

Forecaster comments to the ORTECH Report

The Alberta Forecasting Pilot Project was truly a pioneering and landmark effort in the assessment of wind power production forecast performance in North America. There had never been such a comprehensive evaluation and comparison of real-time forecasts provided by multiple forecasters for wind parks in North America before the Alberta Pilot Project. It will, no doubt, provide interested parties with a tremendous amount of information about and insight into the performance of wind power production forecasts in North America. However, as in most pioneering investigations, there are important limiting factors that one should consider when interpreting the results. It should be noted that many aspects of forecast performance results and the comparison of the performance among the forecast providers was most likely significantly impacted by (1) wind park data quality issues, (2) the lack of a stated forecast performance objective or specific set of intended forecast applications, (3) the methods and metrics used to evaluate the performance, and (4) the different approaches taken by the forecast providers regarding the evolution of their forecast systems during the project.

Specific Comments

1. Section 2, page 11 states

Tables 2-1 to 2-4 provide the monthly and total 10-minute and hourly averaged measured wind speed and measured/calculated power data recovery rates (prior to employing the screen-out criteria detailed in section 2.1) for the 4th quarter (Q4) from February 1, 2008 to April 30, 2008, respectively. Similar tables for the 1st quarter (Q1), 2nd quarter (Q2) and 3rd quarter (Q3) are shown in Appendix C.

The data recovery rates specified in this table are, in general, not the same as the data recovery rates experienced by the forecasters at the time the forecasts were produced. That is, some of the data was not available from Genivar in near real time and was only available to ORTECH (and the forecasters) at a later date and time. Thus, in a number of cases, forecasts were produced for some sites and aggregates without the benefit of recent data from the wind parks. It is likely that this had a noticeable impact on the performance of forecasts for the first few of hours of the forecast look-ahead period since recent trends in power production are important predictors for these look-ahead periods.

Also, the quality control screening done by the forecasters was, in general, different from that used by Genivar and ORTECH and thus there is some difference of opinion on what data should have been included in the final dataset that was used by ORTECH for the evaluation of the forecasts. It is likely that a substantial amount of erroneous or unrepresentative measured data was included in the evaluation sample. The likely inclusion of these erroneous or unrepresentative measured meteorological or power production values in the evaluation sample most likely increased the forecast error (as measured by metrics such as the MAE to RMSE) to a level that was noticeably higher than what might have been reported had a more sophisticated quality control procedure been used to determine the content of the evaluation sample. One significant issue was

the reporting of turbine availability. The forecasters were not provided with any forecast of turbine availability and the actual turbine availability was only reported by two of the existing wind parks. The uncertainty in the forecasted and more importantly the actual turbine availability no doubt degraded the apparent performance of the forecasts as measured by the metrics compiled by ORTECH.

2. Section 2.1, last paragraph page 15 states:

To avoid misleading results ORTECH considered only eleven months (11) from June 2007 to April 2008, inclusive for the existing-facilities in the analyses of power (summarized in the following sections).

The use of the 11-month sample has the effect of placing a heavier weight on forecast performance during the cold season months in the overall project performance statistics. This is because all 6 cold season months were included in the evaluation sample but only 5 of the 6 warm season months were in the sample. The results reported by ORTECH (Section 3.3.2) indicate that forecast performance had difference characteristics in the cold and warm seasons. As a result, the cold season characteristics are more prominent in the overall results than they would be in a true 12-month sample. Since, for example, metrics such as MAE or RMSE (i.e. larger errors) are typically higher in the cold season, the annual results have a slight bias toward MAE or RMSE values than would be obtained from a true 12-month sample.

3. Section 3.1.1, Page 21 states: The South West region contains three meteorological masts. Therefore, three sets of wind data time series are available. These three data sets are combined together irrespectively (of dates/times). The error statistics are then calculated using the same method as for an individual site. With this approach averaged wind speed data from any individual time series are not derived.

This method is not consistent with the concept of the regional aggregation of the power production that is used in the evaluation of the power production forecasts. The “stacked” approach used by ORTECH for the wind speed error calculations provides a measure of the errors at the individual meteorological towers but not of the error in the overall wind speed for the wind parks in the region. That is, there is no “spatial smoothing” incorporated into the ORTECH regional wind speed error calculations because they are based on individual wind speed values at sites and not on a regional aggregate of wind speed. For example, the MAE or RMSE of the stacked wind data represents the typical error at an individual site in a region and not the typical error for the regional wind speed. It is the regional composite wind speed (weighted by the relative capacity of each park) that is more closely linked with the aggregated regional power production. This inconsistency will have an impact (probably modest) on any evaluation of the relationship between wind speed and power forecast errors.

4. Section 3.1.1, Page 26, last paragraph states:

Secondly, the SW region in relative size to the CE region is much smaller

and thus the individual sites are closer together than in the CE region and again may have an effect on forecasting based on topography and spatial smoothing.

There is no spatial smoothing of the wind speed errors because the stacking method used to compute the regional wind speed MAE or RMSE (as discussed in item #3 above) does not incorporate this effect (i.e. it measures the typical error at individual sites not the error of an aggregate forecast). However, it does provide a more regionally representative value for the typical wind park MAE or RMSE in a region than that provided by an individual wind park. It should be kept in mind that the average MAE or RMSE for a group of sites is not the same as the MAE or RMSE of the average (or other composite) wind speed due to the aggregation effect (i.e. random (uncorrelated) errors at individual sites in an aggregate will tend to partially offset each other yielding a lower MAE or RMSE for the aggregate forecast). Thus, none of the MAE or RMSE differences between the CE and SW region that are presented in this report are due to differences in the spatial smoothing, They are mostly due to the differences in terrain complexity (less complex in the CE region) and the absolute magnitude of the wind speed (lower in the CE region).

5. Section 3.1.2, Page 29, first paragraph states:

The accuracy of wind power prediction is dictated by different factors including:

- **The accuracy of wind speed prediction;**
- **The amplification and dampening of the wind speed prediction error through the nonlinear power curve; and**
- **Wind farm efficiency including the turbine availability and performance.**

Although wind speed is the most significant factor, the accuracy of wind power forecasts also depends on the accuracy of the forecasts for other meteorological variables such as wind direction and air density (temperature).

6. Section 3.3.1 (Hours of the Day) page 36 states:

The reason for the less accuracy in the afternoon periods can be accounted for by higher wind speeds and their variability which convert to higher power generation producing larger errors which can be explained by the power conversion curve (see section 3.5, 3.7).

The explanation may be more complex than higher wind speeds in the afternoon lead to higher wind speed forecast errors which result in larger power production forecast errors. Another factor may be that atmospheric events such as rapid transitions in boundary layer turbulent mixing regimes (e.g. from stable to well mixed or vice versa) or thunderstorms which can cause very large errors may be more numerous at this time of the day. A further investigation is needed to gain additional insight into this error pattern.

7. Section 3.3.2 (Seasons of the Year), page 36 states:

The reduced accuracy in the winter season could be due to higher wind speeds and an increased number of weather systems covering the total area during that time of year.

An analysis of power production forecast error by type of weather regime suggests that the magnitude of forecast errors has a significant dependence on the TYPE of weather regimes. This analysis suggested that many of the weather regimes that are associated with large forecast errors occur more frequently and often with greater amplitude in the winter. This may be a significant component of the explanation of the higher forecast errors during the winter.

**8. Section 3.5, page 40, top paragraph states:
The magnitude of error amplification due to power conversion is low relative to those reported in the literature (reference 3 in bibliography).**

The calculated error amplification rate is influenced by the methods used to calculate the regional wind speed MAE and the regional power MAE (see the discussion in item #4 above). The wind speed MAE is based on a stacked method (which essentially measures the average MAE for individual sites) whereas the power MAE is a regional aggregate (which measures the MAE of a regional aggregate). This will tend to underestimate the ratio of normalized power forecast MAE to normalized wind speed MAE since the wind speed MAEs for the individual sites are higher than for a regional aggregate wind speed. Calculations for one project quarter yielded a ratio of 1.33 when using the stacked method for wind speed and 1.51 when using a regional aggregate method, which represents about a 10% increase in the amplification rate. Another factor that may have contributed to a lower apparent amplification rate than reported in the literature is the average slope of the actual facility-scale power curve (i.e. the relationship between the meteorological tower wind speed and total facility power production). This is impacted by a number of factors including the degree of correlation of the wind speed among the turbine sites and the typical amount of turbines that are unavailable. If the wind speed correlation is low the facility-scale power curve will be less steep and the sensitivity of the power forecast to wind speed forecast error will be lower. This is because the uncorrelated variations in wind speed will tend to offset each other and the power production will rise more gradually as the wind speed at the meteorological tower increases. Many of the existing sites that participated in this project were in complex terrain which makes it likely that the wind speed correlation among the turbine locations was lower than for a typical wind park in less complex terrain. The frequent occurrence of turbine outages will also tend to flatten the facility-scale power curve (based on the reported output for a measured wind speed) unless the outages are reported and the power is adjusted to full availability. Since turbine outage reporting was very limited in this project, it is likely that this was a noticeable factor as well.

**9. Section 3.11, 2nd paragraph, page 57 states:
The following windows was initially proposed by ORTECH and then accepted by the working group for this assessment:
1. amplitude of the ramp event is within 80% - 120% of the actual measured**

one;

2. the event is forecasted not more than 12 hours in advance of the actual events:

3. the event is forecasted either 6 hours before or 6 hours after the actual event

It is later stated in this section:

Using the CSI methodology and the window criteria outlined above none of the forecasters did well in predicting the ramp rates as shown by the low percentages of the total ramp rates. However, it must be noted that the forecasters were not given this specific objective, i.e. “tune your models to predict rapid ramp rates” by the working group. It is ORTECH’s opinion the forecasters tuned their models to do well in predicting the general conditions.

More specifically, the forecasters, in general, configured their systems to produce the lowest value of an overall performance metric such as the RMSE or MAE over all forecast intervals. This has a significant impact on the prediction of ramps as defined by the methodology described in this report because the minimization of RMSE or to a less extent the MAE tends to produce a “hedged” forecast in which a ramp had a smaller amplitude and is often stretched out over a longer period of time. This minimizes the magnitude of the error in the frequently occurring situation in which the timing of a ramp event (i.e. a phase error) is slightly or modestly in error. This effect is particularly strong in an RMSE minimization approach since large errors are heavily penalized.

10. Section 3.13, page 59 states:

Are there times (day/month/weather pattern) when there is more uncertainty in the forecasts than other times?

Another approach would be to look at the patterns in the range of the forecasters estimation of the min/max values to see when the forecasters think there is more uncertainty. For example, the min/max range for each forecaster could be examined by the time of day or the month of the year.

11. Section 3.14, page 60 states:

Between 81% and 95% (depending on forecaster and forecast horizon) of the time the measured values fall in between predicted minimum and maximum power values.

There are two factors to consider when viewing the results in Section 3.14. First, there were no instructions given to the forecasters regarding the definition of the min and max values. Therefore, the definitions that were employed were not the same for all forecasters. Thus, the fraction of the time that the error fell into this range does not have much meaning with respect to the reliability of the ranges (do the measured values fall into the stated range at the stated frequency?). The second issue is related to sharpness.

Obviously one could use a range of 0 and 100% of capacity and all values would fall into this range. Thus, even though there is no stated probability for the min/max range, it would be interesting to have seen the average max/min ranges for each forecaster and time horizon. This would provide a crude indication of the sharpness.

12. Section 3.15, page 61 states:

What is the correlation factor between all three forecasts? Is this related to the forecast error? The focus has been on the evaluation of predictions issued by the forecasters against the measured power and wind speed data. The aspect of mutual comparison of forecasters through correlation was beyond the scope of this study.

This is an important question that should ultimately be addressed. The correlation in forecast error between forecast providers is a measurement of how much independent information is provided by each forecaster. This is related to the potential value of using more than one provider and creating an ensemble composite forecast. If all of the forecast errors are highly correlated then there will be little value in creating an ensemble composite. That is, all of the forecasts essentially contain similar information. The added value of an ensemble composite forecast is relatively high if the skill of the forecasters is similar but the forecast errors are relatively uncorrelated.

13. Section 4.1, item (ii), page 64 states:

Some of the measured-data were modified and/or added after the end of each quarter. It was expected to have the data QA/QC'd and ready to be analyzed before they were posted on GENIVAR's (Phoenix Engineering) wind-server for ORTECH to download, which was not the case.

This fact causes there to be a higher level of uncertainty in the representativeness of the forecast performance results.

14. Section 4.1, item (iv), page 64 states:

ORTECH assumed that none of the forecasters did prescreening to their results. The fact that in some cases forecasters did not provide forecasts up to the 48th time horizon (T=48) puts this assumption into question.

It would be useful to have some statement of the extent of this occurrence. From statements in section 2, it implies that all of the missing forecast data may have been in May 2007 and that month was omitted from the analysis (last paragraph of Section 3.0). Is that the case? If it is, then this is a non-issue for the overall power production forecast evaluation for this project although it may have affected the results for the first quarter.

15. Section 4.3, Page 65 (Freezing the Models) states: Provided one of the questions to be addressed is a comparison between forecasters then it is recommended that after an initial training period the forecast model codes be frozen. If they are not frozen the consultant doing the quantitative analysis can not predict whether the output at the start of the project was the same as the out put generated from the forecast model at the end of the project. Another alternative would be to have each forecaster describe in detail the changes made to the model as the project progressed.

This issue is related to the underlying question of “what constitutes a forecast method?” Is it merely the algorithms that a provider uses to generate the forecast? That is, is the intention to compare the performance of one software package (i.e. a fixed set of algorithms) to another? In a broader sense it might be argued that a forecaster’s method consists not only of methods embodied in the software but also of the humans that are part of the forecaster’s team. For example, what if a provider’s method consisted of a human analyzing all of the forecast data and manually producing values for the forecast parameters? Would it then be required for the same human to provide forecasts for all the hours in the evaluation period?

Thus, the objective of an evaluation process one should be clearly defined: is it a software package or set of algorithms that are to be evaluated apart from the skill of the human members of the forecaster’s team or is it the complete service of the forecast provider which can and often does include the use of human learning and modifications to the forecast production system as the team learns more about the nature of the forecasting problem in a specific region. The specifics of the objective were not clearly defined in this project and as a result the approach toward forecast system modifications during the course of the project varied among forecast providers.

16. Page 5, section (IV):

There was no prescreening of the forecasts at least by Forecaster B. This would not make sense for the long term forecasts. The fact that in some cases the forecast horizon is less than 48 hours is due to missing NWP data. Hence, there were no prescreening criteria to be communicated to ORTECH.

17. Page 5, b) Trial period:

A trial period would have been nice. The first 3 or even 6 months suffered from gaps in the measurement data. Therefore, the results of the first 6 months should be reviewed with caution.

18. Page 5, c) Freezing the models:

There is a considerable difference between "freezing the code" and "stopping the tuning": the forecasting system of Forecaster B was adaptive, hence, it includes new information based on better measurement data and "learns" new relations between weather situations and forecasting errors. This is an inherent feature of the forecasting system. In addition, it

was one of the purposes of the pilot project to learn about the forecasting situation in Alberta and improve the prediction accordingly. So "freezing the system", i.e. forecasting with the initial setup that was implemented in the beginning of the project would not have led to new insights into the "most effective method(s) to forecast wind power in Alberta".

19. Page 26, first paragraph:

"Forecaster B remains relatively flat over all horizons. The difference between Forecaster A and C and Forecaster B is probably due to the forecast methodologies used".

It should be noted that there is a difference for the first half year and the second half year. This conclusion is not right. A different approach was not used for the first half year. For the second half year, Forecaster B has a similar behavior to A and C. See the quarterly results in the appendix.

20. Page 29, 3.1.2 RMSE and MAE Power:

The accuracy of the predictions changed over the duration of the pilot project due to improvements of the data quality and the better tuning of the forecasting models to the situation in Alberta. Please see Appendix E where the results of each quarter are shown separately.

21. Page 36, 3.3.2 seasons of year:

The text on the performance of the forecasters is not well-balanced. It is noted that according to figure 3-9 forecaster B is less accurate in the summer in the SW region than A and C. This is correct (though it is due to a technical error in the aggregation of the forecasts). But in contrast to that it is also true that forecaster B has much smaller errors for EF in the winter and summer compared to A and C.

22. Page 46, 3.8.11 wind speed:

Again the text on the performance of the forecasters is not well-balanced. It is pointed out that forecaster A has the lowest RMSE for $T=1$ and $T=2$ in all regions. At the same time forecaster B has the lowest RMSE for $T = 6, 12, 17, 24, 36, 48$ in all regions except CE.

23. Page 47, 3.8.12 power:

Again the text on the performance of the forecasters is not well-balanced. It is pointed out that forecaster B shows larger errors in the first two horizons. However, forecaster B shows the largest improvement compared to persistence for $T \geq 6$ according to figure 3-15.

24. Page 64, (IV): see comment 1.

25. Page 65, 4.3: see comment 3.