

RESEARCH ARTICLE

Generating Short-Term Probabilistic Wind Power Scenarios via Non-Parametric Forecast Error Density Estimators

Andrea Staid¹, Jean-Paul Watson¹, Roger J.-B. Wets², and David L. Woodruff²

¹ Sandia National Laboratories, Albuquerque, New Mexico, USA

² University of California Davis, Davis, California, USA

ABSTRACT

Forecasts of available wind power are critical in key electric power systems operations planning problems, including economic dispatch and unit commitment. Such forecasts are necessarily uncertain, limiting the reliability and cost-effectiveness of operations planning models based on a single deterministic or “point” forecast. A common approach to address this limitation involves the use of a number of probabilistic scenarios, each specifying a possible trajectory of wind power production, with associated probability. We present and analyze a novel method for generating probabilistic wind power scenarios, leveraging available historical information in the form of forecasted and corresponding observed wind power time series. We estimate non-parametric forecast error densities, specifically using epi-spline basis functions, allowing us to capture the skewed and non-parametric nature of error densities observed in real-world data. We then describe a method to generate probabilistic scenarios from these basis functions that allows users to control for the degree to which extreme errors are captured. We compare the performance of our approach to the current state-of-the-art considering publicly available data associated with the Bonneville Power Administration, analyzing aggregate production of a number of wind farms over a large geographic region. Finally, we discuss the advantages of our approach in the context of specific power systems operations planning problems: stochastic unit commitment and economic dispatch. Our methodology is embodied in the joint Sandia – University of California Davis *Prescient* software package for assessing and analyzing stochastic operations strategies. Copyright © 0000 John Wiley & Sons, Ltd.

KEYWORDS

wind power; forecasting; uncertainty; probabilistic scenarios; stochastic unit commitment and economic dispatch.

Correspondence

Andrea Staid, Discrete Math and Optimization Department, Sandia National Laboratories, P.O. Box 5800, MS 1326, Albuquerque, New Mexico 87185 USA. E-mail: astaid@sandia.gov

Received . . .

1. INTRODUCTION

Over the last decade, models of power systems operations problems, such as unit commitment and economic dispatch, have increased in complexity and fidelity, driven both by improvements in algorithms and computational platforms and also out of necessity. An example of the latter is the rapid increase in the deployment of renewable resources over the last decade, particularly wind power [1]. In response to the inherent and often significant uncertainty associated with wind

power production, researchers have introduced a variety of advanced operations planning models that attempt to directly address this aspect of renewables integration. One exemplar is stochastic unit commitment, in which probabilistic scenarios of projected wind power serve as a key input [2]. While such stochastic operations planning models offer the potential for higher degrees of reliability and cost-effectiveness, their advantages cannot be realized without high-quality probabilistic wind power scenarios, which in turn must be extracted from high-accuracy stochastic process models. Specifically, the probabilistic scenarios used as input to stochastic operations planning models must represent the full range of potential outcomes as accurately as possible. If probabilistic wind power scenarios can be constructed in such a manner, a power system can be operated to minimize overall expected costs for the whole range of scenarios, while maintaining system reliability and ensuring maximal utilization (i.e., no curtailment) of available wind power. Including scenarios that are far outside the realm of likely realizations unnecessarily inflates operational costs, while exclusion of relevant scenarios impacts reliability.

Due in part to their central role in realizing the potential benefits of stochastic operations planning models, the construction of probabilistic wind power scenarios has recently gained significant attention from the research community, e.g., see [3]. However, a number of challenges remain, particularly as stochastic operations planning models are approaching the point where they may be ultimately deployed, due to recent advances in mitigating computational concerns, e.g., see [4, 5, 6].

Probabilistic wind power scenario generation methods commonly rely on the availability of a deterministic or “point” forecast – typically the output of a numerical weather prediction (NWP) model – which specifies a single trajectory of wind power over an operational time horizon. Probabilistic scenarios are then constructed by either sampling from parametric forms of assumed forecast error distributions or by analyzing and characterizing historical wind conditions, dynamics, and forecast errors. Realized wind power and forecast errors depend significantly on geographic location and local wind conditions, as well as the characteristics of the farm or set of farms being evaluated. Probabilistic wind power scenarios should be generated in such a way that they capture known relationships present in local wind patterns, such as temporal dependencies, forecast biases, or correlations among farms. Further, they should capture low-probability “tail” events, so that operations planning can address such events should they occur. On the other hand, probabilistic wind power scenario sets should not contain unrealistic behaviors, e.g., extreme and cyclic ramping events not seen in real data. A balance is difficult to achieve, as a set of scenarios should represent low-probability events, but not extremely unlikely events. One can address this balance by selecting scenarios that capture the range of wind behaviors, e.g., steep ramps, that have been historically observed.

In this paper, we present a novel method for generating probabilistic wind power scenarios, focusing on issues related to the use of such scenarios in stochastic power systems operations planning models. Our scenarios are based on non-parametric density estimates of forecast errors, as real-world data does not typically track parametric forms. For characterizing forecast error, we specifically rely on the recently introduced epi-spline basis functions [7]. Epi-splines are piecewise polynomial functions that facilitate non-parametric density estimation and allow the use of exogenous information concerning the qualitative behavior of a problem of interest. Epi-splines can and have been used in a number of applications, and are applicable as long as historical error data is available for estimation. These methods can be applied to probabilistic solar and load scenario generation, as well as to problems where the calculation of the error distribution is critical, such as the task of sizing electric power reserve requirements. We then construct probabilistic scenarios by specifying particular quantiles in the forecast error distributions. This approach allows for careful user control of the degree to which low-probability or “extreme” scenarios are represented. In contrast, sampling-based approaches require significantly more scenarios to capture the same behaviors, significantly inflating the difficulty of any associated stochastic power systems operations planning problem. Our proposed method also relies extensively on data segmentation. Segmentation approaches are typically highly application-specific, and we introduce non-trivial segmentation methods for wind power forecast data. These methods are described briefly in Section 3 and in detail in the Supplementary Material.

We begin in Section 2 by surveying prior efforts on probabilistic wind power forecasting, scenario generation, and associated evaluation metrics. In Section 3, we present our methodology for generating probabilistic wind power scenarios.

We compare the performance of our method with the current state-of-the-art method in Section 4, using a number of statistical evaluation metrics considering publicly available real-world wind power data obtained from the Bonneville Power Administration (BPA) in the US. Finally, we conclude in Section 5 by summarizing our contributions and offering thoughts and directions for future research. We note that our scenario construction methodology is embodied in the jointly developed Sandia – University of California Davis *Prescient* software package for assessing and analyzing stochastic power systems operations strategies; please contact authors for details regarding acquisition and use of this software. Similarly, the probabilistic scenario sets and the input data used in their construction are available for unrestricted public use from the authors.

2. BACKGROUND

Standard practice in power systems operations planning is to use a point forecast (often interpreted as the expected value forecast) for estimating future available wind power and to allocate thermal generator reserves to account for deviations of realized wind power from the forecast. The point forecast serves as input to deterministic optimization models, which are solved to obtain recommendations for operational decisions, including thermal generator on/off statuses and dispatch levels [8]. Similarly, point forecasts are commonly used in power systems simulations, e.g., production cost models and reliability analyses [9]. Despite their widespread use, there are well-known problems with the use of point forecasts in these and related contexts. In particular, available wind power is often the most uncertain quantity in power systems operation; ignoring this uncertainty results in higher costs and potentially unreliable operations. One approach to representing wind power forecast uncertainty is through the use of probabilistic scenarios, which collectively represent a range of potential wind power trajectories over the near term, with associated probabilities. These probabilistic scenarios can then serve as input to power systems operations planning models that explicitly address input parameter uncertainty, such as stochastic unit commitment [2], and Monte Carlo production cost and reliability simulations. Probabilistic scenarios can also serve as the basis of advanced visualizations regarding system operational risk, for presentation to system operators in the role of decision support.

One widespread approach to constructing probabilistic wind power scenarios involves fitting models based strictly on historical, *observed* wind power characteristics. For example, Morales *et al.* propose a methodology based on a time series analysis of historical wind power, while also maintaining spatial correlation across distinct wind farms [10]. However, in order for probabilistic scenarios to be truly effective in power system operations planning contexts, they must be based in at least part on short-term forecasts of available wind power, which provide the best information about near-term conditions. Early attempts at creating probabilistic wind power scenarios from forecasts were fairly simple. For example, Wang *et al.* assume a normal distribution for available wind power, with the point forecast taken as the distribution mean and a chosen percentage of the mean representing the standard deviation [11]. Pinson *et al.* proposed a greatly improved method that accounts for both the interdependence of wind power prediction errors and the predictive distributions [3]. Their method considers a multivariate Gaussian random variable, with the covariance matrix of the prediction errors being used to estimate the multivariate distribution. Although this method represents the current state-of-the-art in probabilistic wind power scenario generation and performs well relative to all known quality metrics, it possesses some shortfalls that may limit applicability and/or effectiveness in certain key practical settings. In particular, the sampling procedure employed can result in very erratic forecasted wind power trajectories (as we discuss later in Section 4), more so than those observed in reality. Further, while a single wind farm may exhibit very sharp ramp events on a regular basis, in an aggregated service area (e.g., consisting of multiple states and a dozen or more wind farms) these ramps are smoothed to varying degrees. Erratic trajectories can significantly impact the solutions obtained from a stochastic operations planning model, as the solution ensures feasibility in all potential scenarios at the expense of increased costs. Further, the Pinson *et al.* approach requires large numbers of samples to accurately represent low-probability “tail” events – and the computational difficulty of stochastic operations models is proportional to the number of probabilistic scenarios considered.

Here, we rely on non-parametric estimates of forecast error densities to construct probabilistic scenarios from the resultant distributions. We perform density estimation using epi-spline basis functions, which have been successfully demonstrated on a number of applications, ranging from financial planning to image reconstruction [7]. Rios *et al.* discuss the general use of epi-splines for probabilistic scenario generation [12]. Feng *et al.* successfully apply this methodology to probabilistic scenario generation for electricity load forecasts [13]. Probabilistic scenarios generated using this type of approach have been analyzed previously in the context of statistical quality metrics for wind scenario applications [14].

Extending the Rios *et al.* and Feng *et al.* probabilistic scenario generation methodologies to the context of wind power is the primary contribution of this paper, in addition to our analysis of the quality of the resulting scenarios relative to competing state-of-the-art methods – specifically that of Pinson *et al.* There are necessarily strong conceptual similarities to scenario generation for forecasted power of any kind, whether generated (e.g., wind and solar power) or consumed (e.g., load). While our method can be implemented using any non-parametric density estimator (including empirical PDFs), we have chosen to leverage epi-spline basis functions. Epi-splines can be highly constrained and highly parameterized based on underlying knowledge or beliefs about the stochastic process being modeled. Royset and Wets refer to such information as *soft knowledge*, which can include knowledge concerning the continuity, smoothness, unimodality, monotonicity, and moments of the expected density [15]. Finally, we observe that data segmentation is critical to the success of our methodology, and that the details are application-specific. Consequently, another contribution of this paper is our description of our segmentation procedure for wind power forecast data.

3. NON-PARAMETRIC FORECAST ERROR DENSITY ESTIMATION AND SCENARIO GENERATION

Our method for scenario generation consists of constructing non-parametric forecast error density estimates for some number of hours in a day, following a detailed segmentation of historical forecast and actual wind power time series. Scenarios are then constructed by (1) specifying quantiles from these error distributions and (2) forming forecasted wind power trajectories by “connecting” (via a process described subsequently) specified quantiles. Because we employ user-specified quantiles, our method results in scenarios that target specific partitions of the error distribution. For example, the resulting scenario sets may focus heavily on low-probability tail events, which are crucial for many applications, including power system unit commitment. Parameters and constraints can be adjusted based on the nature of the specific application, as wind conditions, wind variability, and forecast errors can vary drastically. The spread of the scenarios can be systematically adjusted to achieve sufficient coverage of the expected domain of local wind conditions. Our approach is based on the general scenario generation methodology introduced by [12], with the major difference being that we do not use leading indicators (because they are not readily available). Instead, we make use of exogenously supplied next-day forecasts. Further, we describe segmentation techniques specific to wind power, based on both the magnitude of the forecasted wind power and the pattern of local first derivatives. Finally, we note that although we focus on hourly time resolution and day-ahead scheduling in the presentation below, our methodology is generic with respect to time resolution.

Our general process involves first choosing a set of specific hours in the day-ahead forecast and computing the estimated forecast error densities for those hours in the next-day forecast horizon. The next step involves choosing a set of probability values at which to partition the cumulative distribution function (CDF) of the error density. Within each partition of the distribution, we calculate the representative forecast error based on the probability-weighted error values. These error values are then applied to next-day forecast for the specified hours with all combinations of hour and probability values used to represent the different scenarios. Intra-hour values are interpolated to form full, 24-hour scenarios. Therefore, our method is highly flexible due to the choices made with respect to these parameters.

3.1. Terminology

We begin by introducing key terms associated with our non-parametric forecast error density estimation and associated scenario generation processes:

- **Day Part Separators (DPSs):** The set of specific hours of the forecast horizon for which we compute forecast error densities, e.g., $\{1, 12, 24\}$. As described below, we perform interpolation for hours between DPSs. Given K DPSs, the set of DPSs is denoted by $\{t_1, t_2, \dots, t_K\}$. The specific choice of hours should depend at least in part on the correlation structure of the forecast error, which is generally low (e.g., on the order of hours) for wind [16]. We discuss the issue of wind power forecast error correlation further in Section 4, and provide an analysis of such correlation for BPA data in Figure 3.
- **Cut or Break Point Sets:** Values in the domain of the inverse of a (cumulative) forecast error density, specified as probabilities or quantiles. Every cut point set contains the values 0 and 1. For notational convenience, we typically assume that there is at least one point in the set other than 0 and 1, although this is not strictly required. Generally, we can define a cut point set for each DPS time t_k , denoted as the ordered set $C_k = \{c_1 = 0, \dots, c_{|C_k|} = 1\}$. However, we often assume a special case in which there is a single set of cut points shared among the DPSs $\{t_1, \dots, t_K\}$.

In general, cut points can be *path dependent*. For DPS t_1 , there is one set of cut points in all cases, but in general there can be additional – and conditional – cut point sets for DPSs in $\{t_2, \dots, t_K\}$. For DPS t_2 , there can be one cut point set for each of the cut points associated with DPS t_1 , and so on. Thus, for DPSs with index $k > 1$, we can in general define the path dependent cut point sets

$$C_{k,\ell} = \{c_{1,\ell} = 0, \dots, c_{|C_{k,\ell}|,\ell} = 1\}, \forall \ell \in \mathcal{L}$$

where \mathcal{L} denotes the cross product (all possible combinations) of all cut point sets for all DPSs in $\{2, \dots, K\}$. While our methodology and associated Prescient software package support both standard and path-dependent cut points, we focus strictly on the former in this paper.

- **Skeleton Points:** For each DPS t_k , we define a set of skeleton points \mathcal{N}_k , the individual members of which are defined as those wind power values at time t_k that are *representative* of the forecast error partition between two adjacent cut points c_{i-1} and c_i in C_k for $k \geq 1$, with appropriate generalization for path-dependent cut points. These values represent the mean of the distribution within the partition. For a skeleton point $n \in \mathcal{N}_k$, we denote the corresponding wind power and associated probability respectively by $l(n)$ and $\pi(n)$.
- **Scenarios:** A scenario is defined as a time series of forecasted wind power quantities. Quantities at hours associated with skeleton points are computed using estimated error densities applied to forecasts, while values for hours between skeleton points are computed by interpolating the difference between the bounding skeleton points and the forecast.
- **Historical Database:** Our methodology assumes the availability of a historic database \mathcal{W} of forecasted wind power time series and associated observations, i.e., actuals.
- **Cumulative Forecast Error Density:** For each DPS t_k , we denote the associated cumulative forecast error density by $\Phi_{\mathcal{E}_k}$. Although not identified in the notation (for purposes of simplicity), $\Phi_{\mathcal{E}_k}$ is conditional on the specific hour for which a wind power forecast is issued.

3.2. Non-Parametric Forecast Error Density Estimation

While our approach is independent of the particular method used to fit the forecast error densities $\Phi_{\mathcal{E}_k}$, the accuracy of the resulting scenarios is obviously dependent on the specifics. Based on our experience with BPA and other real-world data sets, non-parametric estimators are critical in wind power applications. Specifically, wind power forecast errors are often skewed, conditional on the power regime (e.g., low, medium, or high), and are even then not easily captured via standard

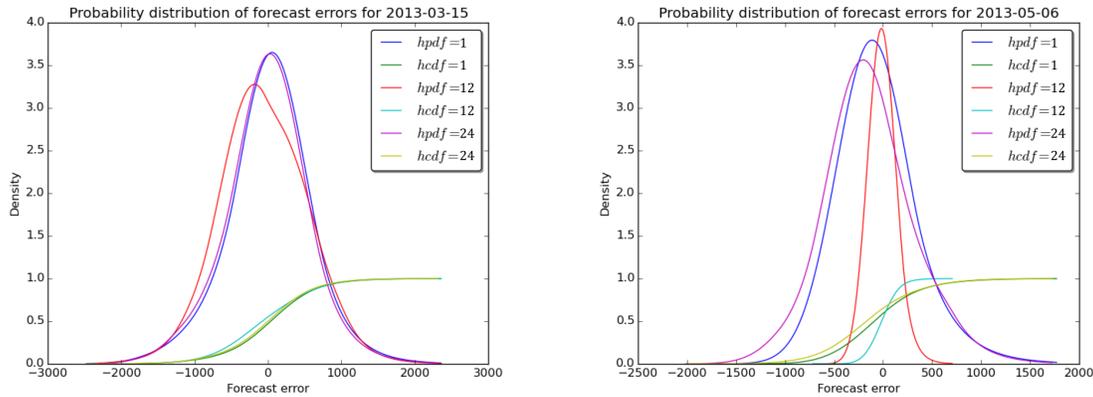


Figure 1. Examples of wind power forecast error distributions within BPA, illustrating variations in skew, variance, and distribution shape across hours.

parametric forms. Figure 1 shows two examples of the distribution of BPA forecast errors for two days, for each DPS. Both the probability density function (PDF) and cumulative density function (CDF) are shown for hours 1, 12, and 24 on the respective days.

In this work, we use exponential epi-splines to compute the $\Phi_{\mathcal{E}_k}$ [7]. For details on this method, we defer to [12]. Our choice was driven in large part due to the accuracy of exponential epi-splines in the face of limited data. However, we observe that simpler, empirical error densities could also be employed. Clearly, the wind power generated cannot be below zero or above the installed capacity, so when the error distributions imply power outside these values the distribution is truncated and the probability outside the boundary is assigned to the boundary. The segmentation methods described in Section 3.4 seek to minimize this effect.

3.3. Scenario Generation

To generate scenarios for a given day, we assume the availability of a wind power forecast (e.g., obtained via a NWP model) \hat{l}_h for $h \in \{1, \dots, 24\}$ for the next day. For each DPS t_k , we identify a set of historical forecast errors \mathcal{E}_k relevant to \hat{l}_{t_k} . By “relevant”, we mean that we identify – through the segmentation process described below and in the supplementary material – historical forecasts and associated realizations that are closely related (specifically in terms of wind power and local derivative) to \hat{l}_{t_k} . We then non-parametrically estimate the cumulative error densities $\Phi_{\mathcal{E}_k}$. In part through inversion of the $\Phi_{\mathcal{E}_k}$, we finally compute the skeleton points \mathcal{N}_k .

Scenarios are formed by combining skeleton points at distinct times t_k , forming a *skeleton* whose probability is straightforward to compute under the reasonable assumption of uncorrelated forecast errors at different DPSs. For the data from BPA used in our experiments, forecast errors are not significantly correlated beyond lags of approximately 4 hours (e.g., see Figure 3 and associated discussion in Section 4), so this assumption does not limit our choice of DPSs in any meaningful way. One can imagine situations where the time lags needed for uncorrelated errors would be quite large, in which case the probability calculations would need to be modified accordingly. One option for doing so is the use of copulas to capture the dependence between random variables [17]. Finally, for hours not associated with a DPS, scenario wind power values are computed by interpolating between the skeleton points using the forecast.

The choice of day part separators and cut points determines the number of scenarios and the extent to which they represent the tails of the distribution. When the cut points are not path dependent, then for each DPS, the number of skeleton points is one less than the number of cut points and the number of scenarios is the product of these numbers over the day part separators. When the cut points are path dependent, there is simply one scenario per path (unless some paths are coincident so they are combined to form one scenario).

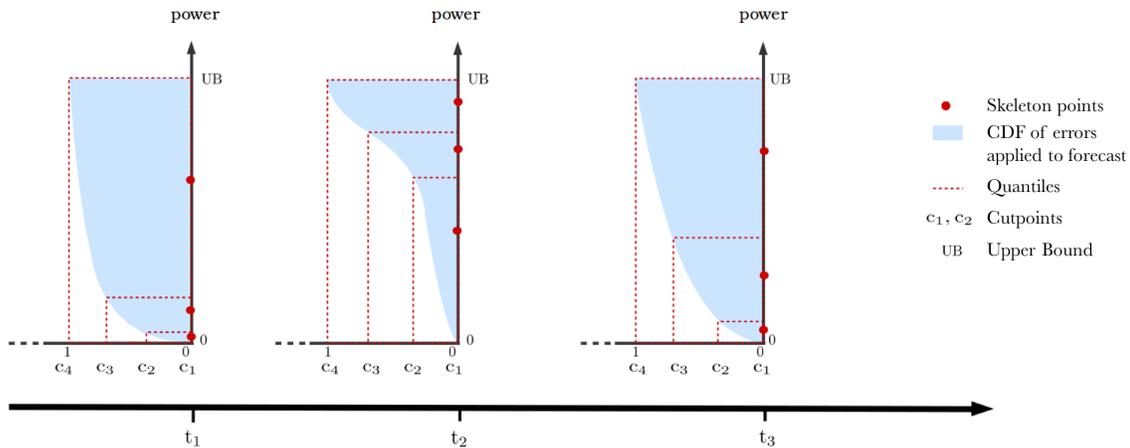


Figure 2. A graphical representation of the process for computing skeleton points.

For simplicity, we continue to assume that the cut points are *not* path dependent, in order to minimize the notational burden. Our scenario generation process is then specified as follows, and is executed for each DPS indexed by k :

1. Inputs: DPSs indexed by k and corresponding cut point sets C_k .
2. Initialize the set of skeleton points: $\mathcal{N}_k \leftarrow \emptyset$.
3. From the historic wind power database \mathcal{W} , identify a set of forecast errors \mathcal{E}_k appropriate for estimating the error distribution around the next-day forecast \hat{l}_{t_k} , by focusing on similar historical data. This segmentation process is described in detail below, and in the supplementary material.
4. Non-parametrically estimate the error density $\phi_{\mathcal{E}_k}$.
5. Integrate $\phi_{\mathcal{E}_k}$ to obtain the cumulative error density $\Phi_{\mathcal{E}_k}$.
6. For $j \in \{2, \dots, |C_k|\}$, add n to \mathcal{N}_k such that:

$$l(n) = \hat{l}_{t_k} + \int_{v_{j-1}}^{v_j} x \phi_{\mathcal{E}_k}(x) dx$$

$$\pi(n) = c_j - c_{j-1}$$

where $v_j = \Phi_{\mathcal{E}_k}^{-1}(c_j)$ for $j \in \{1, \dots, |C_k|\}$

We graphically illustrate the computation of the skeleton points $l(n) \in \mathcal{N}_k$ in Figure 2.

The skeleton points form the basis of scenario generation. Following [12], we combine skeleton points to form scenarios, with associated probability $p(n)$ taken as the product of the composite skeleton point probabilities. As indicated previously, this computation explicitly assumes no correlation in forecast errors between DPSs – an issue that can be rectified through the use of copulas, for example, if warranted. We denote the difference between the value of the scenario skeleton points and the day-ahead forecast values by $(d_k)_{k \in 1, \dots, K}$, such that $d_k = l(n_k) - \hat{l}_{t_k}$. For each hour $h \in [t_{k-1}, t_k]$ that is not associated with a DPS, we compute the linear interpolation $d(\cdot)$ between the points (t_{k-1}, d_{k-1}) and (t_k, d_k) . The interpolated value of the scenario at hour h is then given by $l(h) = d(h) + \hat{l}_h$.

One advantage of our approach over the state-of-the-art multivariate method is that we do *not* rely on sampling to generate probabilistic scenarios. Rather, through the use of cut points and our associated methodology, we are able to provide parametric control to users over the quantiles of the forecast error distribution considered. We argue that this capability is critical in the context of advanced power systems operations planning models, e.g., stochastic unit commitment and dispatch, as we can carefully control for the degree to which low-probability or “extreme” events are considered – without the need to generate large numbers of scenarios as a by-product.

3.4. Data Segmentation

As indicated in the pseudo-code above, we perform segmentation or sub-selection of the historical forecast data \mathcal{W} prior to estimating error densities ϕ_{ε_k} . We leverage a dynamic segmentation process that identifies, for a particular day-ahead forecast \hat{w} and for each DPS t_k of that day, those historical forecasts $w \in \mathcal{W}$ that are “close” in a certain sense to \hat{w} . We define these measures in the supplementary material. Error density estimation is then performed using only similar, relevant historical forecasts. Through segmentation, we attempt to explicitly capture the notion that error density structures can be dependent on the nature of the forecast \hat{w} . In this work, we segment data in \mathcal{W} based on two attributes: the magnitude of the wind power forecast and the local derivative pattern.

We take wind power magnitude into account during segmentation by using only historical data whose forecasted values are within a given *window* of the forecasted quantity \hat{w} . Our approach is a sliding window version of the segmentation procedure proposed by [18, 19], who in turn cite [20] as the original work. The window width is controlled by a parameter defined on the unit interval (typically 0.4 in our experiments) that specifies the fraction of the distribution of observed historical forecasts to consider, centered around the forecast to the extent possible (e.g., for high or low forecasts, obtaining 0.4 of the data requires that the top or bottom 0.4 are used).

We account for derivative patterns in forecasted wind power by developing an approach to clustering forecasts considering their “local shape”. Our approach is based on the empirical hypothesis that forecast errors are linked to the qualitative, local “shape” of the time series. The heuristic idea is that the presence of weather fronts determines the local shape of the time series and these fronts have an impact on forecast errors. Consider a wind power forecast vector for “tomorrow”, denoted $(\hat{l}_h)_{h \in H}$, where $H = \{1, \dots, 24\}$ denotes the hours in a day. For each hour $h \in H$, our goal is to determine the set of historical forecasts $\tilde{\mathcal{W}} \subseteq \mathcal{W}$ that have similar shapes at hour h . Our method for computing the $\tilde{\mathcal{W}}$ for relevant wind derivative patterns is detailed in the supplementary material.

4. APPLICATION AND EVALUATION

We now compare our method to a current state-of-the-art method – that of Pinson *et al.* We use publicly available data from the Bonneville Power Administration (BPA) balancing authority in the US for evaluation, which consists of historical wind power forecast and corresponding actual (measured) time series. BPA has over 4500 MW of installed wind capacity, with most farms located along the Columbia River in Oregon and Washington [21]. This dataset differs from that seen in many wind power forecasting applications and studies, in that we only consider the *aggregated* data, instead of that for individual wind farms (e.g., see [3]). Specifically, forecasts and actuals represent the output of *all* wind farms within the balancing authority area. Aggregated power forecasts are utilized in many power systems applications, specifically in day-ahead and related reliability commitment processes executed by large system operators. We generate and evaluate probabilistic scenarios of aggregated wind power, although the described method can be applied at any scale (both spatial and temporal), depending on the availability of relevant data. The comparison of the two methods is specific to the aggregated dataset used. We use data from June 2012 through September 2013. The results presented here consider daily probabilistic scenarios over a nine month period, from January 2013 through September 2013. All 2012 data is used for training both methods [22].

We generate probabilistic wind power scenarios using both our non-parametric method and a state-of-the-art multivariate Gaussian method proposed by Pinson *et al.*, which we subsequently refer to simply as the multivariate method [3]. For each day, we use the official BPA forecast issued at 11AM local time to generate probabilistic wind power scenarios for the next day (12am to 11pm). This results in forecast lead times ranging from 13 to 36 hours. We use a rolling-horizon training window when assessing both methods, to represent a re-enactment of the data that would have been available to operators in real time, i.e., no future data is used. We start with a minimum of seven months of training data (June through December 2012) to generate probabilistic scenarios for the next day, and then add an additional day of training data (from the current

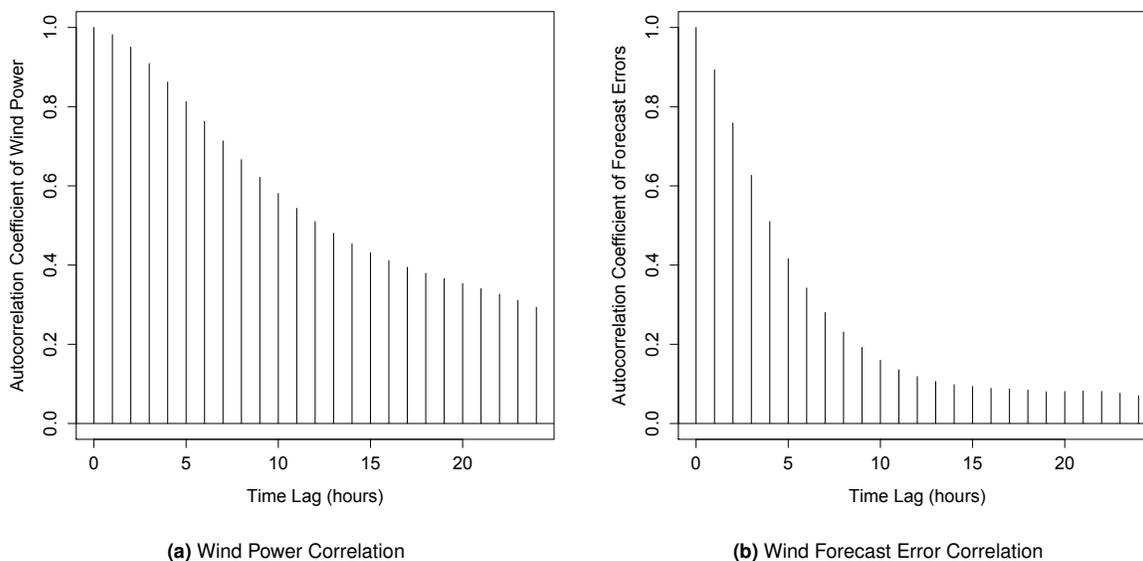


Figure 3. Autocorrelation coefficients for actual values of wind power (3a) and for the associated forecast error (3b). Note that the correlation of forecast errors is significantly lower than that of actual power.

day) for each subsequent day. Therefore, probabilistic scenarios for the most recent days are based on significantly more training data than the earliest days. The computational run-time is negligible for both methods.

We compare the probabilistic scenarios resulting from the two methods using a selection of relevant scenario evaluation metrics. The Energy score, the Brier score, and Minimum Spanning Tree (MST) rank histograms are established methods for comparing sets of probabilistic scenarios [23]. However, these metrics are known to lack sufficient ability to discriminate in some circumstances [24, 14]. For this reason, we also include the Variogram score of order p , which can detect misspecified correlations and is applicable to multivariate, ensemble-based forecasts [25]. In addition to these metrics, we consider a relatively simple measure, the Integrated Distance, discussed subsequently. With the exception of the MST rank histogram, all metrics are probability-weighted. Each scenario has an assigned probability, which is used in the metric evaluations so that higher-probability scenarios are given more weight. The multivariate scenarios are designed to have equal probability, as they are randomly sampled from a distribution. Our non-parametric scenarios, on the other hand, generally possess unequal probabilities. Here, the probability assigned to each scenario is based on forecast error distributions and the probability of falling in a given interval of that distribution. The method is agnostic to where the error distribution comes from and could be applied to distributions covering multiple areas and from multiple data sources.

We note that our non-parametric method is highly configurable. Our initial evaluation was conducted using several distinct sets of scenarios that were created with different cut point sets. In our final evaluation, we chose the parameterization that yielded the strongest and most consistent performance. Specifically, we use hours 1, 12, and 24 as the DPSs, and a shared cut point set $C_k = \{0.0, 0.1, 0.9, 1.0\}$, resulting in sets of 27 scenarios for each day. We also generate the same number of scenarios using the multivariate method for all comparisons presented here. Our choice for DPSs is driven by the observation that there is negligible correlation between forecast errors for these time periods, allowing us to assume that the errors are independent – which is particularly critical when computing scenario probabilities. We show the empirical correlations for our BPA wind power data in Figure 3, which illustrates that the correlation in wind power *errors* drops off more quickly than the correlation in wind power itself.

Regarding the choice of parameters for our non-parametric method, we believe there is still room for improvement, as our experimentation was limited. Further, there are additional parameters that can be modified and constraints that could be considered (specifically associated with fitting the epi-spline basis functions), providing additional degrees of freedom for customization. We observe that different cut point sets result in scenarios that, when used in unit commitment contexts,

yield solutions with varying reliability and cost. Analyses considering the relationship between cut point sets and unit commitment or dispatch solution reliability and cost is beyond the present scope, but is ultimately critical for operational environments.

4.1. Comparative Baseline

We compare our method to that of Pinson *et al.* [3]. We implemented their method following the description in their original manuscript. We first calculate the covariance matrix for the 24-hour forecast horizon based on quantile regression of the training data. To generate scenarios for a given day, we estimate the marginal distribution for each hour in the 24-hour period and then follow the three main steps detailed by Pinson *et al.*: (i) sample from a multivariate Gaussian distribution with $\mu = 0$ and σ equal to the recursively estimated covariance matrix with a forgetting factor of $\lambda = 0.95$, (ii) apply the inverse probit function to obtain the CDF, and (iii) take the inverse of the CDF to obtain forecasted wind power quantities.

During the course of our analysis, we expanded on the basic method of Pinson *et al.* by testing alternative formulas for use in the quantile regression of the training data. Initially, we used the forecast issued for hour $t + k$ to predict hour $t + k$. Subsequently, we used all 24 forecast hours to predict hour $t + k$, computing the first several principal components for all 24 forecast hours, principal components for the slope of the forecast data, and for the extended forecast and slope when using an additional 2 hours before and after the day being predicted. In addition, we used cubic spline formulations of the above mentioned variations, with degrees of freedom ranging from 1 to 5. No notable performance difference was observed in these alternatives relative to the basic method. Thus, all of our subsequent evaluations are based on the simplest variant of the Pinson *et al.* method, with one degree of freedom.

4.2. Qualitative Assessment

We first consider a qualitative analysis of the probabilistic wind power scenarios generated by the two methods, focusing on visual comparison of the results. In Figure 4, we show scenarios corresponding to June 20, 2013. The selection of this particular day is arbitrary; the results are representative of those for other days analyzed. The graphics in Figure 4(a) and Figure 4(b) respectively correspond to probabilistic scenarios generated by the multivariate and non-parametric methods. Each graphic depicts 27 scenarios. Scenario trajectories are colored in blue, while forecasted and actual wind power trajectories are respectively colored in black and red. To simplify the depictions, scenario probabilities are not captured in the figures. We immediately observe that the scenarios generated by the two methods are visually very different from each other. The most striking difference relates to the erratic nature of the multivariate scenarios. Specifically, they exhibit ramping behavior that is not present in the non-parametric scenarios nor the actual wind power trajectories.

Multivariate scenarios are generated via random sampling from distribution models based on historical data. Although the distributions account for error covariance, there is nothing that constrains the between-hour variation in the sampling. Thus, the scenarios can proceed from anywhere along the distribution in consecutive hours, allowing for large and occasionally cyclic ramps. This may be an accurate representation of actual wind conditions in some areas. However, we are evaluating the two methods on an aggregated area with multiple farms, where such erratic behavior is unexpected and unrealistic.

Standard evaluation metrics for probabilistic scenario sets are not always adept at detecting visually obvious differences, and research advances continue to improve such metrics [24, 26, 14]. Nevertheless, existing evaluation metrics do provide a means (if imperfect) to quantify the quality, accuracy, and skill of scenario sets – all of which are necessary qualities of high-quality probabilistic scenarios. Thus, it is critical that the scenarios generated by our non-parametric method achieve state-of-the-art performance levels relative to these metrics – even if the qualitative differences suggest that non-parametric scenarios are qualitatively distinct and possibly superior. We use standard comparative metrics here and leave the development of advanced metrics for scenario quality determination to future research. We note that such research can include both investigations of new metrics that capture the qualitative differences observed in Figure 4 and metrics that investigate quality in terms of their impact on specific power systems operations problems, e.g., stochastic unit commitment and economic dispatch.

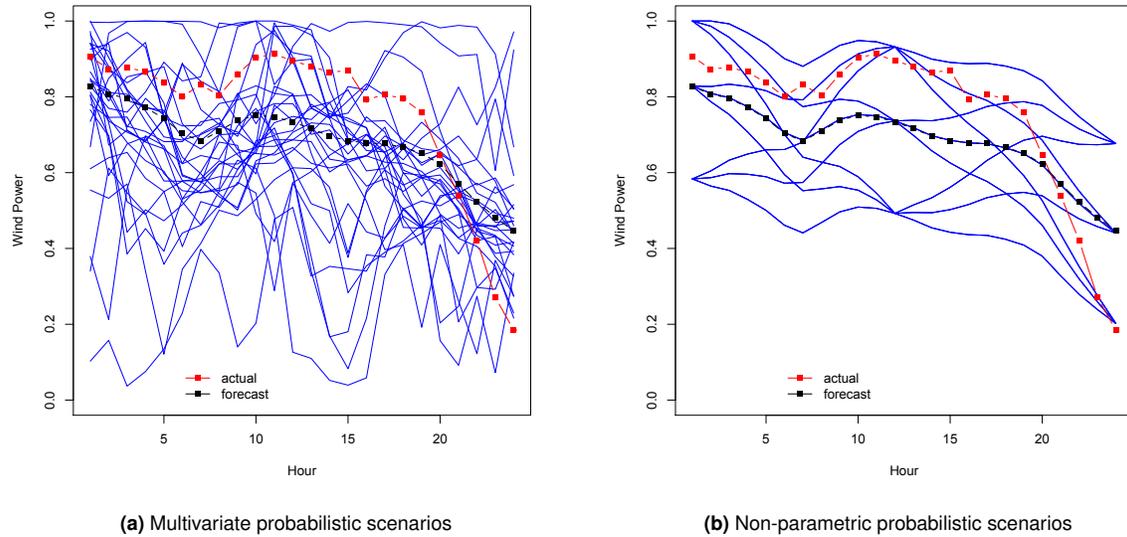


Figure 4. Probabilistic scenarios constructed using both the multivariate method (4a) and the non-parametric method (4b) for June 20, 2013. The plots depict 27 scenarios apiece, in addition to the point forecast and actual wind power.

4.3. Behavior Matching

Any probabilistic wind power scenario set should possess similar statistical properties to that of actual wind data. Thus, the qualitative structure of scenarios associated with a wind farm that routinely sees large fluctuations would not be appropriate when considering an aggregated set of wind farms, in which the fluctuations are somewhat smoothed by virtue of spatial distribution. As discussed previously, the sampling process used to generate multivariate scenarios introduces some large ramps and other erratic behavior. Deviations of this magnitude are not representative of the actual wind characteristics for the BPA balancing authority area. While multivariate scenarios may be better suited to a single wind farm, they introduce unrealistic probabilistic scenarios for BPA applications. In our non-parametric method, we emphasize the need for application-specific segmentation, and only train on data with similar wind conditions to the present forecast. Lacking the random sampling that is integral to the multivariate method, our approximation method for constructing scenarios from non-parametric error density estimates yields smoother wind power trajectories, with more gradual ramps.

Scenarios that better match the behavioral properties of local wind are more appropriate for use when operating a specific power system. Of course, heuristic methods exist that could be used to filter the multivariate scenarios based on information about the historic wind conditions and what is deemed to be realistic wind behavior. However, our non-parametric scenarios better and inherently match the behavioral characteristics of the actual wind, without resorting to heuristics to filter out unwanted scenarios. Information regarding the behavior of probabilistic scenario sets and actual wind observations is reported in Table I. We compare the ramp events occurring in the actual observations and the ramp events found in probabilistic scenarios. The multivariate scenarios tend to have steeper ramps than the actuals. Our non-parametric scenarios do not exactly match the behavior of the actual observations, but more closely match the expected percentages of ramp events, especially for small ramps of 1 or 2%. The multivariate scenarios additionally have more frequent large ramp events. For example, in the actual observations 93.2% of the hourly deviations fall within $\pm 10\%$ of the total wind power capacity. The multivariate scenarios, in contrast, have only 89.9% of ramps within $\pm 10\%$, indicating the relative frequency of larger ramps. Our non-parametric scenarios might prove to be slightly smooth in contrast, over-estimating the proportion of these modest ramps, thus decreasing the frequency of very large ramps. Finally, we note that the multivariate scenarios have significantly more extreme ramps (greater than $\pm 20\%$), which is consistent with the depictions in Figure 4.

Event	Actuals	Multivariate	Non-Parametric
Ramps within ± 1%	37.8%	33.3%	37.5%
Ramps within ± 2%	52.3%	47.9%	54.2%
Ramps within ± 5%	76.7%	72.8%	83.5%
Ramps within ± 10%	93.2%	89.9%	98.1%
Ramps within ± 20%	99.4%	97.1%	99.9%

Table I. Proportion of hourly ramp events within specified wind capacity thresholds, for both actual observations and the probabilistic wind scenario sets. The results for the non-parametric scenarios are based only on one set of outpoints.

Method	Energy score (std. dev.)
Multivariate	0.321 (0.171)
Non-Parametric	0.330 (0.196)
Paired t-test p-value	0.0067

Table II. Energy score means and standard deviations for probabilistic scenarios generated for January through September 2013.

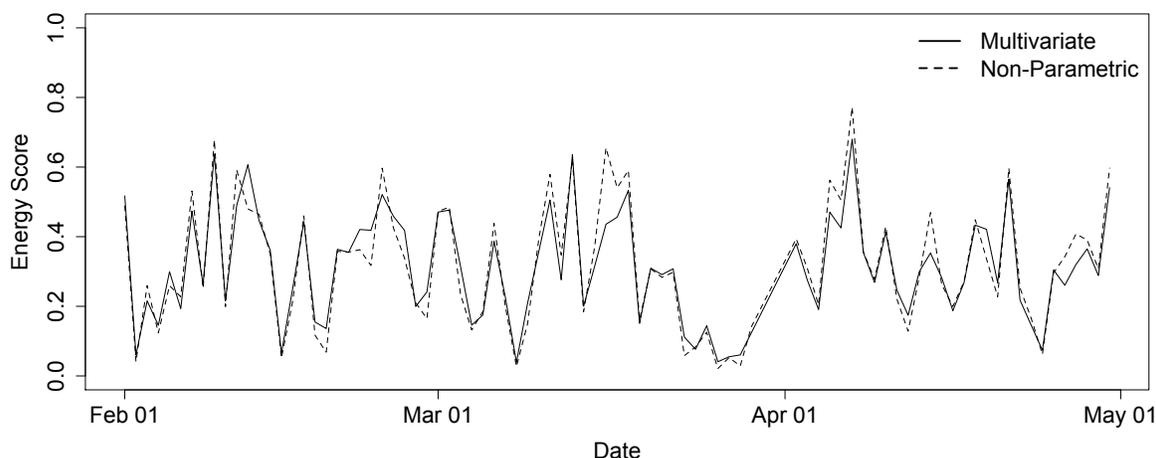


Figure 5. Daily Energy scores for days from February through April 2013.

4.4. Energy Score

The Energy score or metric is a proper (i.e., a perfect forecast will result in the best score), negatively-oriented (i.e., lower is better) score that quantifies both the skill (accuracy) and sharpness (spread) of a scenario set [27, 28]. The Energy score is computed as

$$ES = \sum_{j=1}^J p_j \|z - \hat{z}^{(j)}\|_2 - \frac{1}{2} \sum_{i=1}^J \sum_{j=1}^J p_i p_j \|\hat{z}^{(i)} - \hat{z}^{(j)}\|_2 \tag{1}$$

where J denotes the number of scenarios ($J = 27$ for all comparisons presented here), p_j denotes the probability of scenario j , \hat{z} denotes the set of probabilistic scenarios, and z denotes the actual wind power trajectory. Finally, $\|\cdot\|_2$ denotes the Euclidean norm of the 24-hour vectors. Here, we evaluate the Energy score for each of our test days in 2013 and compare the daily scores of the multivariate and non-parametric methods. The mean and standard deviation of the Energy scores are reported in Table II, while Figure 5 depicts the Energy score over a period of three months. The Energy scores of probabilistic scenarios generated by the two methods are very similar, despite large visual differences. Further, all scenario sets achieve low Energy scores on average, and even the day-to-day variation tracks very closely between scenario sets. Fundamentally, the results indicate that probabilistic scenarios generated by both methods perform well, and identify limitations associated with the Energy score metric in terms of discrimination ability.

Event	h	Threshold ξ	Multivariate	Non-Parametric	Paired t-test p-value
Gradient	3	20%	0.087	0.088	0.5057
Gradient	2	20%	0.046	0.046	0.9321
Ramp Up	1	10%	0.039	0.044	9.79E-07
Ramp Up	2	20%	0.027	0.030	0.0001
Ramp Down	1	10%	0.029	0.027	0.0133
Ramp Down	2	20%	0.015	0.014	0.0818

Table III. Average daily Brier scores from January 2013 through September 2013, for 6 predefined reference events.

4.5. Brier Score

The Brier score is also proper and negatively-oriented. It is an event-based evaluation metric, which assesses scenario sets based on how well they perform with respect to a pre-defined event [29]. In the case of probabilistic wind scenarios, events of interest include ramps (up or down) and the simple gradient over a given time period. The Brier score for a given set of probabilistic scenarios is calculated as

$$BS = \frac{1}{T} \sum_{t=1}^T (P_t[g(\hat{\mathbf{z}}; \theta)] - g(\mathbf{z}; \theta))^2 \quad (2)$$

where T denotes the number of time periods evaluated, θ denotes a parameter set that defines the reference event, and \mathbf{z} denotes the observed wind power trajectory. The quantity $g(\mathbf{z}; \theta)$ indicates whether the defined event occurs in the observed wind power trajectory, while $P_t[g(\hat{\mathbf{z}}; \theta)]$ denotes the probability of the same event occurring in a set of probabilistic scenarios. For example, in the case of a reference event with a parameter set consisting of the overall gradient threshold (ξ) in a given time window (k hours) starting from a given time step (t), the respective formulas are given as

$$g(\mathbf{z}; \theta) = \mathbb{1} \left\{ \left(\max_{i \in \{t, \dots, t+k\}} y_i - \min_{i \in \{t, \dots, t+k\}} y_i \right) \geq \xi \right\} \quad (3)$$

$$P_t[g(\hat{\mathbf{z}}; \theta)] = \sum_{j=1}^J p_j g(\hat{\mathbf{z}}^{(j)}; \theta) \quad (4)$$

where J denotes the number of scenarios and $\hat{\mathbf{z}}$ denotes the probabilistic scenario set. In addition to overall gradient, other events can be evaluated using the Brier score. Table III reports mean Brier scores of our BPA test data for several different events of interest. The Brier score can change drastically based on how an event is defined, and a probabilistic scenario set with very low scores for one event may have very high scores for a different event. Therefore it is important to identify events that are relevant to the problem at hand. In the case of wind power, large ramps in small time frames are often the most critical events for power systems operations planning. In this context, we consider both up- and down-ramps for two different windows and thresholds: a 10% change (relative to capacity) in 1 hour and 20% change in 2 hours. We also consider overall gradient events, which evaluate the difference between the highest and lowest wind power quantity in a given time window. This event definition provides an indicator of variability, but is not confined to the strict definition of a monotonic ramp. In general, we observe neither consistent nor significant overall differences in the Brier scores of the multivariate and non-parametric scenarios. Finally, we show the time series of Brier scores for a gradient event and a ramp down event in Figure 6. While there are rare large daily differences, neither of the methods consistently out- or under-performs the other.

4.6. Rank Histograms

A Minimum Spanning Tree (MST) rank histogram is a visual metric for assessing the quality of probabilistic scenarios [30]. To create the histogram, we first calculate the MST lengths for the set of all scenarios (measuring the distances among all scenarios). We then calculate the MST lengths for each set of scenarios again, except that we substitute in the observed

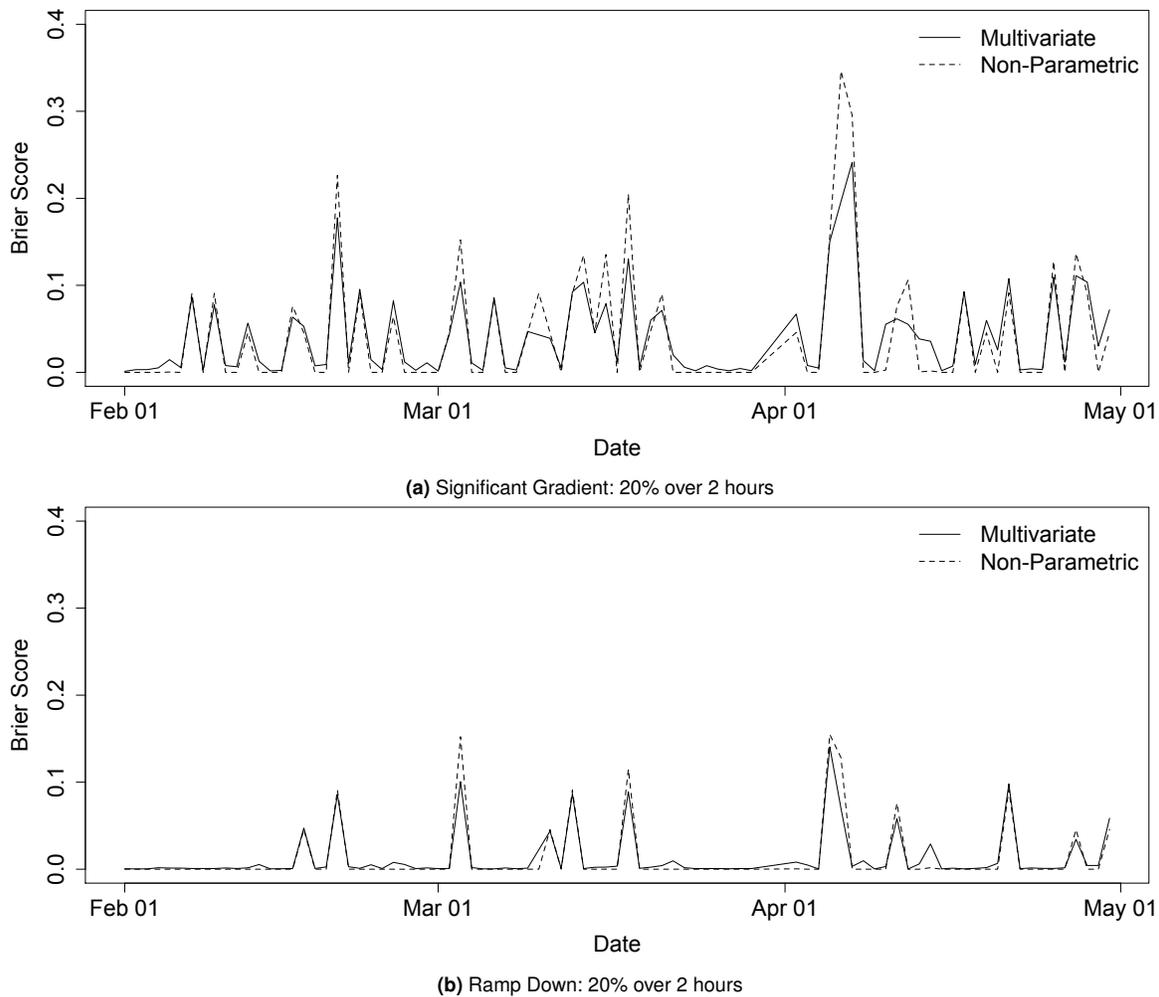


Figure 6. Brier scores for probabilistic wind power scenarios from February 2013 through April 2013, for events of a significant gradient of 20% over 2 hours (6a) and a ramp down of 20% over 2 hours (6b).

trajectory for one of the scenarios and repeat this until we have calculated the MST length for the observed trajectory taking the place of each scenario individually. This results in a set of MST lengths equal to the number of scenarios plus one. We then order these MST lengths and identify the position of the length that corresponds to the scenario-only tree. The value of interest is the rank (position) of the tree length containing only the scenarios among the lengths of all other trees, which include the observation in place of each scenario in turn. The histogram then shows the distribution of these ranks over all of the days being assessed. If the scenarios are drawn from the same distribution as the observation, the histogram should be uniform, since the observation should be indistinguishable, on average, from any one of the scenarios. Deviations from uniformity in the histogram can indicate problems of bias, underdispersion, or overdispersion. MST rank histograms are useful in that they provide this additional information.

Figure 7 shows the MST rank histograms for the multivariate method and the non-parametric method. The multivariate histogram has a slight peak towards the center of the distribution, which may indicate a slight overdispersion, but in general, the histogram looks reasonable. In the case of the non-parametric method, the scenarios are not sampled from a distribution. Since a ‘good’ rank histogram expects to see samples drawn from the same distribution, the non-parametric scenarios perform poorly when put to this test. The non-parametric scenarios follow similar trajectories and are grouped together through specified cut-points. For this reason, MST rank histograms for the non-parametric scenarios show large underdispersion, indicated by the large spikes for the highest and lowest ranks. In most cases, the scenarios are much more

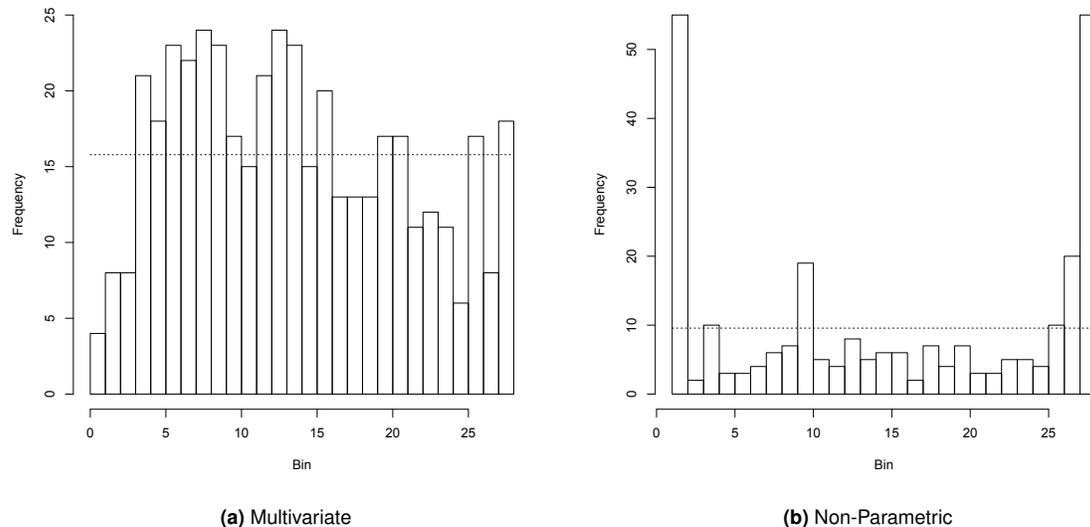


Figure 7. Minimum Spanning Tree (MST) rank histograms. Note the different y-axis scales. The horizontal dotted line represents the height of the ideal uniform histogram.

similar to each other than they are to the observed trajectory, and this results in the within-scenario tree having the shortest length on average.

There has been considerable work done to improve and understand histogram-based metrics for multivariate quantities [31]. Despite improvements, histogram methods are not appropriate evaluation metrics for our scenarios presented here, as they are not designed to capture the range of a distribution and can instead be parameterized to target certain ranges of a distribution.

4.7. Variogram Score

The Variogram score of order p is a proper, negatively oriented multivariate score based on pairwise differences between components [25]. In general, a variogram describes the dependence of data across space and/or time. Thus, in contrast to the Energy score, the Variogram score captures correlations between multivariate components. The Variogram score is analyzed in [25], in the context of probabilistic wind speed forecasts. Later, [32] analyzes the Variogram score in the context of probabilistic solar power scenarios.

The Variogram score metric is formally defined as

$$VS = \sum_{m,n=1}^d w_{mn} \left(|z_m - z_n|^p - E|\hat{Z}_m - \hat{Z}_n|^p \right)^2 \tag{5}$$

where \mathbf{z} denotes the observation (i.e., actual) vector of length d , \hat{Z}_m and \hat{Z}_n denote the m -th and n -th component of a random vector $\hat{\mathbf{Z}}$ (which here represents the scenario set), w_{mn} denotes non-negative weights, and p denotes the order of the variogram. Given a set of J scenarios $\hat{\mathbf{z}}^{(1)}, \dots, \hat{\mathbf{z}}^{(J)}$, the forecast variogram $E|\hat{Z}_m - \hat{Z}_n|^p$ can be approximated by

$$E|\hat{Z}_m - \hat{Z}_n|^p \approx p_j \sum_{j=1}^J |\hat{z}_m^{(j)} - \hat{z}_n^{(j)}|^p \tag{6}$$

for $m, n \in \{1, \dots, d\}$ and where p_j denotes the probability of scenario $j \in J$. We evaluate the Variogram score using values of $p \in \{0.5, 1, 2\}$. We also use two choices for the value of the weights, w_{mn} : (1) equal weights across all hours of the day and (2) weights set to the normalized average correlation of the actual wind production values between hours m and n .

Variogram score (std. dev.)	Correlated Weights			Equal Weights		
	$p = 0.5$	$p = 1$	$p = 2$	$p = 0.5$	$p = 1$	$p = 2$
Multivariate	0.355 (0.216)	0.211 (0.171)	0.074 (0.104)	0.379 (0.239)	0.237 (0.195)	0.088 (0.122)
Non-Parametric	0.351 (0.242)	0.192 (0.189)	0.064 (0.104)	0.372 (0.264)	0.217 (0.215)	0.077 (0.121)
Paired t-test p-values	6.32E-01	1.39E-04	3.20E-06	3.65E-01	2.11E-04	2.61E-05

Table IV. Average daily Variogram scores and standard deviations for both correlated and equal weights, evaluated for $p \in \{0.5, 1, 2\}$. The evaluation period covers January 2013 through September 2013.

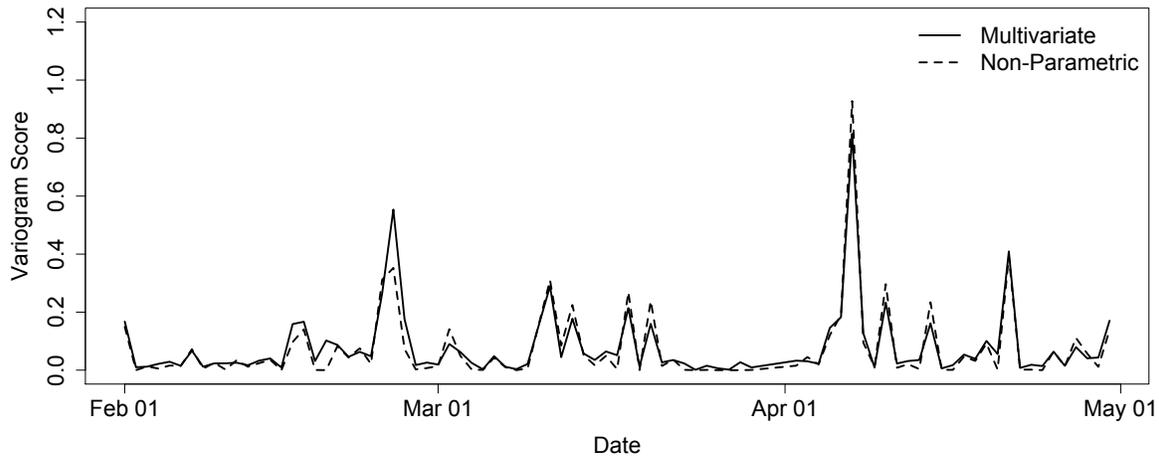


Figure 8. Variogram scores for probabilistic wind power scenarios from February 2013 through April 2013, using order $p = 2$ and correlated weights.

We report statistics for the daily Variogram scores for both the multivariate and non-parametric scenarios in Table IV, using both weighting schemes and a range of p . Figure 8 depicts the Variogram score over a three-month period for both methods using correlated weights and an order $p = 2$. The results demonstrate that the non-parametric scenarios are qualitatively and statistically better (for $p = 1$ and $p = 2$) than the multivariate scenarios, indicating their strength in better capturing temporal correlations presented in real wind power data.

4.8. Integrated Distance

Finally, we consider the Integrated Distance metric, a very simple method to compare probabilistic scenarios with an observed trajectory. Informally, the Integrated Distance metric sums the absolute value of the difference between each observed wind power quantity and the corresponding quantity in a particular probabilistic scenario. The resulting sum is then weighted by probability, and the probability-weighted sums for all scenarios are aggregated. Like the Energy score, the Integrated Distance metric is both proper and negatively oriented. The Integrated Distance metric is formally defined as

$$ID = \sum_{j=1}^J p_j \left(\sum_1^{24} |\hat{z}^{(j)} - z| \right) \tag{7}$$

where J denotes the number of scenarios and p_j denotes the probability of scenario j . Although this metric is not standard in the literature, it does provide a straightforward and intuitive method to evaluate the distance between a probabilistic scenario set and a corresponding observed trajectory. We report the mean and standard deviation of the Integrated Distance metric for both methods in Table V, while Figure 9 depicts the Integrated Distance metric over a period of three months of data. Under the Integrated Distance metric, the non-parametric scenarios perform noticeably better than the multivariate scenarios. The large spikes that are often seen in the multivariate scenarios contribute to higher Integrated Distance

Method	Integrated Distance Metric (std. dev.)
Multivariate	2.29 (1.16)
Non-Parametric	1.89 (1.00)
Paired t-test p-value	2.56E-65

Table V. Integrated Distance averages and standard deviations for probabilistic wind power scenarios from February 2013 through April 2013.

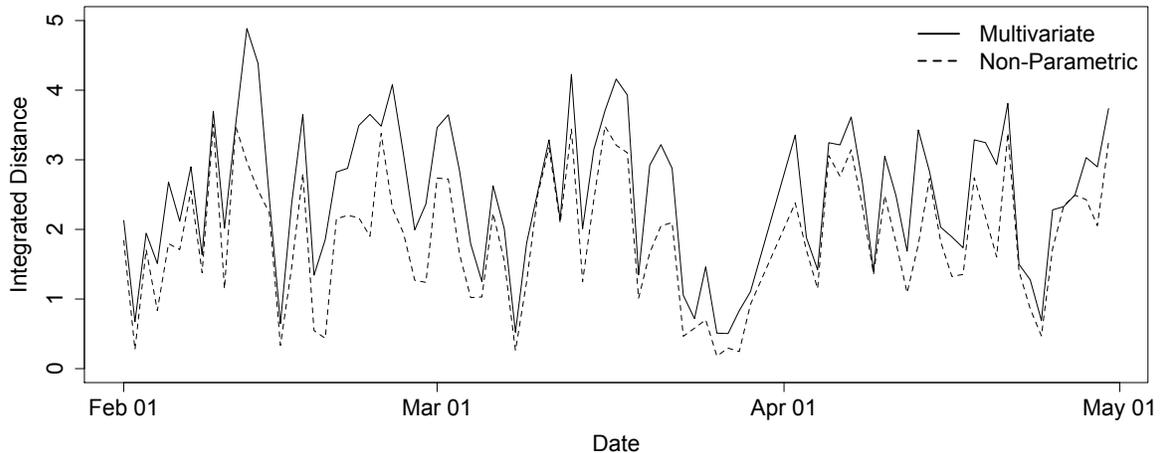


Figure 9. Integrated Distance plotted from February through April, 2013.

quantities, such that the metric is better able to differentiate the performance of the two types of probabilistic scenario sets.

5. CONCLUSION

The non-parametric method we detail here generates high quality probabilistic wind power scenarios that perform well when compared to the current state-of-the-art method, considering both qualitative features and quantitative quality metrics. In contrast to the current state-of-the-art method, our non-parametric scenario generation method is designed specifically for use with stochastic power systems operations planning models, e.g., unit commitment and economic dispatch, which are executed by power system operators hours to days in advance. Specifically, we emphasize the ability to specify scenarios in terms of forecast error distribution quantiles, allowing us to generate scenario sets of minimal size by avoiding sampling. This feature is critical when solving stochastic operations planning models, as their computational difficulty scales in proportion to the number of scenarios considered. Depending on the application, the errors from multiple regions or facilities could be incorporated as well, with appropriate correlation structures captured if needed.

Our non-parametric scenarios perform well in comparison to the multivariate scenarios generated using a state-of-the-art method when assessed using Energy score, Brier score, Variogram score, and Integrated Distance metrics. The MST rank histogram is not an applicable measure for assessing our non-parametric scenarios, as they are not drawn from a distribution. Additionally, we argue that our non-parametric scenarios are visually more appealing and realistic than multivariate scenarios, although these differences are not well-captured by all quantitative metrics. Our non-parametric method does not produce the sharp spikes and large deviations often found in multivariate scenarios, and the resulting time series more closely resemble the behavior of actual wind power observations. The application presented here considers an aggregated set of wind farms, with sharp ramps frequently smoothed out across farms due to geographic dispersion. For this application, our non-parametric scenarios are smoother and more realistic-looking, and more closely match the ramp characteristics of actual wind power. A qualitative difference is that our method produces scenarios that are naturally tree

structured, a feature that could be exploited in multi-stage stochastic optimization – which is central in advanced stochastic unit commitment and economic dispatch models.

In summary, our method offers a fundamentally different approach to probabilistic scenario generation for short-term wind power. We emphasize the critical role of domain-specific segmentation when fitting our forecast error distributions and the need for non-parametric distribution estimates. Our methodology is such that (1) parameters and constraints of the underlying optimization models can be adjusted based on qualitative knowledge of local wind conditions, (2) cut points and scenario “spread” can be adjusted systematically, without sampling, to achieve sufficient coverage of expected wind conditions, and (3) specific knowledge of local-area wind can be taken into account. Ultimately, our probabilistic scenarios are designed for use in stochastic power systems operations planning problems, and the real test of their value will be their performance – in terms of both operations cost and system reliability – in such settings. This analysis is computationally intensive, and the results are beyond the present scope. Thus, we leave this particular assessment to future work.

ACKNOWLEDGEMENTS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. This work was funded by the U.S. Department of Energy’s Advanced Research Projects Agency - Energy (ARPA-E), under the Green Energy Network Integration (GENI) project portfolio, and by the Bonneville Power Administration (BPA).

REFERENCES

1. Energy Information Administration. U.S. Energy Information Administration Wind Data. URL www.eia.gov/cneaf/solar.renewables/page/wind/wind.html.
2. Takriti S, Birge J, Long E. A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems* 1996; **11**(3):1497–1508.
3. 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.
4. Feng Y, Rios I, Ryan SM, Spurkel K, Watson JP, Wets RJB, Woodruff DL. Toward scalable stochastic unit commitment. part 1: Load scenario generation. *Energy Systems* 2015; **6**(3):309–329.
5. Cheung K, Gade D, Silva-Monroy C, Ryan SM, Watson JP, Wets RJB, Woodruff DL. Toward scalable stochastic unit commitment. part 2: Solver configuration and performance assessment. *Energy Systems* 2015; **6**(3):417–438.
6. Papavasiliou A. Coupling renewable energy supply with deferrable demand. PhD Thesis, University of California Berkeley 2011.
7. Royset JO, Wets RJB. From data to assessments and decisions: Epi-spline technology. *INFORMS Tutorial* 2014; URL <http://hdl.handle.net/10945/41758>.
8. Wood AJ, Wollenberg BF, Sheble GB. *Power Generation, Operation and Control*. 3rd edn., Wiley-Interscience, 2013.
9. Bloom A, Townsend A, Palchak D, Novacheck J, King J, Barrows C, Ibanez E, O’Connell M, Jordan G, Roberts B, *et al.*. Eastern renewable generation integration study. *Technical Report*, NREL (National Renewable Energy Laboratory (NREL), Golden, CO (United States)) 2016.
10. Morales JM, Minguez R, Conejo AJ. A methodology to generate statistically dependent wind speed scenarios. *Applied Energy* 2010; **87**(3):843–855.
11. Wang J, Shahidehpour M, Li Z. Security-constrained unit commitment with volatile wind power generation. *IEEE Transactions on Power Systems* 2008; **23**(3):1319–1327.

12. Rios I, Wets RJB, Woodruff DL. Multi-period forecasting and scenario generation with limited data. *Computational Management Science* 2015; **12**(2):267–295.
13. Feng Y, Gade D, Ryan SM, Watson JP, Wets RJB, Woodruff DL. A new approximation method for generating day-ahead load scenarios. *IEEE Power and Energy Society General Meeting (PES)*, IEEE, 2013; 1–5.
14. Sari D, Lee Y, Ryan S, Woodruff D. Statistical metrics for assessing the quality of wind power scenarios for stochastic unit commitment. *Wind Energy* 2016; **19**(5):873–893, doi:10.1002/we.1872. URL <http://dx.doi.org/10.1002/we.1872>.
15. Royset JO, Wets RJB. Fusion of hard and soft information in nonparametric density estimation. *European Journal of Operational Research* 2013; **247**(2):532–547.
16. Tastu J, Pinson P, Kotwa E, Madsen H, Nielsen HA. Spatio-temporal analysis and modeling of short-term wind power forecast errors. *Wind Energy* 2011; **14**(1):43–60.
17. Bessa RJ, Miranda V, Botterud A, Zhou Z, Wang J. Time-adaptive quantile-copula for wind power probabilistic forecasting. *Renewable Energy* 2012; **40**(1):29–39.
18. Al-Awami AT, El-Sharkawi MA. Statistical characterization of wind power output for a given wind power forecast. *North American Power Symposium (NAPS)*, 2009, 2009; 1–4, doi:10.1109/NAPS.2009.5484044.
19. Bludszuweit H, Domínguez-Navarro JA, Llombart A. Statistical analysis of wind power forecast error. *IEEE Transactions on Power Systems* 2008; **23**(3):983–991.
20. Bofinger S, Luig A, Beyer H. Qualification of wind power forecasts. *2002 Global Windpower Conference, Paris*, vol. 2(5), 2002.
21. Bonneville Power Administration. Wind projects 2013. URL <http://www.bpa.gov/transmission/Projects/wind-projects/Pages/default.aspx>.
22. Bonneville Power Administration. Wind power forecasting data 2014. URL <http://www.bpa.gov/Projects/Initiatives/Wind/Pages/Wind-Power-Forecasting-Data.aspx>.
23. Pinson P, Girard R. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy* 2012; **96**:12–20.
24. Pinson P, Tastu J. Discrimination ability of the energy score. *Technical Report* 2013; URL http://orbit.dtu.dk/fedora/objects/orbit:122326/datastreams/file_b919613a-9043-4240-bb6c-160c88270881/content.
25. Scheuerer M, Hamill TM. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review* 2015; **143**(4):1321–1334.
26. Thorarinsdottir TL, Scheuerer M, Heinz C. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *arXiv preprint arXiv:1310.0236* 2013; .
27. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 2007; **102**(477):359–378.
28. Gneiting T, Stanberry LI, Grit EP, Held L, Johnson NA. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST* 2008; **17**(2):211–235.
29. Wilks DS. *Statistical methods in the atmospheric sciences*, vol. 100. Academic press, 2011.
30. Wilks DS. The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review* 2004; **132**(6):1329–1340.
31. Thorarinsdottir TL, Scheuerer M, Heinz C. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics* 2016; **25**(1):105–122.
32. Golestaneh F, Gooi H, Pinson P. Generation and evaluation of space-time trajectories of photovoltaic power. *Applied Energy* 2016; **176**:80–91.