Improving Wind Power Prediction Intervals Using Vendor-Supplied Probabilistic Forecast Information

Sabrina Nitsche
University of Duisburg-Essen, Germany
César A. Silva-Monroy, Andrea Staid, and Jean-Paul Watson
Sandia National Laboratories, Albuquerque, NM
Scott Winner
Bonneville Power Administration, Portland, OR
David L. Woodruff
University of California Davis, Davis, CA

***Abstract*—We describe experiments concerning enhancing a simple, yet effective method to compute high-accuracy prediction intervals (PIs) for day-ahead wide area wind power forecasts. The resulting PIs are useful for operators and traders, to improve reliability, anticipate threats, and increase situational awareness.**

I. Introduction

The potential value of probabilistic forecasts for wind power, particularly in the short term, has been discussed for some time (see, e.g., [1]). Recent research describes methods for generating wide area probabilistic forecasts as the basis for constructing scenarios for use in day-ahead and hours-ahead commitment of thermal generating units and dispatch of energy (see, e.g., [2]–[5])

The Bonneville Power Administration (BPA) employs probabilistic forecasts for a different, albeit related, purpose: prediction intervals for day-ahead wind power forecasts. These intervals assist operators and traders to anticipate threats, identify market opportunities, and generally enhance situational awareness for power system operation. In this paper, we begin by describing a simple, yet effective method to compute prediction intervals for wide area wind power using data that is readily available to utilities and balancing authorities. We then describe a data-driven enhancement of our base method that makes use of a proxy for weather variability and stability, to improve overall accuracy.

Specifically, we describe methods to compute (100-α)% prediction intervals (PI) for wide area wind power forecasts that are linked to *actual* generation. The quality of our PIs is assessed by various metrics, based on the observation that measured wind power quantities should reside within the PI exactly (100-α)% of the time. In contrast, traditional vendor-supplied wind power forecast PIs are based on the NWP (Numerical Weather Prediction)-generated *forecast* traces used to calculate projected wind power. At BPA, PIs are issued hourly for the next 168 hours, including the 24 hours of the day-ahead planning window. An example of a PI generated by our method for BPA is shown in Figure 1.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94-AL85000. This work was funded by the Bonneville Power Administration (BPA).

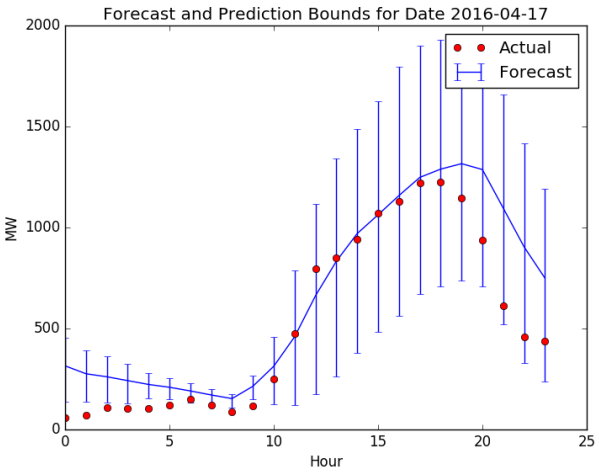


Figure 1. An example of a 70% prediction interval for forecasted BPA wind power. The PI is created day-ahead. Actual (measured) quantities are depicted as points.

II. Computing Prediction Intervals

To describe our method, it is useful to introduce the notion of a date-time pair for which forecasted wind power data are available. In the computational experiments described below, times are always hours of the day, although our method is general and can be applied using data with an arbitrary time resolution. For each date-time pair, we define the *forecast error* (often simply referred to as the *error*) as the difference between the observed (measured) wind power and the forecast. We assume the availability of a historical database of forecasts and corresponding observations, which is generally maintained by system operators.

We compute errors and construct an empirical non-parametric error distribution based on the historical data. We then estimate various empirical *order statistics*. The $\frac{\alpha}{200}$ and $1-\frac{\alpha}{200}$ order statistics provide the estimates of the prediction interval (PI), which is placed around the "point" forecast of a future date-time pair in order to compute PI limits. In cases where very wide PIs (e.g., 99% or higher) are desired, it may be necessary to fit an analytic (but still non-parametric) error distribution when there is insufficient data, such that empirical order statistics are unreliable estimates. In our motivating use

case, 70% intervals are desired (i.e., $\alpha = 0.3$) and empirical order statistics provide effective estimates.

All historic date-time pairs for which there is a reported forecast and a corresponding measured quantity are referred to as the *pre-window*; this allows for missing / dropped quantities. In general, we do not use all of the data in the pre-window. Instead, we use categorization methods to down-sample the data in the pre-window, identifying only the most relevant date-time pairs for the current forecast and computing order statistics and PIs based on this data. The specifics of our categorization methods are described subsequently.

A. Categorization by Megawatt Window

It is well known that many statistical procedures involving wind power benefit from either segmenting the data by power level (see, e.g., [6], [7], which are in turn based on [8]) or by transforming the data to stabilize the variance across power levels (see, e.g., [9]). Here, we consider segmentation methods because they are consistent with our proposed enhancements to compute PIs. When segmenting by wind power, we only use data in the pre-window for which the *historical* forecast lies inside a *MW window* around the corresponding next-day forecasted wind power quantity.

To determine the MW window, we estimate the (normalized) empirical order statistic function – denoted "eos" – of the forecasts in the pre-window. The function $\text{eos}(x; P)$ returns the i^{th} value in the ordered set P for $i \in \{1, \ldots, |P|\}$ if $x = i/|P|$ and an interpolated value otherwise; $|P|$ denotes the number of elements in P. By convention, for all P, $\text{eos}(0; P)$ equals zero and $\text{eos}(1; P)$ equals the installed wind power capacity. We denote the inverse of this function by eos^{-1}, with $\text{eos}^{-1}(x) = 1$ for all x greater than or equal to the installed capacity and $\text{eos}^{-1}(x) = 0$ for $x \leq 0$.

An interval with a *category width* $s_{mw} \in [0, 1]$ is placed around the normalized empirical order statistic of the forecasted wind power quantity v. The category width dictates the number of data points in the MW window. If there are W points in the pre-window, the MW window should contain approximately $s_{mw}W$ points, with half of the points both smaller and larger than the forecasted quantity. Because the window implied by s_{mw} may extend below zero or above the maximum power, we may obtain fewer data points than expected. Finally, the corresponding forecasts of the prediction interval limits are computed by inverting the empirical order statistics to obtain the limits for the MW window. The lower bound of the window is given as

$$eos(\max\left\{0, \text{eos}^{-1}(v; P) - \frac{s_{mw}}{2}\right\}; P)$$

and the upper bound is given as

$$eos(\min\left\{1, \text{eos}^{-1}(v; P) + \frac{s_{mw}}{2}\right\}; P)$$

B. Categorization by Vendor Prediction Interval Width

Categorization by power level provides reasonably good prediction intervals, but incorporating information about the stability of the weather and the weather forecast improves quality further, as as we will show. Vendor-supplied power forecasts provided to BPA and numerous other system operators provide estimated PIs based on ensembles of weather forecast models, in addition to their "point" forecasts of wind power. The idea of using model ensembles to create probabilistic forecasts has been discussed for some time (see, e.g., [10]) and the details are beyond the scope of the present work. Our point here is that we can use the width of these vendor-supplied PIs to further improve categorization. If the width of the vendor-supplied PI for the forecasted date-time pair is small, we only consider in the pre-window those historic date-time pairs for which the PI width is also small; if it is large, we only consider date-time pairs with wide PIs. Informally, we consider vendor-supplied PIs as a proxy for weather stability and model confidence.

We estimate the empirical order statistics of the widths of vendor-supplied PIs for those date-time pairs in the appropriate MW window, as identified via the segmentation procedure described above. We then place an interval around $\text{eos}(w; P)$ using a category width $s_{vpi} \in [0, 1]$. The category width parameter is intended to control for the number of points in the window, in the nominal case. If there are W points in the MW window, the *vendor PI window* contains at most $s_{vpi}W$ points. Nominally, half of these points correspond to PIs that are smaller and larger than the width of the given PI. However, there are cases other than the nominal case. For example, there may be fewer than $s_{vpi}W/2$ points on either side because the vendor-supplied forecast PI width may be near the largest or smallest value for date-time paris contained in the MW window. Finally, we compute the corresponding widths of the interval limits by inverting the empirical order statistics, yielding the limits of the width window.

C. Implementation

To rigorously describe exactly how empirical order statistics are computed, we temporarily drop explicit specification of the list P from the function eos (eos with P implicit). Consider a list of values $[x_1, \ldots, x_n]$ with $x_1 < x_2 < \cdots < x_n$. The order of x_i, $i = 1, \ldots, n$, then is i and the probability that the outcome is smaller or equal to x_i is $\frac{i}{n+1}$. Now consider $x \in \Re$. If $x = x_i$ for an $i \in \{1, \ldots, n\}$, we have $\text{eos}(x) = \frac{i}{n+1}$. If $x_i < x < x_{i+1}$, the probability that the outcome is smaller than or equal to x is

$$\begin{aligned} \text{eos}(x) &= \text{eos}(x_i)\frac{x_{i+1} - x}{x_{i+1} - x_i} + \text{eos}(x_{i+1})\frac{x - x_i}{x_{i+1} - x_i} \\ &= \frac{i(x_{i+1} - x) + (i+1)(x - x_i)}{(n+1)(x_{i+1} - x_i)}. \end{aligned}$$

If $x < x_1$ we take the linear function

$$g(t) = \frac{1}{(n+1)(x_2 - x_1)}t + \frac{x_2 - 2x_1}{(n+1)(x_2 - x_1)}$$

through the points $\left(x_1, \frac{1}{n+1}\right)$ and $\left(x_2, \frac{2}{n+1}\right)$ and set

$$\text{eos}(x) = \max\{0, g(x)\}.$$

Now assume $x > x_n$. We take the linear function

$$h(t) = \frac{1}{(n+1)(x_n - x_{n-1})} t + \frac{n(x_n - x_{n-1}) - x_n}{(n+1)(x_n - x_{n-1})}$$

through the points $\left(x_{n-1}, \frac{n-1}{n+1}\right)$ and $\left(x_n, \frac{n}{n+1}\right)$ and set

$$\text{eos}(x) = \min\{h(x), 1\}.$$

To estimate the empirical order statistics we require at least 20 data points. Because we require at least 72 data points in the pre-window, we are always able to estimate the empirical order statistics for the computation of the MW window. If there are less than 20 data points in the MW window or the vendor PI window, we increase the single category width for the window until there are 20 data points in that window. Then we are able to compute the empirical order statistics of either the errors for computing the prediction intervals or the widths for computing the PI width window.

Note that it is a little confusing to discuss the vendor PI width window because of two uses of the word "width." The window has a width, but it is being constructed based on the width of the vendor-supplied prediction interval. To help with this we refer to points selected by the categorization as being in the window and use "window width" to describe the width of the window used to select the points. This window cordons off a section of the vendor-supplied PI width values.

If we categorize by the width of the vendor-supplied prediction intervals and there are fewer than 20 data points in the resulting window, we increase the single category width for the window until there are 20 data points in it. Since we start the computation of the vendor PI window with at least 20 data points, we can be sure that there is a single category width for the PI width window that fulfills this property. In the worst case, the category width will be incremented until it is big enough to cover the whole MW window.

III. Computational Experiments

Our tests use data from the Bonneville Power Administration (BPA). The BPA is a federal Power Market Authority that owns 75% of the installed transmission in the U.S. Pacific Northwest. A good wind resource and easy access to transmission has resulted in the development of 33 wind generation facilities (also known as projects) – presently ~4,500 MW – in the BPA service territory. Historical wind power forecasts and actual generation values from both individual BPA wind projects and the aggregate fleet were made available to the research team. The BPA's Centralized Wind Power Forecasting Initiative utilizes two commercial wind power forecasting vendors. Vendor forecasts are evaluated for quality and the better of the two is published to BPA systems as the BPA Official Forecast. The second forecast serves a reliability/back up function. The focus of this research is on the forecasts provided by the primary BPA vendor and include the average or expected generation values and associated prediction intervals.

To evaluate the quality of our computed prediction intervals, we perform a rolling horizon *re-enactment*. By re-enactment, we refer to a walk forward through date-times in the past, computing prediction intervals using only data available prior to that date-time. In doing so, we compute prediction intervals using only relevant historical information, and are able to assess prediction interval quality using actual observations not used in the computation of those prediction intervals. In all of our tests, we consider 70% prediction intervals, to mirror existing BPA practice.

A. Evaluation of Prediction Intervals

To assess the quality of our computed prediction intervals we evaluate them based on their skill and sharpness [11], using re-enactments where we determine what would have happened had we been using our prediction intervals in actual operations. Skill refers to the degree to which the fraction of new observations that fall inside a prediction interval matches $1 - \alpha$. Sharpness refers to the average width (in MW) of a prediction intervals.

We compute the percentage of the date-times where the observed (actual) value lies outside the prediction interval. An observed value lies outside of the prediction interval on the "left side" ("right side") if and only if the observed value is strictly smaller (larger) than the lower (upper) limit. These percentages indicate how well our prediction intervals contain the observed values. Our objective is to generate prediction intervals for which left and right percentages are as close as possible to $\frac{\alpha}{2}\%$. As we show below, prediction interval behavior can be significantly different on the left and right sides. Thus, we choose to report and analyze the left and right quantities separately. Finally, we use the average width of prediction intervals to quantify sharpness. If prediction intervals are too wide, they lose their utility for situational awareness and operations planning.

B. Numerical Results

We use BPA data ranging from 11/01/2015 to 04/24/2016. Our rolling horizon re-enactment starts on 02/01/2016, such that sufficient historical data is available. The installed wind BPA capacity for this period is 4500 MW.

BPA vendors provide forecasts and prediction intervals (based on NWP forecast trace statistics) for a 168 hour rolling interval. Because we are focused on day-ahead prediction intervals, we use the forecasts for the hours of day d that are released at 11:00 AM on day $d-1$. In the experiments below, we report absolute error quantities as our analysis indicates that there is no advantage to using the relative error.

We summarize the quality of our prediction intervals computed both with and without categorization by vendor-supplied prediction interval widths in Tables I and II . They record the (1) the percentage of points that fall outside of the prediction intervals on both sides of the forecast; an ideal value would be 15, and (2) the average width of the prediction intervals both overall and broken out by the width below (left) and above (right) the prediction; obviously, narrower is better.

In Table I, we observe that the prediction intervals computed with our method – both with or without categorization by vendor supplied PI width – are significantly better in terms of

both skill and sharpness than the vendor supplied prediction intervals (official). In Table II, we see that both the skill and sharpness of the prediction intervals computed with our approach are significantly improved when categorizing by vendor-supplied PI widths. In particular, the left side skill is significantly worse when vendor-supplied PI width categorization is not employed.

Table I
EVALUATION OF 70% PIS COMPUTED WITHOUT VENDOR-SUPPLIED PI WIDTH CATEGORIZATION

	$s^*_{mw} = 0.1$	$s^*_{mw} = 0.2$	Official
Out left (%)	11.54	11.54	22.68
Out right (%)	14.72	14.55	6.37
Avg width (MW)	709.67	705.32	980.27
Avg width left (MW)	378.72	371.09	454.17
Avg width right (MW)	330.95	334.23	526.10

Table II
EVALUATION OF 70% PIS COMPUTED WITH VENDOR-SUPPLIED PI WIDTH CATEGORIZATION

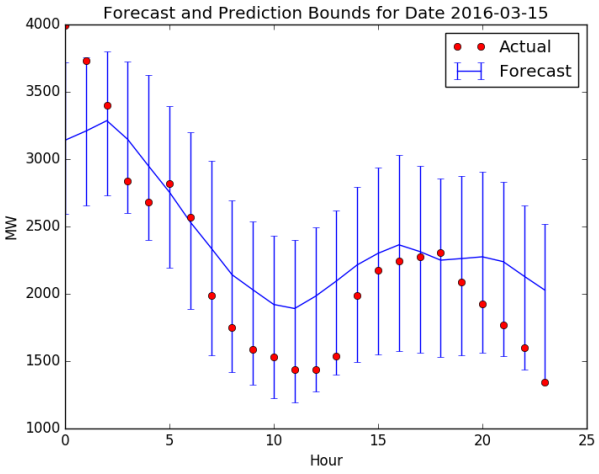
	$s^*_{mw} = 0.4$		$s^*_{mw} = 0.5$	
$s^\#_{vpi} =$	0.4	0.5	0.4	0.5
Out left (%)	14.19	13.90	14.66	14.72
Out right (%)	14.90	14.96	15.02	14.90
Avg width (MW)	715.02	716.43	712.71	713.83
Avg width left (MW)	377.83	379.12	375.66	376.43
Avg width right (MW)	337.18	337.31	337.05	337.40

* category width for the categorization by MW window.
# category width for the categorization by the width of vendor supplied approximate prediction intervals.

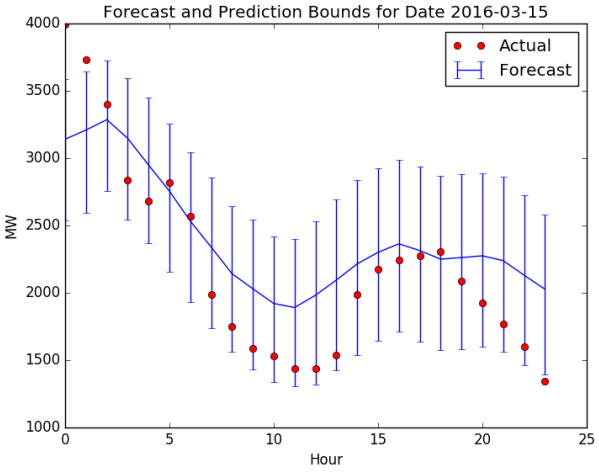
Overall, our computed prediction intervals with vendor-supplied PI width categorization have good skill on both the left and right sides. All of our computed prediction intervals have a smaller average width and better skill than the approximate prediction intervals provided by the the official values generated by BPA (in the column labeled "Official").

Figures 2 and 3 show prediction intervals for the days 03/15/2016 and 04/17/2016, respectively. The wind power that was actually observed for each hour are shown as red dots. In each figure, subplot (a) on top shows the prediction intervals that are computed without categorization by vendor-supplied PI widths and with categorization by MW window using s_{mw}=0.1. The bottom subplot (b) shows the prediction intervals that are computed using categorization by vendor-supplied PI widths using s_{vpi}=0.4 and MW window categorization using s_{mw}=0.5. Both the s_{mw} and s_{vpi} parameters used were the best obtained after limited experimentation.

On 03/15/2016, we observe that the prediction intervals computed using vendor-supplied PI width categorization are somewhat tighter than the prediction intervals computed using only MW-based categorization, but they still approximately contain (with the exception of hour 1 and 24, which are slightly outside the corresponding PI) the observations. On 04/17/2016, we observe that the prediction intervals computed using vendor-supplied PI width categorization are tighter than those computed using only MW window-based categorization, such that the lower bound of the prediction interval is greater



(a) Computed prediction intervals for 2016-03-15, without vendor prediction interval widths; s_{mw} = 0.1.



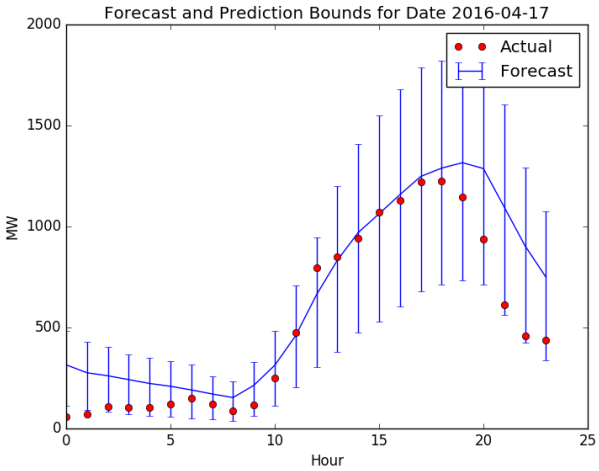
(b) Computed prediction intervals for 2016-03-15, with vendor prediction interval widths; s_{mw} = 0.5 and s_{vpi} = 0.4.

Figure 2. Prediction Intervals for 2016-03-15

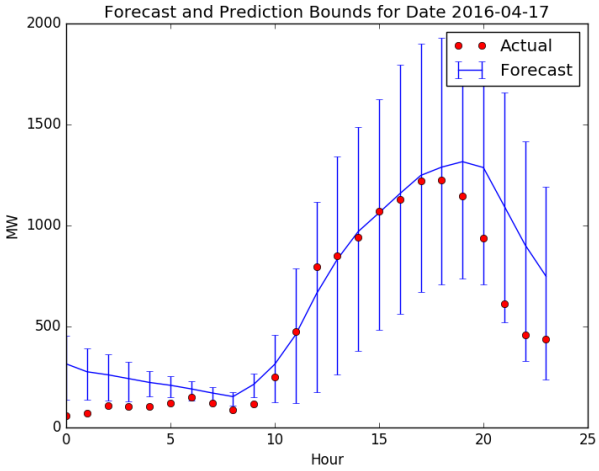
than the observation for hours 1 through 5. This result illustrates why the left-side skill of our computed prediction intervals improves if we use vendor-supplied PI width categorization in addition to MW window-based categorization. Further, we see that the prediction intervals computed with the vendor-supplied PI width categorization are wider (less sharp) than those computed using only MW window-based categorization, for the hours in the second half of the day.

IV. CONCLUSION

We have described a novel method for computing prediction intervals for forecasted wide area wind power, using data from BPA for experimental testing. By analyzing historical error distributions and leveraging straightforward MW window-based categorization, we are able to significantly improve over the skill and sharpness provided by vendor-supplied prediction intervals. We can further improve the accuracy



(a) Computed prediction intervals for 2016-04-17, without vendor prediction interval widths, s_{mw} = 0.1



(b) Computed prediction intervals for 2016-04-17, with vendor prediction interval widths; s_{mw} = 0.5 and s_{vpi} = 0.4.

Figure 3. Prediction Intervals for 2016-04-17

of our method by adding categorization based on the width of vendor-supplied prediction intervals –which provide some indication of the stability of a particular forecast. Using both categorization schemes, we are able to obtain prediction intervals with very high skill, e.g., within 0.5% on both the low and high ends of 70% prediction intervals. Further, our prediction interval widths (sharpness) are significantly improved over that obtained by the vendor prediction intervals.

Our method leverages both historical forecasts and corresponding actuals readily available to a system operator, in addition to prediction interval data provided by many forecasting vendors. Thus, integration into existing system operations is relatively straightforward.

A number of topics remain as future research. One major practical issue involves how to best display prediction intervals, specifically while varying α. One possibility is shown in

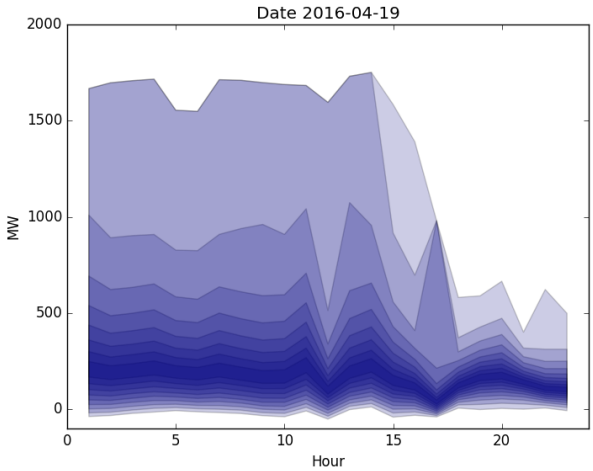


Figure 4. Prediction intervals for wind power forecast at various values of α, on a single graphic. Lighter colors correspond to lower α values (i.e., larger % PI).

Figure 4, in which we show prediction intervals over various α quantities. Further, there is the challenge of extending our methodology to the computation and display of prediction intervals for solar power and other sources of uncertainty.

References

[1] P. Pinson, C. Chevallier, and G. N. Kariniotakis, "Trading wind generation from short-term probabilistic forecasts of wind power," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1148–1156, Aug 2007.

[2] P. A. Ruiz, C. R. Philbrick, and P. W. Sauer, "Wind power day-ahead uncertainty management through stochastic unit commitment policies," in *Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES*, March 2009, pp. 1–9.

[3] A. Papavasiliou and S. Oren, "A stochastic unit commitment model for integrating renewable supply and demand response," in *Proceedings of the 2012 IEEE Power and Energy Society Meeting*, 2012.

[4] Y. Feng, D. Gade, S. M. Ryan, J.-P. Watson, R. J. Wets, and D. L. Woodruff, "A new approximation method for generating day-ahead load scenarios," in *Proceedings of the 2013 IEEE power and energy society general meeting*, 2013.

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

[6] A. T. Al-Awami and M. A. El-Sharkawi, "Statistical characterization of wind power output for a given wind power forecast," in *North American Power Symposium (NAPS), 2009*, Oct 2009, pp. 1–4.

[7] H. Bludszuweit, J. A. Domínguez-Navarro, and A. Llombart, "Statistical analysis of wind power forecast error," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 983–991, 2008.

[8] S. Bofinger, A. Luig, and H. Beyer, "Qualification of wind power forecasts," in *2002 Global Windpower Conference, Paris*, vol. 2(5), 2002.

[9] D. L. Woodruff and G. Slevogt, "Variance stabilizing transformation of wind forecast errors," *Wind Energy*, vol. 19, no. 10, pp. 1845–1852, 2016, we.1954. [Online]. Available: http://dx.doi.org/10.1002/we.1954

[10] J. L. Anderson, ":a method for producing and evaluating probabilistic forecasts from ensemble model integrations," *J. Climate*, vol. 9, pp. 1518–1530, 1996.

[11] P. Pinson, G. Kariniotakis, H. A. Nielsen, T. S. Nielsen, and H. Madsen, "Properties of quantile and interval forecasts of wind generation and their evaluation," in *Proceedings of the European Wind Energy Conference & Exhibition*, Athens, 2006, http://www.ewea.org. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4250