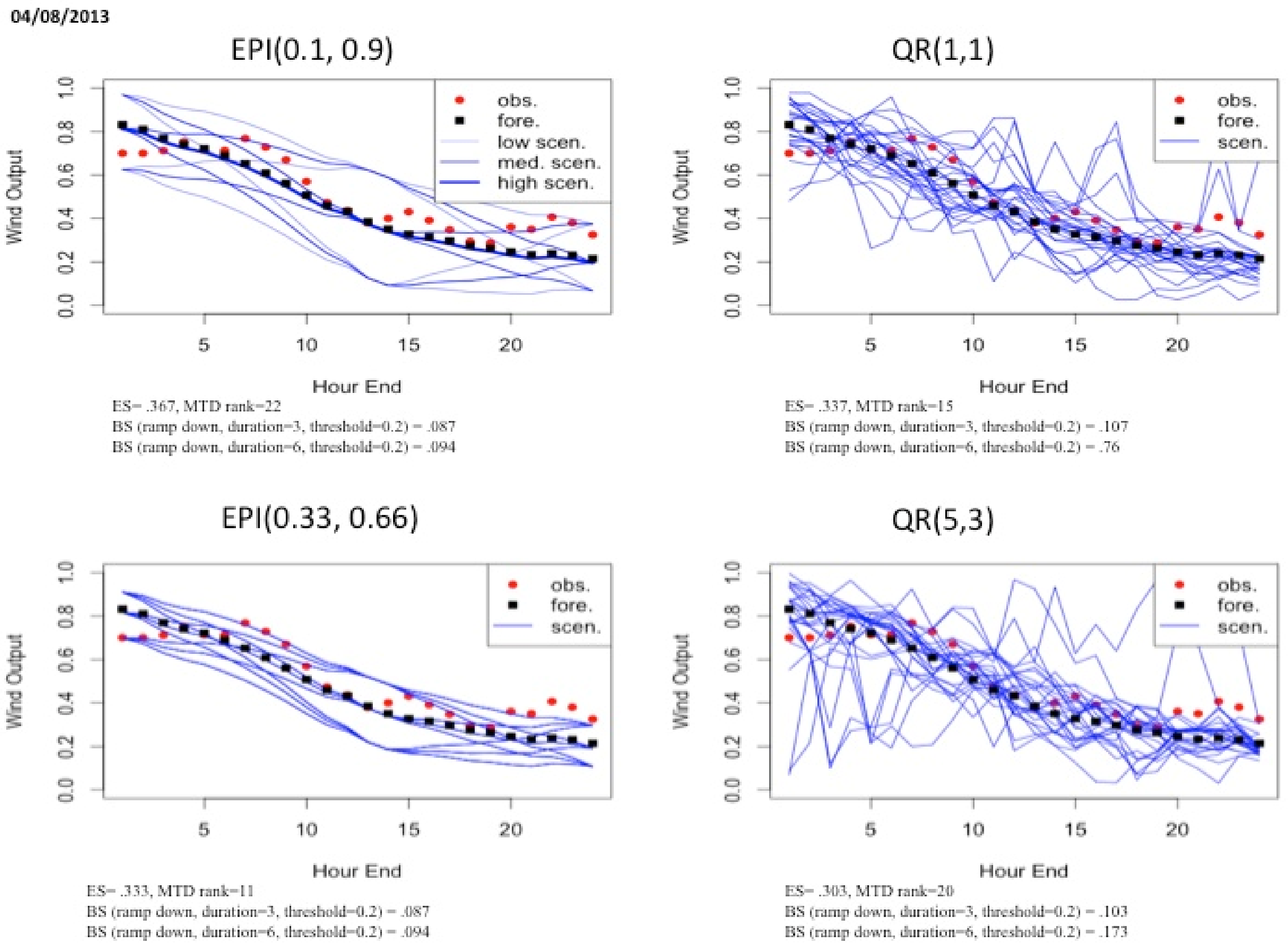
QR(5,3)

ES= .061, MTD rank=22

BS (ramp down, duration=3, threshold=0.2) = .001

BS (ramp down, duration=6, threshold=0.2) = .001

Fig. 12 Wind power scenarios generated for day 2013-03-26. Note the difference in scale between the left- and right-hand panels.



ES= .367, MTD rank=22
BS (ramp down, duration=3, threshold=0.2) = .087
BS (ramp down, duration=6, threshold=0.2) = .094

ES= .337, MTD rank=15
BS (ramp down, duration=3, threshold=0.2) = .107
BS (ramp down, duration=6, threshold=0.2) = .76

ES= .333, MTD rank=11
BS (ramp down, duration=3, threshold=0.2) = .087
BS (ramp down, duration=6, threshold=0.2) = .094

ES= .303, MTD rank=20
BS (ramp down, duration=3, threshold=0.2) = .103
BS (ramp down, duration=6, threshold=0.2) = .173

Fig. 13 Wind power scenarios generated for day 2013-04-08

The means and standard deviations of energy scores for all scenario sets are provided in Tables 2 and 3. In general, they are very similar and fairly low. Because the differences in score means between the different methods are small compared to the standard deviations, it is hard to draw any conclusion. We applied the pair-wise statistical tests for equal performance [32] to see if there is a significant difference among the scenario sets. According to paired t-tests there were only a few significant differences. The QR(4,3) scenarios had a higher mean ES than all of the other quantile regression scenario sets. Also, EPI(0.1, 0.9) had higher mean ES than EPI(0.33, 0.66) and all of the quantile regression scenario sets except QR(4,2), QR(4,3), and QR(5,3). The mean ES of the EPI(0.33, 0.66) scenarios was lower in the pairwise comparison than any of the other scenario sets. Thus, according to the energy score, EPI(0.33, 0.66) has the most skill.

Table 2

Means and standard deviations of energy scores for scenarios generated by quantile regression with Gaussian copula approach with various combinations of models and nonlinearity

	Model 1	Model 2	Model 3	Model 4	Model 5
DF=1	0.304528 (.180)	0.304323 (.179)	0.303317 (.181)	.303545 (.184)	.305380 (.182)
DF=2	0.303961 (.179)	0.304481 (.185)	0.307880 (.183)	.309423 (.186)	.307906 (.185)
DF=3	0.303177 (.184)	0.305345 (.182)	0.307104 (.184)	.316676 (.188)	.309297 (.182)

Table 3

Means and standard deviations of energy scores for epi-spline approximation approach scenarios with different cutting probabilities

	ES (std. dev. of ES)
Shape: (0-0.1-0.9-1)	0.31621 (0.209)
Shape: (0-0.33-0.66-1)	0.29373 (0.195)

We conjecture that for problems such as stochastic unit commitment, the reliability of scenarios is more important than their sharpness. If the actual wind power level exceeds the highest level among scenarios, an opportunity cost would be incurred by having committed too many thermal generators on the day ahead. Actual wind power falling below all the scenarios would likely necessitate the dispatch of expensive peaking units. Thus, it could be risky to only depend on the ES, which encompasses both reliability and sharpness. A low ES that is obtained due to sharp scenarios (which are not perfectly reliable) could misleadingly indicate high quality of scenarios, which do not actually represent the uncertainty properly. The MTD rank histogram better identifies whether the scenarios having lower ES are well calibrated.

MTD rank histograms after de-biasing and scaling are displayed in Fig. 14. The Mahalanobis transformation scales the data according to the sample covariance. The sample covariance matrices computed for the epi-spline scenarios were not actually positive definite as shown by slightly negative eigenvalues. To remedy this and allow the computation of the required square roots, we employed a method to find the nearest positive definite covariance matrix [33]. The Cramèr-von Mises statistics for all four rank histograms indicate rejection of a hypothesis of uniformity at the 1% significance level. However, the statistic for the rank histogram in (d) is very close to the critical value of 0.33. The smallest rank is overpopulated in MTD rank histogram (b) for scenarios generated by EPI(0.33, 0.66), which means the temporal dependence structure in the observation is overestimated. The one in (a) generated by EPI(0.1, 0.9) is relatively flat. In (c) we can observe a hill shape, which indicates that scenarios with both lower and higher levels of autocorrelation than the observation are generated by QR(1,1). The histogram generated by QR(4,3) in (d) is flatter than (c) and, overall, indicates the best match of autocorrelation and variance levels between scenarios and observations. The hill shapes observed in (c) and (d) might be attributed to the linear interpolation of the marginal tails, which caused excursion of scenarios far from the pack. This increased the number of scenarios with lower autocorrelation than the observation. With a better modeling of the tails by finding a suitable parameterization, we could obtain more realistic scenarios and expect flatter MTD rank histograms.

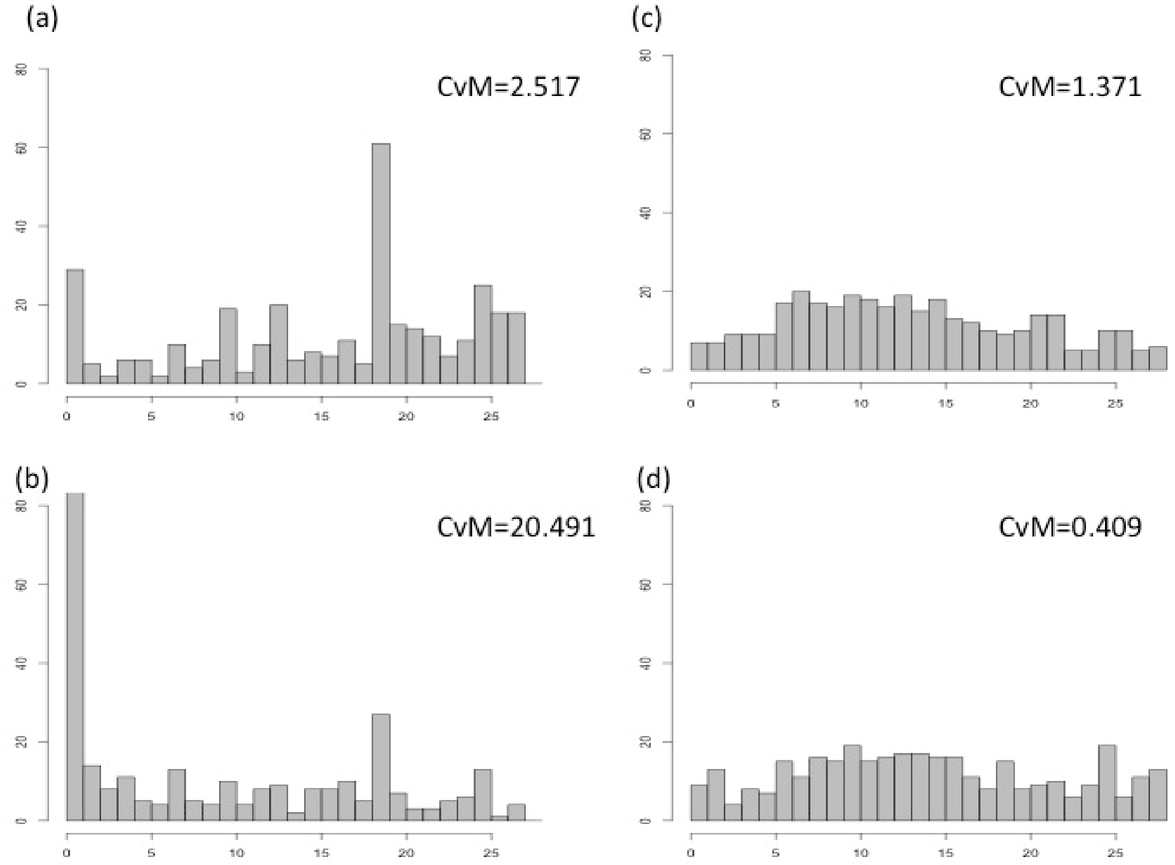


Fig. 14 Mass transportation distance rank histograms for scenarios generated by (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) QR(1,1) and (d) QR(4,3).

Fig. 15 shows how the MTD rank histogram differs from the MST rank histogram for scenarios with unequal probabilities. Results are shown for the EPI(0.1, 0.9) scenarios without de-biasing or scaling.

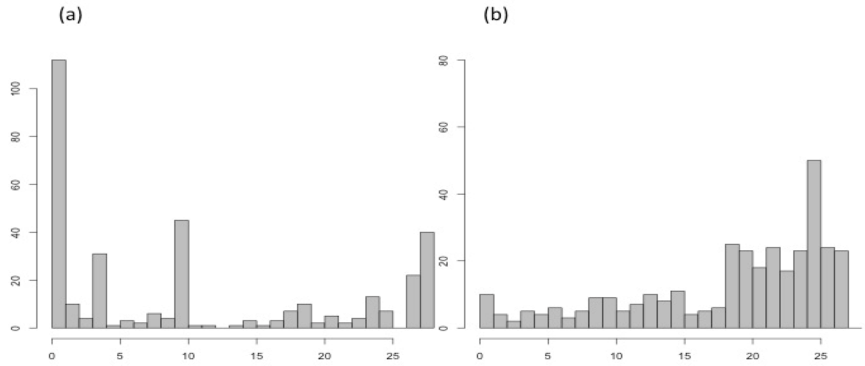


Fig. 15 (a) MST rank histogram when the scenario probabilities are not considered, (b) MTD rank histogram incorporating scenario probabilities.

Average daily Brier scores for all events and all parameters tested are fairly low for all of the scenario sets, as shown in Table 4, and quite similar across the scenario generation methods. For the shortest duration of one hour, the epi-spline scenarios have lower scores but these events are very rare overall. For event 9 (20% ramp down within 6 hours), although the average daily Brier score is slightly higher for scenario set generated by EPI(0.1, 0.9), hourly Brier scores are lower than 0.1 for more than 90% of the hours for the same scenario set, whereas for other scenario sets this proportion is approximately 85%. By changing the parameters, we can capture slight differences among scenario sets. The parameters corresponding to critical events should be determined according to the unit commitment problem results.

Table 4

Average daily Brier scores for ramp down and ramp up events with magnitudes ξ = 0.2, 0.4 and durations κ = 1, 3, 6 for scenarios generated by two variants of the epi-spline approximation approach and by three variants of the quantile regression with Gaussian copula approach.

Events:	EPI(0.1, 0.9)	EPI(0.33, 0.66)	QR(1,1)	QR(3,2)	QR(5,3)
1-RampDown(κ =1, ξ=0.2)	0.0015	0.0015	0.0023	0.0025	0.0031
2-RampDown(κ =1, ξ=0.4)	0.0000	0.0000	0.0002	0.0002	0.0004
3-RampUp(κ =1, ξ=0.2)	0.0046	0.0046	0.0050	0.0052	0.0057
4-RampUp(κ =1, ξ=0.4)	0.0001	0.0001	0.0003	0.0004	0.0005
5-RampDown(κ =3, ξ=0.2)	0.0335	0.0325	0.0320	0.0314	0.0321
6-RampDown(κ =3, ξ=0.4)	0.0024	0.0024	0.0029	0.0030	0.0030
7-RampUp(κ =3, ξ=0.2)	0.0452	0.0433	0.0398	0.0401	0.0402
8-RampUp(κ =3, ξ=0.4)	0.0066	0.0064	0.0064	0.0067	0.0069
9-RampDown(κ =6, ξ=0.2)	0.0645	0.0595	0.0614	0.0615	0.0614
10-RampDown(κ =6, ξ=0.4)	0.0133	0.0131	0.0140	0.0143	0.0142
11-RampUp(κ =6, ξ=0.2)	0.0641	0.0601	0.0593	0.0584	0.0602
12-RampUp(κ =6, ξ=0.4)	0.0322	0.0312	0.0297	0.0303	0.0300

The results in this section show the value of employing multiple verification metrics to assess different characteristics of scenarios. The energy score may be appealing as a single number but its emphasis on sharpness could introduce risk. For example, although the lowest ES is obtained from the EPI(0.33, 0.66), the resulting MTD rank histogram is not flat which means it is not as reliable as the other variants of approaches. Thus, we predict that this variant may give the highest cost in SUC problem among all of the scenario sets. The MTD rank histogram identifies whether both variance and autocorrelation in the scenarios match the observations. QR(4,3) is expected to result in slightly lower cost compared to QR(1,1) because its MTD rank histogram is flatter. Brier scores may be very useful but their relative values depend on the definition of critical events. Because high impact events such as steep ramps occur only rarely, the usefulness of statistical assessments may be limited.

5. Conclusions

High quality short-term wind power scenarios are very important for achieving cost savings by stochastic unit commitment. Aiming to assess the quality of probabilistic scenarios for wind power trajectories, we employed some existing verification approaches and introduced a mass transportation distance rank histogram to assess calibration of unequally likely scenarios. We applied them to scenario sets that were generated by two very different approaches, one of which produces unequally likely scenarios. On-going research focuses on finding relationships between the verification approach results and unit commitment problem results. We expect MTD rank histograms to predict SUC cost savings better than the ES, because reliability is a more crucial property of wind power scenarios than sharpness. Because scenario sets that are too sharp do not adequately describe the uncertainty, we do not view sharpness is one of the most desired properties of wind power scenarios. It appears that EPI(0.1, 0.9) and QR(4,3) are more competitive than the others presented in this paper.

Another way to predict how scenarios may perform in SUC is by examining Brier scores as well. However, it is very important to decide which events should be considered to distinguish between these scenario sets. The events should be chosen in such a way that they can distinguish whether the scenarios capture steep ramps which (a) may result in costly dispatch decisions in the recourse stage of SUC and (b) are present in the observations. Once these critical events are identified, the decomposition of Brier scores into reliability and resolution components can help to distinguish among sets of scenarios. The ramping event parameters to use in these scores should be determined by careful SUC simulations.

Acknowledgment: This material is based on work supported by the ARPA-E initiative of the U.S. Department of Energy under the Green Electricity Network Integration (GENI) program and by the Iowa Energy Center. We are grateful to Andrea Staid and anonymous reviewers for valuable comments.

References:

[1] S. Takriti, J. R. Birge, E. Long. A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems*, Vol. 11, No. 3, August 1996.

[2] A. Tuohy, P. Meibom, E. Denny, M. O’Malley. Unit commitment for systems with significant wind penetration. *IEEE Transactions on Power Systems*, Vol. 24, No. 2, May 2009.

[3] A. Papavasiliou, S. S. Oren, R. P. O’Neill. Reserve requirements for wind power integration: A scenario - based stochastic programming framework. *IEEE Transactions on Power Systems,* Vol, 26, No. 4, November 2011.

[4] K. Cheung, D. Gade, C. Silva-Monroy, S. Ryan, J-P Watson, R. Wets and D. Woodruff. Toward Scalable Stochastic Unit Commitment - Part 2: Assessing Solver Performance. Forthcoming in *Energy Systems*, 2015.

[5] Y. Feng, I. Rios, S. Ryan, K. Spürkel, J-P Watson, R. Wets, and D. Woodruff. Toward Scalable Stochastic Unit Commitment - Part 1: Load Scenario Generation. Forthcoming in *Energy Systems*, 2015.

[6] J. Wang, A. Botterud, R. Bessa, H. Keko, L. Carvalho, D. Issicaba, J. Sumaili, and V. Miranda. Wind power forecasting uncertainty and unit commitment. *Applied Energy*, 88(11):4014-4023, 2011.

[7] P. Pinson, H. Madsen, A. H. Nielsen, G. Papaefthymiou, and B. Klöckl. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12(1):51-62, 2009.

[8] J. O. Royset and R. J.-B. Wets. From data to assessments and decisions: Epi-spline Technology. INFORMS: *TutORials in Operations Research,* 2014.

[9] P. Pinson and R. Girard. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20, 2012.

[10] T. Gneiting, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211-235, 2008.

[11] P. Pinson and J. Tastu. Discrimination ability of the energy score. Technical report, Technical University of Denmark, 2013.

[12] D. S. Wilks. The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, 132, 1329-1340, 2004.

[13] D. Gombos, J. A. Hansen, J. Du, and J. McQueen. Theory and applications of the minimum spanning tree rank histogram. *Monthly Weather Review*, 135, 1490-1505, 2007.

[14] L. A. Smith. Disentangling uncertainty and error: on the predictability of nonlinear systems. *Nonlinear Dynamics and Statistics,* A. E. Mees, Ed., Birkhauer Press, 31-64, 2001.

[15] S. T. Rachev. Probability metrics and the stability of stochastic models. John Wiley, Chichester, UK, 1991.

[16] S.T. Rachev, L. Rüschendorf. *Mass Transportation Problems*, Vol. I and II. Springer-Verlag Berlin, 1998.

[17] J. Dupačová, N. Gröwe-Kuska, and W. Römisch. Scenario reduction in stochastic programming An approach using probability metrics. *Math. Program.,* Ser. A 95: 493–511, 2003.

[18] G. Brier. Verification of forecasts expressed in terms of probability. *Mon Weather Rev,* 78:1-3, 1950.

[19] I. Rios, R J-B Wets, and D.L. Woodruff. Multi-period forecasting and scenario generation with limited data. *Computational Management Science,* 12: 267-295*,* 2015.

[20] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69, 243–268, 2007.

[21] W. Hsu and A. Murphy. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2: 285-293, 1986.

[22] G. J. Szekely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:55-80, 2005.

[23] T. Thorarinsdottir, M. Scheuerer and C. Heinz. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. Forthcoming in *Journal of Computational and Graphical Statistics*, 2015.

[24] J. O. Royset and R. J.-B. Wets. Nonparametric density estimation via exponential epi-eplines: Fusion of soft and hard information. Technical report, Naval Postgraduate School, 2013.

[25] A.T. Al-Awami, and M.A. El-Sharkawi. Statistical characterization of wind power output for a given wind power forecast. *North American Power Symposium (NAPS)*, 2009

[26] H. Bludszuweit, J.A. Dominguez-Navarro and A. Llombart. Statistical Analysis of Wind Power Forecast Error. *IEEE Transactions on Power Systems*, vol. 23, no. 3 983-991, 2008.

[27] S. Bofinger, A. Luig and H. G. Beyer. Qualification of wind power forecasts. *in Proc. Global Wind Power Conf.*, Paris, France, 2002.

[28] http://www.bpa.gov/Projects/Initiatives/Wind/Documents/20140625-BPA-Super-Forecast-Methodology.pdf, viewed 12 March 2015.

[29] http://transmission.bpa.gov/Business/Operations/Wind/default.aspx, viewed 12 March 2015.

[30]http://www.bpa.gov/Projects/Initiatives/Wind/Pages/Wind-Power-Forecasting-Data.aspx, viewed 12 March 2015.

[31]http://transmission.bpa.gov/Business/Operations/Wind/WIND_InstalledCapacity_DATA.pdf viewed 12 March 2015.

[32] T. M. Hamill. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, 14, 155-167, 1999.

[33] N. J. Higham. Computing the nearest correlation matrix – A problem from finance. *MA Journal of Numerical Analysis* 22, 329–343, 2002.