WEPROG

Weather & wind Energy PROGnosis

Aahaven 5 ▪ DK-5631 Ebberup ▪ tel/fax +45 6471 1763/2094 ▪ www.weprog.com

AESO Wind Power Forecasting Pilot Project

Final Project Report

Responsible Officers:
Jess U. Jørgensen
Corinna Möhrlen

May 2008

Executive Summary

The scope of the final report for the AESO wind forecasting pilot project has been to provide detailed answers on a number of questions that have been discussed within the project and that are relevant for the future handling of wind power in Alberta.
The analysis of the results achieved in the pilot project showed us that there is an unusual challenge and opportunity at hand. The opportunity being the wind resource and the challenge being how to sustainably integrate an increasing installed capacity.
In this executive summary, we will briefly discuss our results and findings in the project and provide a recommendation for a future wind power forecasting solution for Alberta. Finally, we will discuss to which wind power penetration level the described solution can scale.

Project Summary

In the AESO wind power forecasting pilot project, WEPROG has been setting up 3 MSEPS 75 member ensemble systems for the Alberta region producing a total of 225 weather forecasts every 6 hours
These systems were dedicated to the project and the configuration was kept unchanged throughout the project. Each of the 225 weather forecasts were converted to wind power with 4 different assumptions. This resulted in a generation of 900 wind power forecasts every 6 hours with a forecast horizon of 3 days ahead for each facility. The first verification of this system was carried out after 3 months in August 2007, and indicated a systematic sensitivity of the forecast error to the spatial model resolution. While this also applied to the prediction horizon 0-4 hour, it was less significant than for the longer horizons (24-72h).
The first event analysis at the 6th of September 2007 confirmed this sensitivity to spatial resolution and opened discussions on the applicability of the approach and the impact of such unpredicted events for the future forecasting system in Alberta.
Another very difficult event in the 1st week of December 2007 showed that all forecasts in the project produced ramping forecasts, where no concurrent ramps were observed. Instead the facility areas remained cold and calm apart from some localised warm short lasting downdrafts. Following this event, WEPROG decided to invest in an increased ensemble size of 600 weather forecasts every 6 hours for the last 3 months of the project in order to find a better configuration for Alberta. An "AESO forecasting test-bed" was

setup and a number of experiments carried out. The results showed that it is possible to improve the forecasts with increased spatial resolution.
In fact, the 75 ensemble members in a 6km resolution experimental setup showed less error growth, lower phase errors, better predictability of errors and a better frequency distribution of the error. Additionally, we achieved a very good understanding of the wind pattern in the area, because the weather maps generated by this 6km ensemble were confirmed by the individual measurements from the wind facilities.
However, the error remains at a level that is significantly higher than in other regions with similar dispersion levels and average wind power generation. And, we do not expect that the 6km ensemble after optimisation is going to change this pattern. It is only the forecast horizon 0-6 hours where we expect long-term improvement, because the algorithms are relatively new and observational and ensemble information has not yet been fully exploited.

How can we be sure that major improvements will be impossible to achieve ?

We have verified 8925 model years in the period May 2007 to April 2008 and 1800 model years for year 2006. Each model configuration has been tested on it's skills at various forecast horizons at each facility. If there would have been some configurations showing significantly better forecasts than others, this would have been found. There are small differences in quality when analysing the facilities, areas, monthly and diurnal cycles. However, the pattern is chaotic and random, indicating only that the spatial resolution helps and that the error is predictable to a large extend by the ensemble spread.

Key findings in the Pilot Project

The analysis of the Alberta test-bed results led to a number of findings:

1. The *initial conditions* are an important and essential factor for the forecast quality in Alberta.

2. The *lateral boundary conditions* need to be provided from a high resolution model ($<= 22km$), because all verification scores of 45-60km resolution simulations indicate high error.

3. The 6km ensemble does not require advanced *statistical training*, as it can produce the wind natively to a much higher accuracy than the lower resolution models. The statistical training can be regarded as an option to use in order to tune on the economic value.

4. Only the 6km ensemble produced the same structures of the wind field as observed: sometimes scattered downdrafts and sometimes a smooth laminar flow gently moving southward and northward over the facility area. The movement is driven by

the Coriolis force due to changes in the wind speed arising from differences in the pressure difference across the mountains.

5. The 6km ensemble was capable of producing long narrow patterns parallel to the mountains of cutoff wind speeds in the south west region. They were nearly always confirmed by measurements when they approached a measurement mast.

6. The large-scale flow frequently amplifies extremely strong gradients along the 50 deg latitude to a level that moderate resolution models cannot simulate correctly

7. All tested error measures improved by using the 6km ensemble, which is impressing considering that higher resolution models respond faster

8. Statistical processing suppresses the model signal in extreme events. The effect is strong in Alberta, because there are considerable phase errors in the training. These errors caused the statistical corrections to dampen the forecast signal and most important ramps

9. Ensembles with large amounts of members give the best statistical measure on the likelihood of sudden generator stops at extreme conditions

10. The ensemble spread is an important measure for the *uncertainty of amplitude and phase of steep ramps*. The correlation between spread and actual error increases with spatial resolution and does not reduce with increased forecast horizon

11. There exists a *seasonal relative performance difference* between the ensemble members mainly due to the snow cover

12. The *difficult forecasting days* consist of triple series of ramps each associated with one low pressure system with a narrow cold front. The most difficult events ramp up and down withing a two hour interval

13. An "extreme ramp analysis" indicates that approximately 90% of the extreme ramp events were predicted by some ensemble members with the correct amplitude, but 20% of these had phase errors between 1-3 hours

14. Extreme static stability has caused systematic forecast errors with nearly all model formulations because of incorrect vertical mixing

The Challenges

Using a single forecast with a normalised mean absolute error of 15% for the bid on the day-ahead market creates an expectation. All market participants know that this forecast is error contaminated and will look for opportunities to exploit the error. Why is this ?

Energy prices suffer from volatility, if there is either urgency, risks, monopolies, variable supply and demand or other factors such as congestion on the grid. We have reason to believe that the wind power forecasting errors will contribute significantly to volatility. Unless determined steps are taken to "hide" forecast errors from the market, the volatility will increase with increasing installed capacity of wind power, even if all new capacity is installed optimal for minimum forecast error.

The intermittent generator is the weakest party during volatility and the reason why wind power has been dependent on production incentives until now. The key challenge however is to facilitate wind power integration, so that wind power can operate on market terms and independent of production incentives. The next section outlines how this milestone can be met.

The recommended Forecasting Solution for Alberta

The required components in the recommended forecasting solution can be outlined as follows:

1. All wind power capacity is included in a "Wind Generation Pool" (WGP) for maximum inertia from spatial smoothing. The WGP is balance responsible for wind power

2. A reliable measurement network is established for enhanced 0-4 hour forecasting with independent met and power measurements

3. A daily, weekly or monthly market for reserve (R_{BPP}) of wind power is established, in which "Balance Providing Parties" (BPP) are chosen via tenders.

4. A weather forecast component is established with capabilities of producing a large number of different high resolution NWP forecasts as a representative ensemble to estimate the forecast uncertainty.

5. A wind power module is established computing wind power (P) and ramp rate ($RR = \frac{dP}{dt}$) from the ensemble of weather forecasts, which will result in an optimal forecast (P_{opt}), a lower limit forecast (P_{min}), an upper limit forecast (P_{max}) and maximum ramp-rate (RR_{max}) forecasts, where the limits are approximated with quantiles for the chosen ensemble size (see the red and green line on Figure 1.

The first aspect to consider is how the WGP should use the above components in the day-ahead market. The WGP will base the bid rather on the extreme forecast than the most likely (P_{opt}), i.e. P_{max} should be chosen by default. There may be times were the BPP is limited and can not provide the required reserve R_{BPP}, which is ($P_{max} - P_{min}$). The WGP can then only bid in with $P_{min} + R_{BPP}$. It is also necessary for the WGP to verify that the BPP can ramp with at least RR_{max}. A failure of this test will require

the WGP to limit the bid to match the available ramping capabilities. By using a large ensemble size we ensure that P_{max} is a smooth function. The WGP can however smooth further by bidding in with unused reserve capacity to maximise the daily base generation of WGP and the BPP.
The red line (P_{max}) on Figure 1 shows the default bid, if the ramp rate is moderate and the BPP has sufficient balancing capacity.

The purpose of the short-term forecast (0-4 hours) P_{opt} is to schedule the reserve of the BPP according to yesterday's bid of the WGP. For this, the BPP will receive new forecasts of the expected deficit between the wind power generation and yesterday's bid every 15 minutes. The forecast update must be frequent, because the error is significant already on the 2 hour horizon, as can be seen on Figure 2.

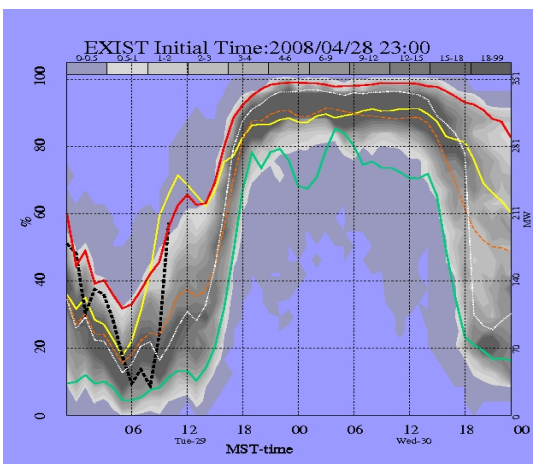

Figure 1: Example of a wind power probability forecast computed from 10^3 members. The upper 80% quantile is shown as the solid red line, the yellow line as the so-called best forecast and the green line as lower 20% quantile. The black dotted line is the measurements.

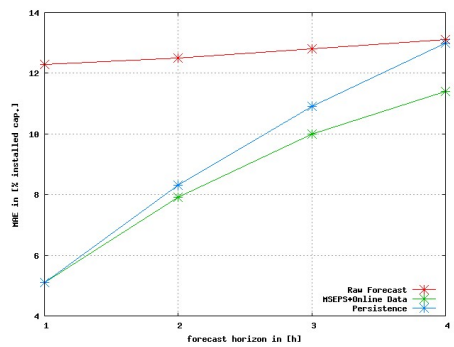

Figure 2: Comparison of MAE error for the 1-4h horizon at existing facilities using *WEPROG's ensemble short-term algorithm* for 05/2007-04/2008. The red line is the forecast computed on the basis of historic measurements and weather forecasts, the green line combines the weather forecasts with online measurements, the blue line is the persistence forecast.

The above bidding scheme minimises the impact of the intermittent nature of wind power for all market participants. The BPP(s) will absorb all variability and the market will experience only moderate ramps from the WGP and no severe imbalances. The pooling concept for wind power will therefore suppress price volatility, because the BPP is contracted in due time and open competition. The capacity a BPP may provide to the WGP is estimated to approximately 50% utilisation averaged over time.
The WGP also has to pay for the flexibility of the BPP. However, this price will always be lower than the market price for primary reserve, because it is a dynamic reserve bought in

advance that enables the WGP to deliver the maximum possible power production from wind. This balancing power is required much more regular than primary reserve. In other words, the BPP may get a lower average price per produced MWh, but the BPP will also have more production hours per time unit. As long as the price paid by the WGP is above the achievable market price following the merit-order (especially in windy periods), this will be an attractive alternative for the BPP.

We believe the WGP concept can scale to a wind power penetration level of up to 50% without compromising grid security and economic feasibility. The economic feasibility of wind power is expected to increase relative to fossil fuel driven generation, because of the increasing demand world wide. We estimate 50% to be an upper penetration limit, because Alberta's grid is not strongly interconnected and the wind power generation is correlated. This implies many hours, where wind will have to cover nearly all demand in Alberta. A 50% penetration level also implies a large over-capacity of generation in periods with a lot of wind. Although this will make participation in the WGP attractive for the scheduled units, the over-capacity will lead to an increased average market price per energy unit. This increase will however be significantly lower than in other areas with lower wind resources. Thus, the electricity price can be expected to still be competitive in a global perspective.

From the above description it is obvious that the forecasting solution has to be centralized and for security reasons most likely located in the AESO. A decentral solution, where all wind facilities forecast individually will neither secure safe operation of the grid nor the economic feasibility.

Conclusion and Outlook

The main challenges in the Wind Power Forecasting Pilot Project was found to be of meteorological nature. The lack of weather forecast accuracy turned out to be an obstacle to the quality of wind power forecasts. The wind power computation relies fully on the accuracy of the weather forecast. Thus, before tuning on the wind power computation step, the quality of the weather forecast should approach a level to which further improvements are marginal.
WEPROG had no prior experience with the specific conditions in Alberta and a significant amount of resources were therefore used to identify the important parameters for high quality weather forecasts. What works very well operationally in Europe and Australia was not applicable in Alberta. Therefore, some basic research had to be conducted to understand the driving forces of the local wind conditions. This was a time consuming process. The results delivered to the project did not benefit from this work, because a fixed setup was used. However, a graphical user interface had been setup to enable the AESO to follow the various test setups. One of them was a 75 member ensemble running in 6km resolution.

We consider the lessons learned from the 6km ensemble as a very important breakthrough in the understanding of the variable nature of the forecast error. In meteorology it is a common understanding that the initial conditions are the main cause of forecast errors. Our studies indicate that there are at least two additional major aspects causing significant forecast errors in Alberta. These are:

- Lower resolution models develop stronger gradients than they can resolve correctly. Consequently their adiabatic equations give wrong tendencies and the weather forecast quality degrades
- Extreme static stability has caused systematic forecast errors with nearly all model formulations. An example is the difficulty in simulating the breakdown of the stable planetary boundary layer

The meteorological forecast error at the end of the project is still high. We got an understanding and an improvement, but there is a serious error to deal with.
Nevertheless the wind power resource is huge in Alberta and the forecasting quality of the day-ahead forecast should not be the limiting factor for the development of wind power. Therefore, we developed a pooling concept, which is taking advantage of probabilistic information from the ensemble forecasting methodology that was used in the project.
With this concept we expect wind power in Alberta to scale to ca. 50% penetration level without loosing grid security nor economic feasibility compared to other countries.
The "Wind Generation Pool Concept" almost prevents balance costs and fits into the current market structures. Wind power in a pool can act in competition with scheduled units. It should also not be regarded as a dramatic difference to wind power's present status, but rather as a hybrid between AESO's current handling of wind power and a traditional market implementation based on a single forecast. In fact, the main difference lies in the management and utilisation of the required balancing power.
A traditional market implementation of forecasting in Alberta is not going to bring economic efficiency, because the error will not be reduced significantly with time.
The reason why the pool concept can work is that the day-ahead market does not reflect marginal costs, but the level of competition in each of the next day settlement intervals. Within a pool there is possibility to adjust the generation not only once per day, but every few minutes at nearly marginal costs. This process increases the effective market value of wind power.
To conclude, the choice of forecasting and market implementation strategy will have significant impact on how wind power will develop in Alberta. While a traditional forecast solution will likely require production incentives to make investments in wind power feasible once a certain penetration level has been reached, we are convinced that the pooling concept has the potential to scale to a penetration level of up to 50% from wind on market terms.

Acknowledgements

First of all, we wish to acknowledge the AESO, the industry group and CANWEA for taking the initiative for the wind forecasting pilot project. This project was not only an interesting project, but also posted an enormous challenge to all participants, the forecasters, GENIVAR- Phoenix Engineering, ORTECH and the project coordinator. However, we also think it was a project where all participants learned a lot.
Special thanks to Darren McCrank for keeping this project on track, making sure that deadlines are met and finding compromises with all parties involved. This was certainly not an easy task at times.
We wish to acknowledge the team in GENIVAR-Phoenix Eng. for the data handling and update of availabilities. Bringing order into the data, masts, breaking connections and missing data streams, was certainly a challenge of the greater order.
We also wish to acknowledge the analysis team in ORTECH, that has probably been writing more report pages about time-series analysis with high variability than a university department would in 3 years. It has not been an easy task to make sense out of the enormous amounts of data, considering the level of insight that was provided in the methodologies and the expectations to the quantitative analysis. Every Met-service around the world is challenged by showing forecast skills in a useful way to justify the progress of their modellers. Therefore, we, as forecasters know the amount of effort it takes to take advice from so many parties and to get a meaningful result out of the available data.
Last, but not least, thanks to all those that have helped making this project a success behind the scenes.

Contents

General Project Overview

The structure of the Rocky Mountains between British Columbia and Alberta in conjunction with the distribution of wind power in Alberta at first glance gives the impression of moderately complex conditions. Wind power positioned closest to the mountains would be expected to be considerable more difficult to predict than capacity further away from the mountains.

These were the initial expectations when designing a suite of forecasts to be used during the pilot project for wind power forecasting in Alberta.

Historical data from year 2006 was made available for training at the outset of the project. The following training phase showed that the expected error pattern was misjudged. A multi-scheme ensemble prediction system (MSEPS) in 60km horizontal model resolution was first used to evaluate the data. Here, it was noted that the error was considerably higher in winter than summer relative to the installed capacity. There were some periods with a very large error, while the overall picture did not indicate severe problems except in a handful of events.

It was interesting to observe that March and April in 2007 were already significantly more difficult than what was observed in 2006. Thus, the weather pattern had become more variable already in the beginning of 2007.

Year 2006 data were evaluated on the basis of the oldest facilities. However, some of the newer facilities moved the centre of facilities, which actually had some impact on the forecasting results, because this added variability in the generation output in that region. For confidentiality reasons it is however not possible to go further into detail.

Some extra experiments were carried out for the winter 2006/2007. They showed sensitivity to the representation of the mountains in the model world. After more testing it was decided to run year 2006 with a 22km ensemble with two different orography formulations. The result pointed to the ensemble with the steepest orography and its configuration was then chosen for the project as a nested ensemble in an outer 45km ensemble providing the lateral boundary data.

As the forecasting went on, it became clear that 2006 was not a good reference year. There were many surprises to be encountered in 2007. In the autumn, it became visible that the 22km ensemble was repeatedly very different from the 45km ensemble. Further sensitivity experiments showed also sensitivity to even higher resolution. The same equations applied in different resolutions even with interpolated orography and roughness gave fundamentally different results. This was certainly a worry, because that meant that the setup chosen for the project suffered from numerical truncation errors.

After sufficiently many failed forecasts, it was decided to setup a 6km ensemble with a

considerable higher computational cost. In fact, 4 additional 75 member ensembles were tested over a period of 3 months in 22km resolution in order to explore the effect of a some changes in a so-called "test-bed", which will be described in more detail in Section 2. In this test-bed, 600 weather forecasts were generated every 6 hours and more than 2/3 of the CPU time went to the high resolution 6km ensemble.

Fortunately, the difficult conditions continued over the entire 3-months of the intensive test. It is therefore worth noting that year 2007 and the beginning of 2008 were both more difficult forecasting conditions than the training year 2006.

After the pilot project it appears that more radical solutions should have been applied earlier in the project, not only because the error at the end of the project is still high, but because the flow across the mountains does not have the smooth laminar character as a moderate resolution of 20km expects as input and delivers as output.

Using a 20km model formulation to simulate a chaotic flow on a smaller scale does imply the risk of a failure in the model formulation, because there is a special parameterisation of the small-scale motion required. Traditionally, a gravity wave drag parameterisation and the vertical diffusion process should handle this. The project however demonstrated that neither of the tested formulations was capable of modelling the flow correct and in a satisfactory way.

With increased spatial resolution we tried to simulate the chaotic flow explicitly. It was neither the expectation nor the idea that this would give a sustained correct time-series of data, because we do not know the true initial conditions. What we can however achieve is a flow with a correct structure. The different ensemble members can be regarded as perturbations around the true flow, whereas it cannot be expected that they are exact when the flow is most scattered.

As year 2007 and 2008 developed, it can be postulated that forecasts with a sustained high accuracy will not be achievable in Alberta except on the 0-4 hour forecasting horizon. What we have learned in brief is:

- Out of 600 forecast, 375 independent forecasts had errors of similar magnitude and 225 were not good enough to add value to the ensemble
- The forecast error is large for most horizons and in all areas
- The wind flow is occasionally scattered and not predictable

These results suggest that it is crucial to find a forecast methodology in which the accuracy of a single day-ahead forecast is of minor importance for the future wind power integration in Alberta.

The AESO previously proposed a power management scheme along with the requirement of "reasonable good forecasts". Knowing that the forecast error is high and that the result will most likely be increased volatility on the market pricing with growing amounts of wind power, it will be more important to implement methods to handle the forecasting error than to reduce it i small steps.

Having carried out hourly updated 48 hour forecasts over a year with 225 independent weather forecasts, 600 weather forecasts over a period of 3 months and having studied the wind power generation for more than two years, the following analyses and recommendations for the future handling of wind power in Alberta can be considered well founded.

Lessons learned from Forecasting in Alberta and Findings

The scope of the Pilot Project was to test different methodologies over one year in order to explore what kind of forecasting methodology should be used in a future operational environment in Alberta.
For this reason different methodologies were selected in the project. To have consistency in the final results this also meant that there should only be modest changes to the setup. The forecasters had one year of historic forecasts available to train and adopt their model systems.
WEPROG followed this principle with one exception. Following the first special event analysis of the 6th of September 2007 case, two updates were made to the original system. On enforced the 1-4 hour forecasting system follow the measurements stronger. The second included new power curve based on the summer 2007 data. Both were however reverted to nearly the original formulation in December 2007 as they had not been very successful. The two changes were modest and were tuned on the period May to August 2007, but they didn't take account for the more variable and less predictable conditions that characterised October 2007 to January 2008.
Apart from these two changes listed above there was a period of 4 weeks starting on the 19th of June 2007, where there was an error in the handling of the lateral boundary conditions on the large scale ensemble, which reduced the quality of the day-ahead forecasts in that period.
Some technical changes were also required on the handling of measurements throughout the project. Otherwise the forecasts were carried out according to the original schedule over the course of the year with fixed setup.
At the outset of the project WEPROG was running the following systems:

- System A - a 45km resolution 75 member setup with envelope orography
- System B - a 45km resolution 75 member setup with mean orography
- System D - a 22km resolution 75 member setup with envelope orography

The envelope orography has effective filled all valleys on model sub-grid scale with some solid material. In the mean orography we simply smooth out the mountains until the valleys are filled up. The mean orography need a higher roughness field to block the flow, while the envelope orography is taller and block the same with less roughness. System

A, B and D can be regarded as different parameterisations of the mountains. There is at certain places up 800m altitude difference between A and B.

System D was the primary system to deliver forecasts and the ensemble mean from this system was used to deliver the weather weather variables. Note, that a more advanced statistical processing around the ensemble mean was used to predict their power, but the ensemble mean itself has a strong influence on this result as well.
The 3 systems delivered 225 weather forecasts every 6 hours, where only one forecast of 48h was delivered to the project evaluation. Therefore, WEPROG additionally set up a graphical user-interface *PLATON* for AESO to be able to follow the trends and additional information from ensemble forecasts in real-time. Appendix C provides some screen-shots and further explanations to PLATON.
Note that the discussions and conclusions presented in the following are drawn from evaluation of all of these 225 forecasts. System D can be regarded as the delivered data, but we have found out that around every 12th forecast were delivered from system B. The reason is that system B is done before D and if system D would be just a few minutes delayed at 5, 11, 17 and 23 hours in UTC time, then the delivery process picks up system B instead of D, because B was in that case the newest.

2.1 The adjustment and training of the System

The target for the power prediction was to provide reasonable probabilistic forecast of primary power, reserve and ramping. This was done by tuning on the frequency distribution of the forecasted power generation. That means the forecast should produce the correct amount of hours with no generation and full generation. The forecasts were calibrated to produce the correct amount of hours with a given level of power generation. The technique allows to pair any forecast horizon with measurements without any degradation and in that way take account for bias. This should have been done to take account for the drifting bias.
The assumption was then that if the forecasts contain every significant ramps then $\delta P/\delta v$ should be reasonable accurate. However, we know that the forecasts are not perfect and there are ramps that are predicted, but on the other hand also forecasted ramps that do not exist. The assumption is that the cancel each other over long time.
The algorithm can be tuned by running a low pass filter on the measurements, which will reduce the variance to the same as the forecast. That will however reduce some of the extreme ramps with little amplitude and short wave length. These would in real life be balanced by primary or secondary reserve and are therefore not so important from a grid security and market perspective.
The forecast horizon used in the training included the horizon 6-18 hours in meteorological context, which would be 1 to 13 hours in end-user context. Thus the forecast should on average be best around the 6-hour horizon.
The applied calibration technique has the following capabilities:

- The power forecasts act responsive to any change in the wind speed
- Phase errors in the training have no impact on the estimation accuracy
- Every ensemble member's power curve looks similar to the ideal power curve, but is shifted left or right depending on the models wind speed bias
- The probabilistic power forecast is consistent with the ensemble measured as wind speed probabilistic forecast
- Derived products such as reserve prediction can be composed directly from the minimum and maximum values of the ensemble of power forecasts
- Ramp rate forecasts have nearly the same quality at any forecast horizon. The steepness of the ramp does not changes, but the time window at which the ramp may take place increases over the forecast horizon.
- The algorithm is not optimising the individual forecasts for minimum RMS error. If required, it is however relatively easy to compute an optimal smooth forecast from the ensemble, utilising the members that have the best long term statistics.

From the above, it can be concluded that the power forecast methodology is most optimised for responsive ramping although the method does not guarantee that the ramps are of the correct magnitude or are in correct phase.

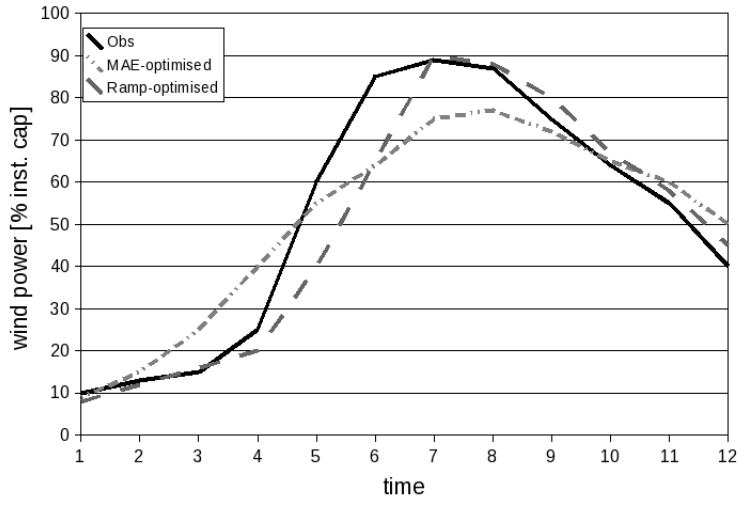

Figure 2.1: Schematic of the difference of optimising with MAE/RMSE and of optimising on ramps.

The ramp rate has impact on the actual costs and the approach give the possibility to estimate the ramp rate within a certain interval. It was found in recent studies that the economic optimisation in a future system should rather focus on a cost function than on optimising the mean absolute error (MAE) or root mean square error (RMSE) ([15]). We expect that the way to implement this idea in real life would be to compute the cost of each forecast and choose the cheapest. Here the cost is defined as the average balance cost found by assuming all forecasts are equal likely representations of the truth.
This means that the forecast we will choose in the end by this algorithm is the most defensive in cost space and not in MAE or RMS. This will cause that we choose a slightly steeper ramp than the average, because it is cheaper to slow down a ramp than accelerate a ramp. A scheduled generator with wind in the pool can benefit from this computation, because the marginal costs of ramping is known more accurate than for parties who rely on the market.
Figure 2.1 shows an schematic example of a steep ramp, where one forecast is optimised on MAE/RMSE and the other forecast is optimised on ramps, or simply on predicting the weather as is with a purely physical methodology,

2.2 Forecast Performance in Alberta and Comparison to other Counrties

The forecast performance in the short-term forecasting horizons up to 4 hours have been satisfactory. In fact there has been made good progress in applying our ensemble approach together with online measurements to make use of the ensemble spread to extrapolate observations into the future.

Figure 2.2 shows the verification of mean absolute error (MAE) of the delivered short-term wind power forecasts on existing facilities (green line) together with a persistence forecasts (blue line) and the raw weather forecasts from the 22.5km D model setup. It can be seen that already after one hour, the short-term algorithm outperforms persistence, which is an encouraging result in respect to the intra-day market in Alberta, which is working with a 2-hour time frame. Note, that more detailed statistics of the ultra short-term forecasting verification can be found in Appendix A.9.

The results of the day-ahead forecasts based on the the 225 weather forecasts from system A, B and D did however not meet the performance level that is known from other places with similar wind resources and the same or similar model configurations. System B is similar to what WEPROG uses in Europe and Australia, where the error level is approximately half of the error level in Alberta and at some places even less.

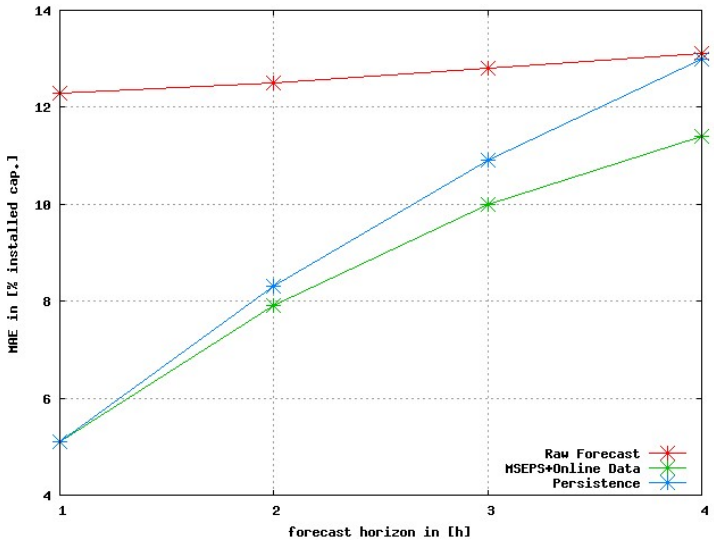

Figure 2.2: Comparison of MAE error for the ultra short-term forecasting at existing facilities using the *WEPROG ensemble algorithm for short-term predictions* for the forecasting horizon 1-4 hours ahead delivered to the AESO pilot project in the period May 2007 - April 2008. The red line denotes the raw forecast, the green line denotes the MSEPS algorithm making use of online observations and the blue line denote persistence.

Figure 2.3 shows a comparison of forecasts in the pilot project period (May 2007 - April 2008) with forecasts delivered to Australia and Ireland. Further statistics can be found in Appendix A.1.
These two countries were chosen, because of a similar wind resource and integration structure of wind power as Alberta.
This lack of quality of the forecasts was identified already in the preparation of the project, but became more and more evident in the course of the pilot project. Especially, the first special event analysis in September prompted to the need to investigate other solutions for the weather forecast ensemble setup than the selected. An analysis of the error pattern was conducted for August and September 2007, where we carried out approximately 200 sensitivity experiments with different horizontal resolutions from 3km up to 60km, different vertical resolution and a number of different model formulations. These experiments indicated that a number of improvements could be achieved by using a different model configuration setup.

The conclusion of these experiments carried out in September 2007 was that the chosen model resolution of 22.5km was insufficient to solve the Alberta day-ahead forecasting problem satisfactory. However, the project was almost half way through and it was close to the change into a new season, where the weather had already started to become more variable and changeable. To make significant model changes would have made it difficult to trust in the final statistics and caused also a lot of technical changes to a running system, which had proven stable and reliable.

The decision was therefore to conduct sensitivity experiments in order to investigate the most suitable system configuration for Alberta and to find out how much error reduction can be gained by better weather forecasts rather than making changes to the real time system.

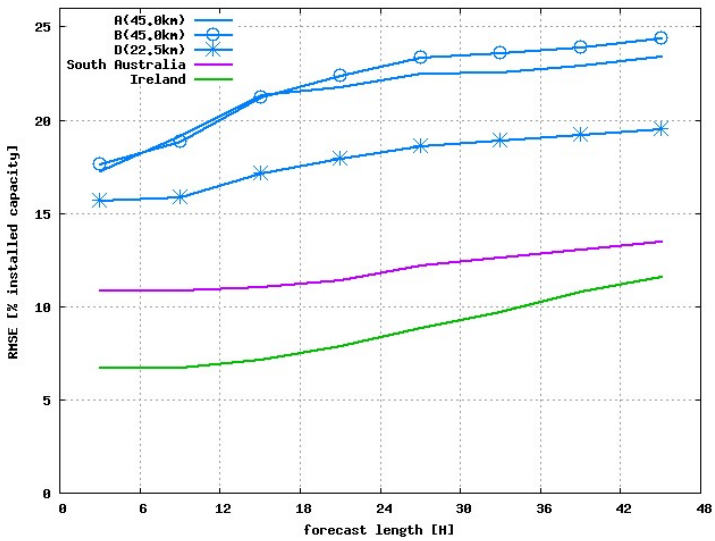

Figure 2.3: Comparison of RMSE error over the forecast length 0-48h of the forecasts delivered to the AESO pilot project and forecast delivered to Australia and Ireland in the period May 2007 - April 2008.

A note on comparisons of forecasts and penetration to other countries

When comparing the forecasting error of Alberta with other countries, there are a few important aspects to be noted:

- The higher the dispersion level and/or the more uncorrelated the wind power, the lower the forecast error
- The higher the wind resource the more competitive is wind power
- The more condensed the wind power is installed, the less the grid integration costs

When evaluating these aspects, it becomes clear that there is no perfect location, where all positive aspects are met. Figure 2.4 shows a frequency distribution of a number of areas with similar or significantly more installed capacity than in Alberta.
In Australia we observe an almost ideal near-Gaussian distribution of the wind power generation. This is ideal, because it implies relatively small spot market price variations due to the varying wind power generation. The dispersion of wind farms is also high and hence the power generation mostly uncorrelated. However, the wind facilities are often

far away from where the demand is large enough to absorb the wind power. This may introduce transmission constraints.

Ireland, like Australia has a much more uniform generation pattern. While the wind power is also distributed but not really dispersed, the generation is not as uncorrelated as in Australia. The high resource also makes wind power economically more feasible as in the Danish and German case. Ireland also have some grid constraints and low interconnectivity.

This is opposite to the western part of Denmark, where wind power is dispersed, but also condensed on a small area. Here, the wind power is rather correlated. The forecast error is relative small, which should mean high efficiency. The correlated generation however creates strong price volatility. The west-danish energy market is getting saturated with energy during windy periods due to the large installed capacity, which is higher than the minimum demand. Consequently, the energy price is very low or even zero in windy periods and high in other periods. Such conditions makes it difficult to reach the marginal cost level of wind energy on market terms without production incentives.

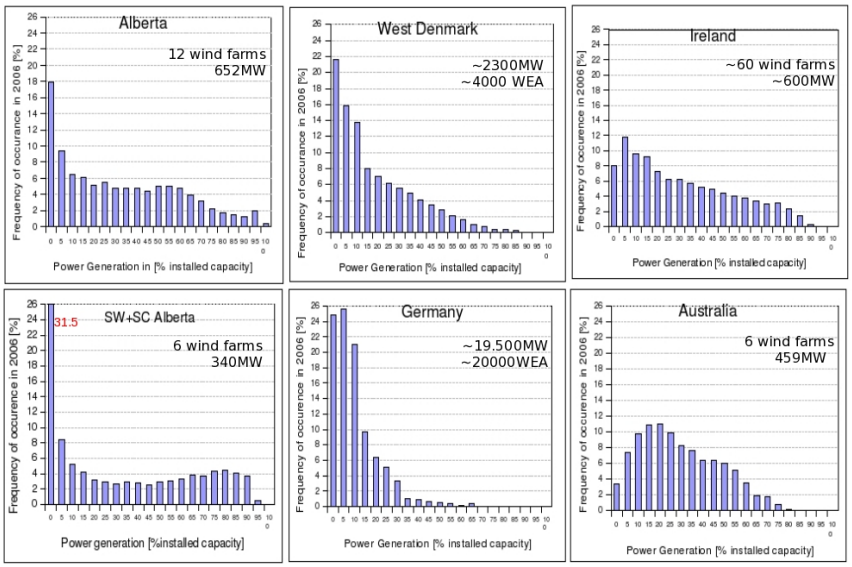

Figure 2.4: Frequency Distribution of wind power production in different countries. The frequency distribution of the actual generation is computed by counting the number of hours with generation within a certain interval. The possible output values lie between 0 and 100% of the full capacity. This interval is divided into 20 equally sized sub intervals/bins.

This is the same for the German wind power, where the distribution is relatively similar in the lower bins, but differ in the higher bins, indicating the higher load factors achieved in Denmark. Nevertheless, in both countries a high penetration level would imply a large over-capacities and consequently many hours of low prices. In Denmark, there is already more wind capacity installed than average demand to reach a 20% supply level from wind. To reach the currently set goal of 50% of average consumption from wind (ca. 1800MW), 250% wind power capacity are required compared to the average demand. This will most likely result in many hours with energy export at very low or zero prices, while there is still need for a large amount of scheduled capacity.

Taking all aspects into consideration, it has to be concluded that Alberta's energy system will be able to function very efficient at high penetration levels, because of the higher capacity factor and relative ease of increasing the transmission capacity. Thus, the challenge will rather be to find a forecasting strategy that is able to either improve significantly or alternatively deal with the high day-ahead forecasting error without compromising grid security or economic feasibility.

2.3 The AESO Forecasting Test-bed

The end of November and beginning of December 2007 were weeks with extremely variable weather conditions that made forecasting for Alberta extremely difficult. The sensitivity experiments together with these weather conditions prompted for a more structured test environment, because they didn't give a consistent picture of the forecasting issues at hand.
WEPROG therefore decided to establish an "AESO forecasting Test-bed". The test-bed was made ready for operation by the 1st of February.
The process of establishing the AESO Test-bed included the following phases:

- Acquisition of new hardware, suitable for the task: a HP blade system
- Installation of operating system and cluster maintenance software
- Installation of the MSEPS software and first test runs
- Migration into a hosting centre
- Final setup and testing of the "AESO forecasting Test-bed"
- Setting the new system into real time operation

The system carried out forecasts for the period 1st of February to 1st of May 2008 and the verification included 2nd of February to 30th of April.

2.3.1 The MSEPS Test-System Description

Table 2.1 summarises the configurations of the different delivering setups and experimental setups. The systems A, B and D were the delivering systems to the project with D as the primary. Systems E - J were test systems and K is a large-scale North-America setup in 65km. System J refers to the "super ensemble"formed by aggregating all system's ensemble members except K's, which forecast quality in 65km resolution is too low to add value to the "super ensemble".

ID	hor. model resolution	# of members	resolution lat. bnd	of lat. Bnd	Analysis	Resolution of Analysis	Orography
A	45km	75	1deg[1]	1	NCEP	1deg	envelope
B	45km	75	1deg	1	NCEP	1deg	mean
C	45km	75	1deg	1	NCEP	1deg	envelope
D	22km	75	45km	75	NCEP	1deg	envelope
E	22km	75	45km	75	NCEP	1deg	envelope
F	22km	75	45km	8	NCEP	Variable	envelope
G	22km	75	22km	8	NCEP	1deg	envelope
H	6km	75	22km	8	NCEP	1deg	envelope
I	22km	75	22km	8	CMC	60km	envelope
J	-	675	-	-	-	-	-
K	60km	75	1deg	1	NCEP	1deg	mean

Table 2.1: Summary of the various ensemble systems in the AESO test-bed for the last quarter of the project.

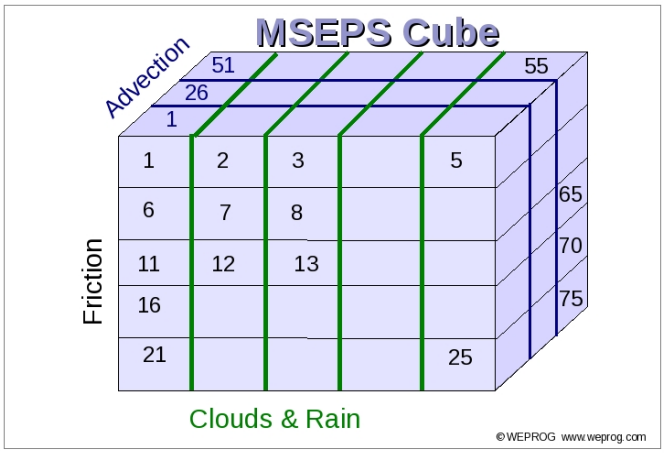

Figure 2.5: "MSEPS cube": illustration of the 5x5x3 matrix of the MSEPS system.

System C uses the same ensemble members as B, but a newer version of the statistical processing. The same relationship applies to D and E. For this reason, only A, B, D, F, G, H, I and K can be regarded as independent 75 member ensembles and are part of the following verification.

Each of the 75 member ensembles can be understood as a 3D matrix of dimensions 5x5x3 where each matrix element should be interpreted as one NWP model formulation. Figure 2.5 gives an illustrative overview of the MSEPS matrix. Each cell updates the previous state of itself with information from the global state estimate every 6 hours.

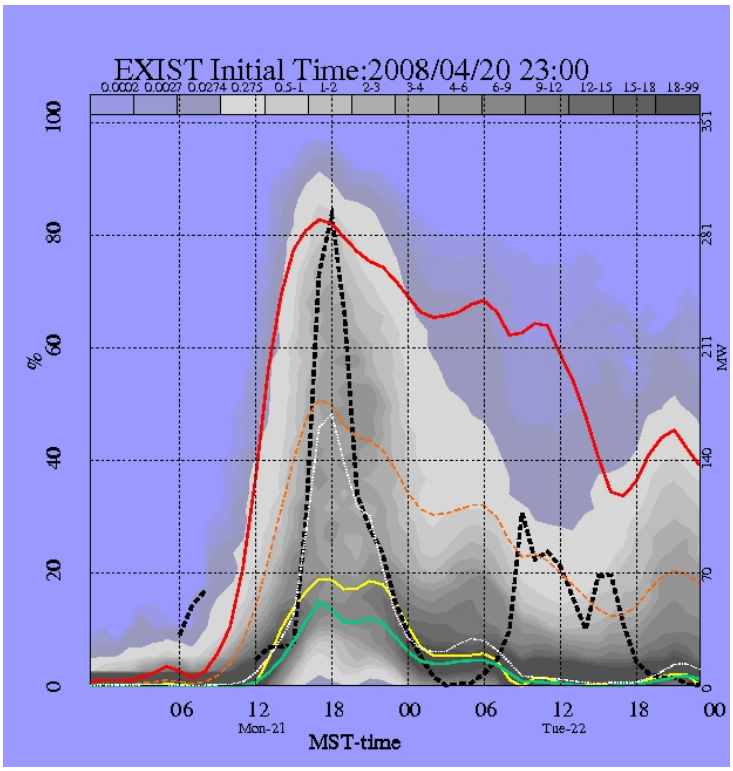

Figure 2.6: Example of a probability distribution of wind power from the "Super Ensemble" here based on 525 weather forecasts, but 2700 power forecasts. The black dashed line are the measurements, the red line denotes the maximum, the green line the minimum, the yellow line a statistically optimal forecast, the red dashed line the mean, the white dashed line is a best guess following the highest probability. The gray shading denotes the probability density, where the bluish outer colours contain the lowest and highest 5 ‰ of the 675 forecasts.

The 75 ensemble members in setup F have some additional differences. These members should rather be seen as a 6 dimensional sparse matrix, where the average distance between cells is maximised. The aim with setup F was to get a better understanding of which parameters had most impact in Alberta.

The 8 members providing lateral boundary conditions were not included in the wind power evaluation. The focus with these members were to simulate the large-scale weather pattern over an area of 4000x6000 km.
System I, which is based on CMC global analysis was approximated by the 6 hour forecast from the previous run at 06 and 18 UTC, because these state estimates are currently not available. The error level has a tendency to start higher, but there is no visible effect of this after about 18 hours.
Setup J, the "Super Ensemble", which is based on 525 weather forecasts gives a very smooth forecast. The PMT filter nevertheless produces consider table ramps at certain times. By applying quantiles for the lower and upper 5 $^o/_{oo}$ of the ensemble members, the information is filtered better and hence more useful. Figure 2.6 shows an example of a probability figure of a ramp event, where the EPS mean and pmt-filter generated best guess are too soft because of the coarser resolution systems that drag the average down. Although the likelihood did not seem to be too high for the ramp, the "Super Ensemble" provided an informative overview for the possible ramp at 18 hours MST.

2.4 Summary of the Findings

The analysis of the Alberta test-bed results led to a number of interesting findings. The detailed results are given in Appendix B

1. The *initial conditions* are an important and essential factor for the forecast quality in Alberta.
   The location of a low pressure system or a high pressure system is difficult to determine in a mountainous region. Additionally, there is often imbalance, which tends to deform the lows or the highs. The first 12-15 forward integration hours have hardly any error increase on system F, G, H and I, which shows that it is no longer the forecast step that is the major problem. The uncertainty of the initial conditions seems to dominate.

2. The *lateral boundary conditions* need to be provided from a high resolution model ($<= 22km$), because system G, H and I are better than D and F.

3. The 45km ensemble is unsuitable for Alberta, as it cannot resolve the mountains and the details in the facility regions sufficient well and hence the weather on the lee side of the mountains. The result is consistent for the entire year and for Q4 alone. However the steeper orography in A does help on the score on all parameters and the error growth of A has been relatively low in Q4.

4. The *local details* within the facility region are important. The ensemble mean usually gives the best position of the extreme gradients and the structure of the horizontal gradients. This can be concluded from system H, because this resolves 4-5 times sharper gradients with the potential of causing more error, but this is compensated by more correct model dynamics.

5. The 6km ensemble does not require much *statistical training*, as it can produce the wind natively to a much higher accuracy than the lower resolution models. The statistical training can be regarded as an option to use in order to increase the economic value.

6. To achieve more realistic and reliable forecasts the target is that the high resolution 6km ensemble produces equally scattered winds and equally smooth winds as observed. The ensemble's variability must reflect the variability of the measurements. This is achieved if the flow structure is on the same scale in the model and in the atmosphere. More precisely the resolution must be sufficient to resolve the downdrafts that are observed. It is considered as good progress that this target can be met without increasing any common error measure, which would be the expected result after introducing a more responsive high resolution model.

7. The ensemble mean is a better forecast than the individual members in very certain conditions. Even during such times the location of extreme values will natively be split within a region of at least 4 different grid points. The average of the members will give a better average position within the 4 points. Thus, ensemble forecasts can actually add resolution to the resolution of the individual members.

8. The ensemble spread is an important measure for the *uncertainty of amplitude and phase of steep ramps*. The correlation between spread and actual error increases with spatial resolution and does not reduce with increased forecast horizon.

9. There exists a *seasonal performance difference* of ensemble members. The ensemble members that are best in winter are not the same as those that are best in summer. Different numerical formulations are superior to others at different times. Simple non-local formulations are better in the winter, while this is the opposite in the summer.

10. The *most difficult forecasting days* consist of triple series of down-ramps and up-ramps. The first down-ramp creates a low pressure system right over the facility region due to potential vorticity conservation in the flow over the mountains. The second up-ramp is when cold fronts arrive from north towards the facilities. The third down-ramp is generated when the cold front finally moves south-eastward.

11. Very *cold conditions* in the winter can result in extreme static stability in which nearly all model configurations have a tendency to over-predict. System H was found to give the best results in these events.

12. *Northerly flow* results in strong cold fronts with strong ramps in the SE and SC region. Very high spatial resolution is required to produce a correct ramp rate in these situations. It is expected that the ramp rate can reach 300-600MW in a 5 minute interval in maybe 1-3 events per year in the future

13. Several hours of northerly or northwesterly flow leads to an unstable atmosphere with significant *potential vorticity*. This implies a risk for more frequent short lasting ramps at a number of facilities, which can be derived from this type of large scale weather pattern.

14. The *risk for large ramps is correlated to the strength of the jet stream*, which is weakest in the very late summer and early autumn and again in spring time. The flow over the mountains is not maintained for very long in these periods. As the flow is blocked, new low pressure systems are born more frequently with the risk of large ramps. Due to the taller mountains further north, the blocking starts there, whereas warmer air continues to slip through at the facility regions. Cold air from North-East is sucked down on the northwest side of the low and develops to a sharper and sharper front that can interact suddenly with a low pressure system now starting to move eastward. Many of the low pressure systems that are born in the facility region may reach as far as Russia.

15. The "extreme ramp analysis" that was carried out indicates that approximately 90% of the extreme ramp events with phase errors of less than 4 hours seem to be predictable by the ensemble. It was also found that if non of the ensemble members caught the ramp within 3-4 ours, the event was not predictable at all. This was confirmed by e.g. the probability of detection, which did not increase further.

16. The *friction process* is sensitive to a correct temperature profile in the planetary boundary layer within the mountains. The depth of the boundary layer and the roughness determine how much the mountains block the flow. Detailed modelling of the vertical circulation is hence required the net blocking right.

17. *Statistical processing suppress* the signal in extreme events. The effect is strong in Alberta, because there is considerable phase error. Cutoff wind speeds, extreme temperatures and ice-up occur so rare that it is better to describe the events physically in a nearly unbiased ensemble and use the probabilistic result instead.

18. *Forecasting of negative ramp rates* are required to determine how many MW scheduled capacity is required to balance the ramps. Statistical methods used to make empirical relationships between weather forecasts and measurements have a tendency to suppress the ramp rate and can lead to that insufficient capacity would be allocated for the ramp itself.

19. Tuning on correct *frequency distribution* of the power generation is a first order approach to get reasonable correct ramps. This constraint works reasonable, if the measurements are time filtered to hourly resolution. Then the model and measurements will contain about the same amount of ramps. It is the more frequently observed power output ranges, which determine the power curve as they have the highest weight in the estimation process. The ramping is less accurately estimated,

because there are few hours with near half-generation and more hours with near no-generation or full generation and is mostly high forecast uncertainty of these.

20. The *minimisation of phase errors* should in theory be possible by giving weight to the ensemble members with the lowest dispersion term in the long-term error statistics. This has been tried, but resulted in an improvement on the dispersion term only on the second decimal. From that finding it can be concluded that there is no pattern in which forecasts have a positive or a negative phase error. Attempts to subjectively correct for phase errors are therefore rather likely to fail. This test was applied to the 2-year data set from the project (including the historical data) at all horizons up to 48 hours. We conclude from this that phase errors are random and most important model formulation independent.

21. Ensembles with large amounts of members give the best statistical measure on the likelihood of sudden generator stops at extreme temperatures or load, if the same physical value is used as threshold in all ensemble members. Some members over-predict and some under-predict. This gives all in all the best risk evaluation in extreme events.

22. Many hours have been used to study the *flow pattern* by comparing potential wind power based on the model wind speed in 100m altitude with with the actual measurements. This leaves the impression of a sometimes extremely scattered wind field.
    Each forecast error or "scattered measurement" appears as a triangle if the background color, which is the forecast and the foreground color, which is the measurement, deviate from each other. Two years of data (17000 hours) has been studied. The conclusion is that the wind field is very often scattered and that a 10-20km resolution model system will not be able to produce a correct scattered field. If it would, the model would be defined as contaminated with numerical noise. Only higher resolution modelling can and should produce the same scattered wind field as the observations show.

23. The large scale flow in the facility region has a structure that repeatedly builds up *gradients in temperature and wind speeds*. This also takes place in a model resolution of 100km, because it is controlled by large scale flow. If we take the 20km resolution as an example, then each model is again and again building up strong temperature gradients that it cannot handle correct. The gradients imply numerical truncation errors in the dynamical equations and develop wrongly. The 6km model formulation can deal with at least 3 times stronger gradients before similar truncation errors are introduced, thus the model is more correct. The question is, if it pays off to further increase the resolution. For the time being no, because the weakest point should always be eliminated first and system H has initial conditions as the weakest point.

The above list of findings has been compiled in order to give assistance to the design of a future forecasting system. Appendix A and B contain a more detail analysis of the

verification behind most of the findings mentioned above, but some of the findings are rather result of event analysis.
We have also explained how the southern part of Alberta plays a central role for weather forecasting for a large fraction of the northern hemisphere. A large number of the low pressure systems that spend their life at the midlatitutes are born in Alberta. Some of them were maybe perturbations on the 500hPa flow over the Pacific, but the mountains essentially convert them to low pressure systems that are visible in a deeper layer of the troposphere.

2.4.1 Objective Verification

In this section we will only present and discuss the most important results and associated findings. The detailed statistical results are presented partly in Appendix A and Appendix B, which all details the power conversion process that was used to make a fair evaluation for the 3 months period.

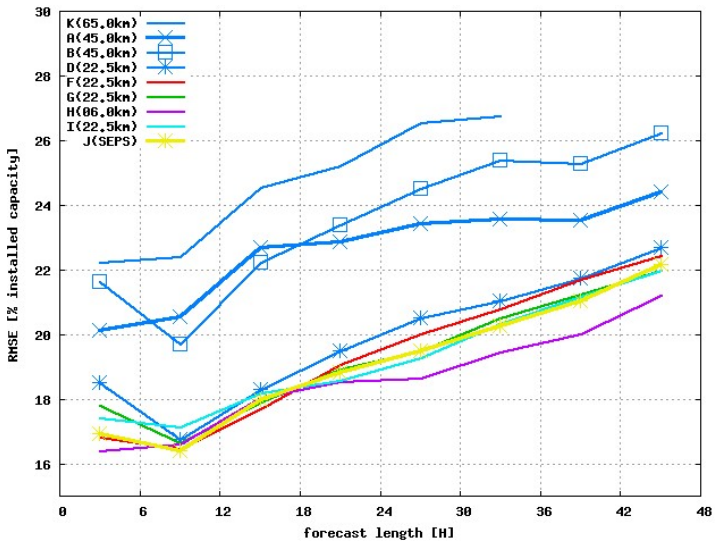

Figure 2.7: Error statistics over the forecast length 0-48h for the delivering ensemble systems (A-D), the experimental ensemble setups (E-J) and the large-scale North-America ensemble K.

RMSE Verification Results

Figure 2.7 shows the root mean square errors of the systems that delivered to the project (A-D), the experimental systems (E-J) and the large-scale 65km North-America ensemble. The impact on resolution, both on the forecasting systems resolution, but also on the boundary providing system to the forecast error development over the forecast length is significant. The blue lines of the K(65km), A and B (45km) systems demonstrate that this resolution is unsuitable for Alberta. The 4th blue line with stars, the D(22.5km) system

that delivered one optimal forecast to the pilot project is the system that performed worst among the 22km resolution systems (F,G and I), which however show similar performance. It is interesting to note that the CMC based ensemble with the "fake" analysis at 06 UTC and 18UTC does not perform worse than the other two systems.
Nevertheless, it is obvious that the high resolution ensemble H has a clear advantage over all other systems in the day-ahead forecast, while the results in the first day are screwed up by problems in the initial conditions (see also 2.4.2).
In Appendix B.4 and B.5 the advantage of the H system is discussed in more detail. It is worthwhile mentioning, that the dispersion, the part of the RMS error that describes the phase errors is lowest in the H system, while the SDBIAS, describing the variability of the system is strongest for the H system, when compared to all other system. This is somewhat surprising, as higher resolution models normally have a higher variability and higher phase errors than their coarser resolution equivalents. The reason why this is not so, seems to be due to the inherent "smoothing effect" of using an ensemble instead of a single model.

Forecast Error Prediction

The ability to predict errors can be measured with the correlation between forecast spread of each ensemble and the actual error of the forecast as verified above.

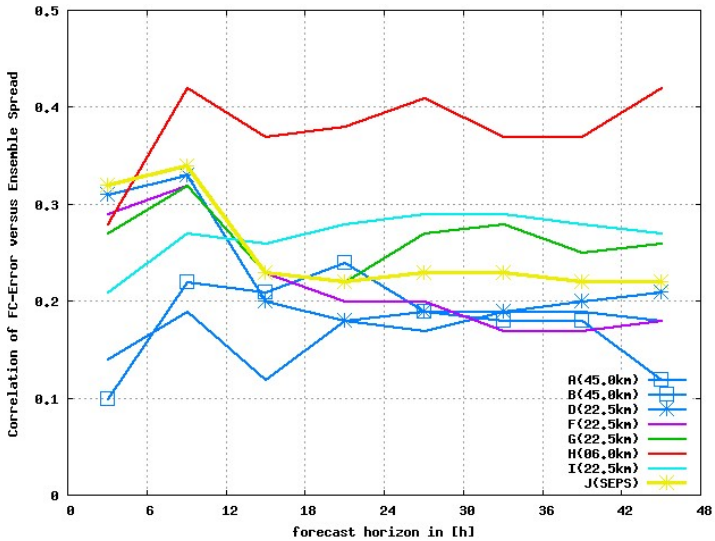

Figure 2.8: The figure shows correlation between error and ensemble spread over the forecast length 0-48h for the delivering ensemble systems (A-D), the experimental ensemble setups (E-J) and the large-scale North-America ensemble K. The higher the correlation the better the ensemble members warn about when the error is going to be high. System H has therefore very good skills to predict price volatility. One would in real life use the ensemble spread to define the amount of reserve required for wind power.

Figure 2.8 shows that the higher resolution systems, especially system H, are significantly better than the coarse resolution models (A,B and K). However, it is interesting to observe that the correlation hardly reduces with forecast horizon for the coarse resolution models and the very high resolution system H, while there is a deterioration of correlation from day one onwards in D and F, but G and I does not degrade with horizon. This shows that the high resolution lateral boundaries are the important factor for good ensemble spread on the longer horizon. A correlation between error and spread that is 25-30% better for H is a very strong skill. A correlation of 0.4 is not perfect, but exactly that result made the experiment worth the effort, because it shows that the model system have errors for an acceptable reason, which is uncertainty in the initial conditions and some model error. This is much different than the other systems with truncation errors. They cause the whole ensemble to be incorrect and not even very good predictors for the error. We could not have known this without doing the H experiment where we have eliminated the truncation error.

What is characteristic for D, F, G and I is that all of them produce power at the right time, but often not the correct amount, because they tend of have the same wind everywhere in the facility area. Thus, all facilities have some power, while system H rather give full generation at some facilities and no generation at other facilities. This is more representative for the actual generation pattern.

The constant correlation is important for the allocation of reserve. Although, the correlation is still relatively low, there is considerable daily high frequent moderate amplitude variability, which drags down this correlation. This background error is not dependent on the mean wind speed, but on a variety of meteorological and non-meteorological factors. The result also suggests *that reserve can be bought day ahead* and on a shorter notice in more frequent intervals, because the accuracy is independent of the forecast length. This shall be understood as follows. If forecasts have the correct ramping structure, then there is no correlation difference between the members that represent the actual uncertainty, because most error is due to phase error. This is demonstrated in B.

Frequency Distribution of the Wind Power Forecast Error

Figure 2.9 shows the frequency distribution of the wind power error in the project period 05/2007-04/2008 for system D. The blue colored bars denote the frequency of the forecasts, where the observations were within the minimum and maximum (spread) forecast of the ensemble. The bordeaux colored bars are frequency in which the observations were lying within the lower and upper 5 ‰ of the low probability bands. The purple colored bars denote the frequency of outliers, i.e. where the observations were outside the Ensemble spread. This type of frequency distribution indicates how well the ensemble is capable of defining the uncertainty of the forecasts correct. The ideal distribution would have only blue bars, indicating that the observations always lie within the ensemble spread. Additionally, the blue bars would be Gaussian distributed, indicating that there are equally

many hours in each range of the production intervals. The distribution is however flat with intervals near the extremes being most populated, because of the lower and upper boundary of the power curve. The ensemble spread is one-sided at every hour in which there is either no or full production and the ensemble spread is only symmetric during hours with medium generation.

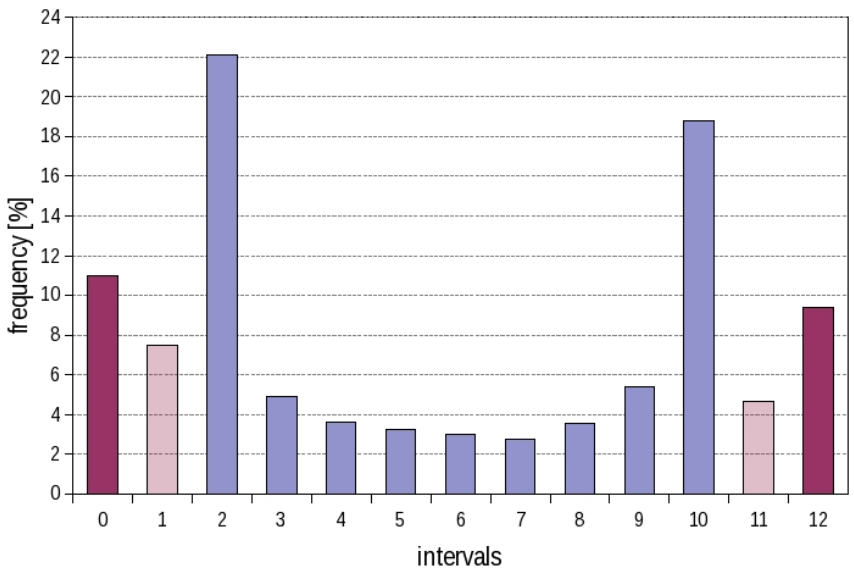

Figure 2.9: Frequency Distribution of the wind power error period 05/2007-04/2008 for system D. The bordeaux colored bars are the lower and upper 5‰ and the purple colored bars are random outliers. Bin 1 contains the events where the true value was under the minimum forecast (equivalent to over prediction). Bin 12 contains the events where the true value was above the maximum forecast (equivalent to under prediction).

Extreme Ramp Analysis

Steep ramps can be generated by either cold fronts from north or from non-stationary lee waves in westerly flow. Both will continue to cause some risk in the future. The up-ramping is not so critical as the down ramping, because other power producers can stop their production, or turbines could be curtailed if there is a risk that the wind generation will ramp down very shortly again. In the down-ramping case there could be production capacity missing to fill the gap.

To analyse the ramping capability of the MSEPS probabilistic forecasts that were not delivered to the project, we extracted those events, where the measurements showed a change in power of equal or more than 80MW per hour in a two period. The 80MW is not in itself critical, but if it is scaled to future capacity levels, then this threshold value is

realistic. Instead of only looking at the minimum and maximum, we now verified, whether any at least one of the ensemble members ramped correct.

Month 2007/08	positive ramps	negative ramps	Total per month
May	3	1	4
Jun	4	4	8
Jul	1	1	2
Aug	6	2	8
Sep	2	5	7
Oct	1	1	2
Nov	5	2	7
Dec	1	2	3
Jan	3	0	3
Feb	6	4	10
Mar	6	3	9
Apr	3	3	6
Total	41	28	69
Avr	3.4	2.3	5.7
$< +/-100MW$	23	17	40
$> +/-100MW$	18	11	29
Avr (MW)	98.58	-96.83	+/-96.47
Max (MW)	146.53	-124.82	-
Min (MW)	87.58	-88.21	-

Table 2.2: Overview of the observed extreme ramp events $> 80MW/hour$ in the project period May 2007 - April 2008.

Table 2.2 provides an overview of the extreme ramp events found within the project period May 2007 - April 2008. There were 69 events found with a ramp rate of $> 80MW/hour$, which were studied in more detail.
We also studied, whether these extreme events had specific characteristics such as season or time of the day. While there was no significant seasonal pattern found, one characteristics is that the ramping events have the same occurrence rate as the Chinooks in Alberta, and the recorded events can almost all be connected to a Chinook event. The exceptions to this are associated with cold fronts from North. Another characteristics is that the ramps are less likely in the winter months, where there are longer periods with strong wind and near-full production. December and January had very few ramps.
The test was carried out for hours of observed extreme ramp events and allowed for:
- magnitude errors of $< 5\%$
- phase errors of 0h, 1h, 2h, 3h and 4h

ID	allowed phase error	% of 80MW/h	fail rate [# / 69]	fail rate [%]	success rate [# / 69]	success rate [%]
1	+/-0h	95%	25	36.2%	44	73.8%
2	+/-1h	95%	13	18.8%	56	81.2%
3	+/-2h	95%	9	13.0%	60	87.0%
4	+/-3h	95%	7	10.1%	62	89.9%
5	+/-4h	95%	7	10.1%	62	89.9%

Table 2.3: Summary of the extreme event statistics. The test included 69 extreme events and allowed for 5% magnitude errors and 0h-4h phase errors.

Table 2.3 summarises briefly the results. It should be noted that the amount of correctly forecasted extreme events over one year increased from 73% for a 0h phase error allowance to almost 90% of the extreme events when allowing for a phase error of 3h. There was no improvement found, when allowing for 4h phase errors, which is most likely due to events that were not forecasted at all. This result is consistent with that we had seen events where we regarded the 22km ensemble as completely failed.
A contingency table was generated and a number of statistical tests carried out with the input from the contingency table. These can be found in Appendix A.11.

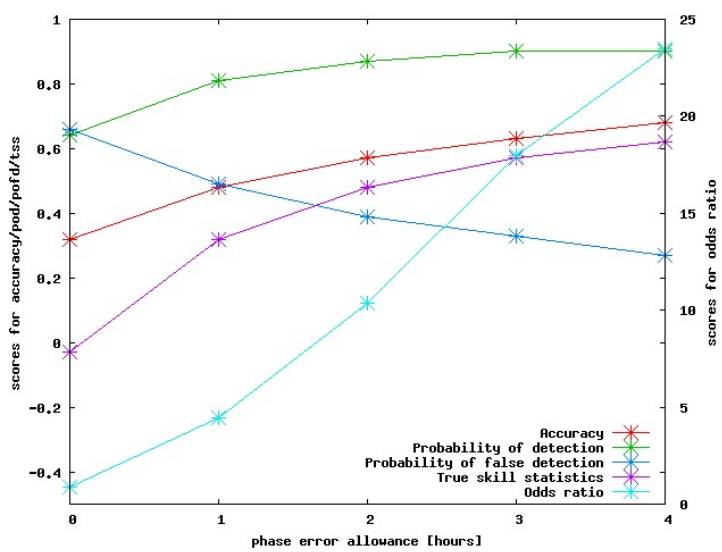

Figure 2.10: Results from the extreme ramp event analysis carried out for ramp events $> 80MW/hour$.

The questions that we tried to answer were:

- What fraction of the forecasts was correct?
- What fraction of the observed *yes* events were correctly forecasted?

- What fraction of the observed *no* events were forecasted as *yes*?
- How well did the forecast separate the events from the non-events?
- What is the ratio of the odds of a forecast being correct or being wrong?

The test verified, whether there is any one ensemble member that matched an observed ramp of that size with an allowed magnitude error of 5% and different allowed phase errors (y-axis). The verified tests were conducted with help of a contingency table containing hits, misses, false alarms and correct negatives over one year.
Figure 2.10 shows the result of the statistical tests carried out with the data from the contingency tables. The first question refers to the accuracy measure (red line), the second question refers to the probability of detection (green line), the third questions refers to the probability of false detection (blue line), the forth question refers to the true skill statistics (pink line) and the last question refers to the odds ratio (light blue line). Further details on the tests can be found in Appendix A.11.
What is characteristic to observe is that for the accuracy, the probability of detection, the probability of false detection and the true skill statistics tests the curves become flat and seem to have a threshold at the 3 hour phase error allowance. That is, when looking at all ensemble members and allowing a phase error of more than 3h, no further improvement can be achieved. In other words, if non of the ensemble members has a ramp within a range of 3-4 hours, the event is not predictable/predicted by the forecasting system day-ahead.
This is an interesting result and confirms the observations from Table 2.3 that there are approximately 10% of the events that seem to be unpredictable and that while the "odds" that the event might be forecasted still increase, the probability of detection, the probability of false detection and the true skill statistics remain nearly constant.
This is an important finding, although we count it as a hit if just one of the ensemble members had a matching ramp. However, it is encouraging to further study the possibilities that exist to make sure the end-user has access to this kind of information.

2.4.2 The requirement for high spatial resolution

Throughout the project, we discovered a number of cases, where it was obvious that the atmosphere had motion on a scale that is impossible to simulate without higher spatial resolution.
A first step in this project was then to demonstrate that a high resolution ensemble in 6km or 3km horizontal resolution can run stable. This was not obvious as we are not aware of an operational 3D model system that runs with nearly 4000m tall mountains between British Columbia and Alberta before except in research experiments. Figure 2.11 and Figure 2.12 show the orographic differences and the sensitivity of the wind speed and potential vorticity of the 45km, 22.5km and the 6km resolution models A,D,G and H, respectively.

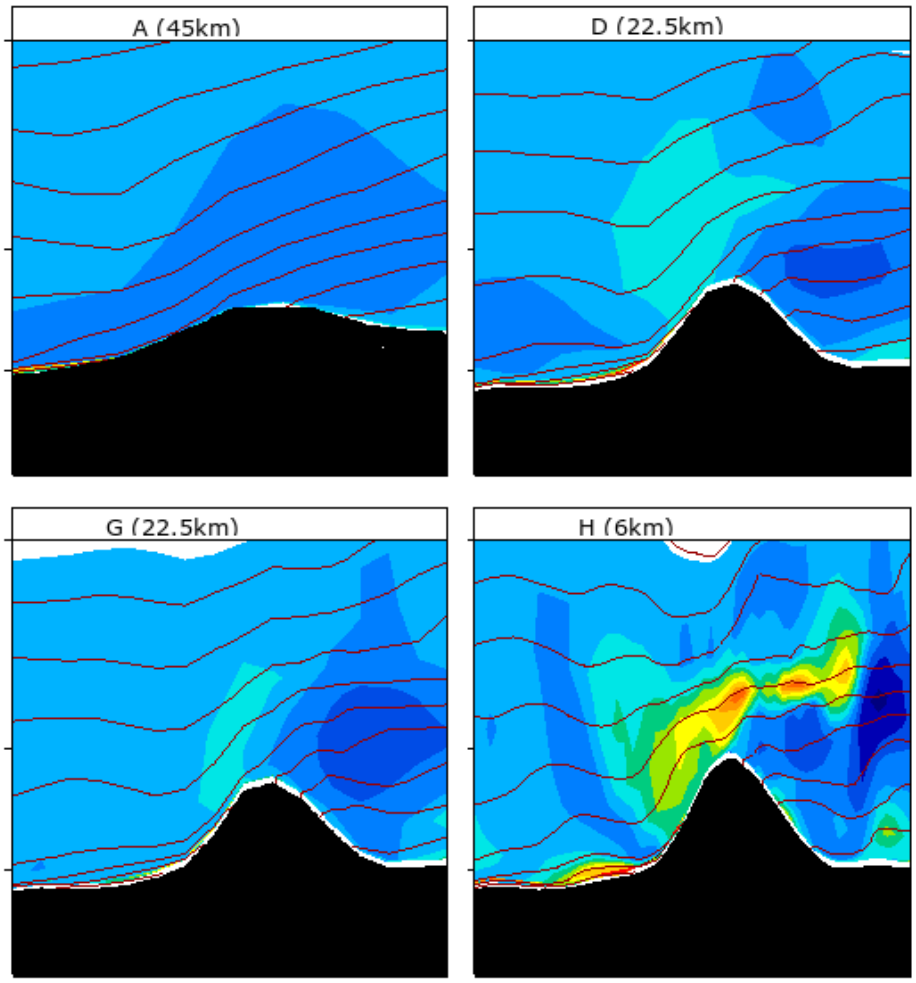

Figure 2.11: Cross sections including the model orography of potential vorticity in the model systems A(45km), D(22.5km), G(22.5km) and H(6km). This is just before a very steep ramp, where one would expect to see some high values like H has it. There is also a strong potential vorticity signal present in the flow on the BC side of the mountains.

The tests were successful, although there were a number of events, where the model time step was reduced by a factor of 3, because the vertical velocities of the lee waves were so extreme that the vertical stability criteria was exceeded for the advection term and some models had to restart with lower numerical time steps. This is directly related to the steepness of the mountains in the model. What was a horizontal wind became suddenly a vertical component on the lee side of the mountains, where it is accelerating and maybe developing to a Chinook.
The next task was to demonstrate the quality of the high resolution models. Here, it was encouraging to see a very small model bias. The roughness in a period with snow cover may though have a different bias than in summer. By the time of writing up the report

there is no result for summer conditions.

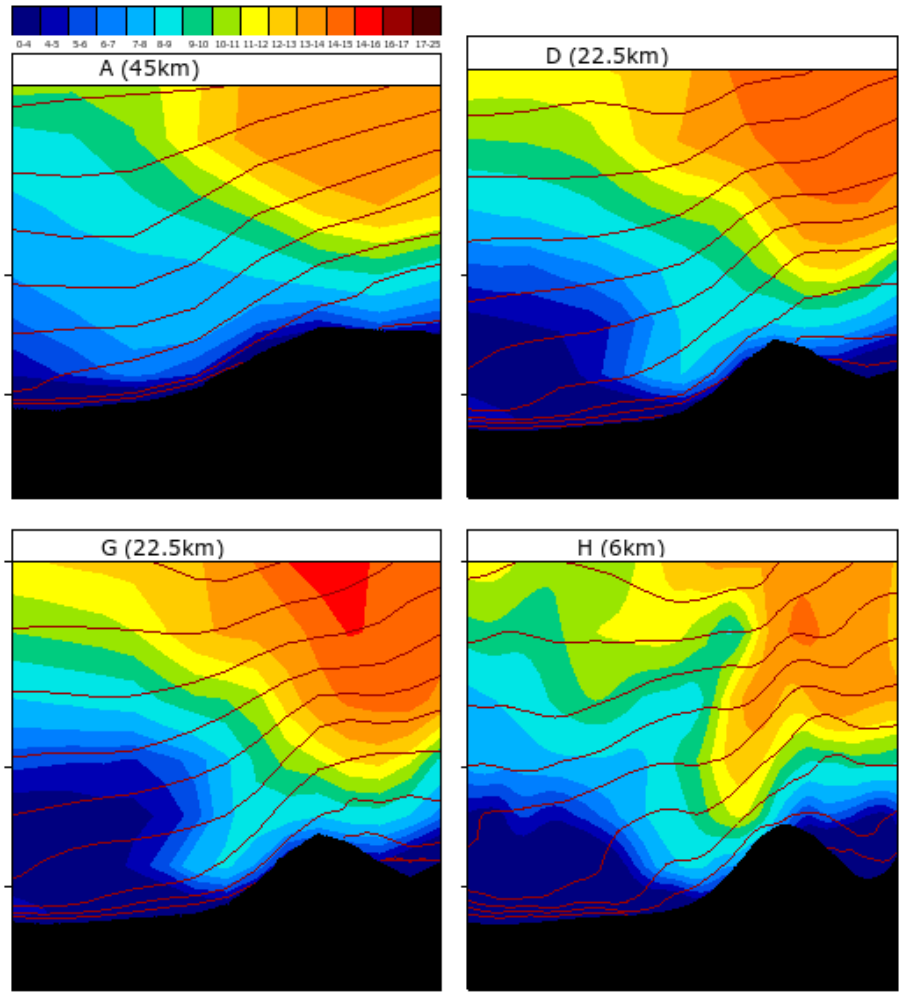

Figure 2.12: Cross sections of wind speed in the model systems A(45km), D(22.5km), G(22.5km) and H(6km).

In figure 2.7 it can be seen that the error of the day-ahead forecast improved much more than the current day. This is a direct indication of that we need to focus on the initial conditions. It is however not a surprise that the initial conditions cause trouble. Starting a model state in 50-100km resolution with 6km model resolution grid with very different shape of the mountains is not trivial to do right.
The most obvious to take hand about this is to compare the large scale state with the small scale state in pressure coordinates and determine the high resolution state that fits the large scale best. This will not introduce imbalances in the model. There is plenty of room for improvement and exactly this type improvement will improve the forecasts on the 1-12 hour horizon.

The impact on the short-term prediction is also expected to be positive, but no objective verification has been made so far. The major benefit is that the wind speeds can be trusted. The operator can directly evaluate the cutoff risk on region basis from horizontal maps of wind or potential wind power computed with reference power curves applied in every grid point as shown in Figure 2.13 and Appendix C. The 6km ensemble does also ease the quality check of measurements.

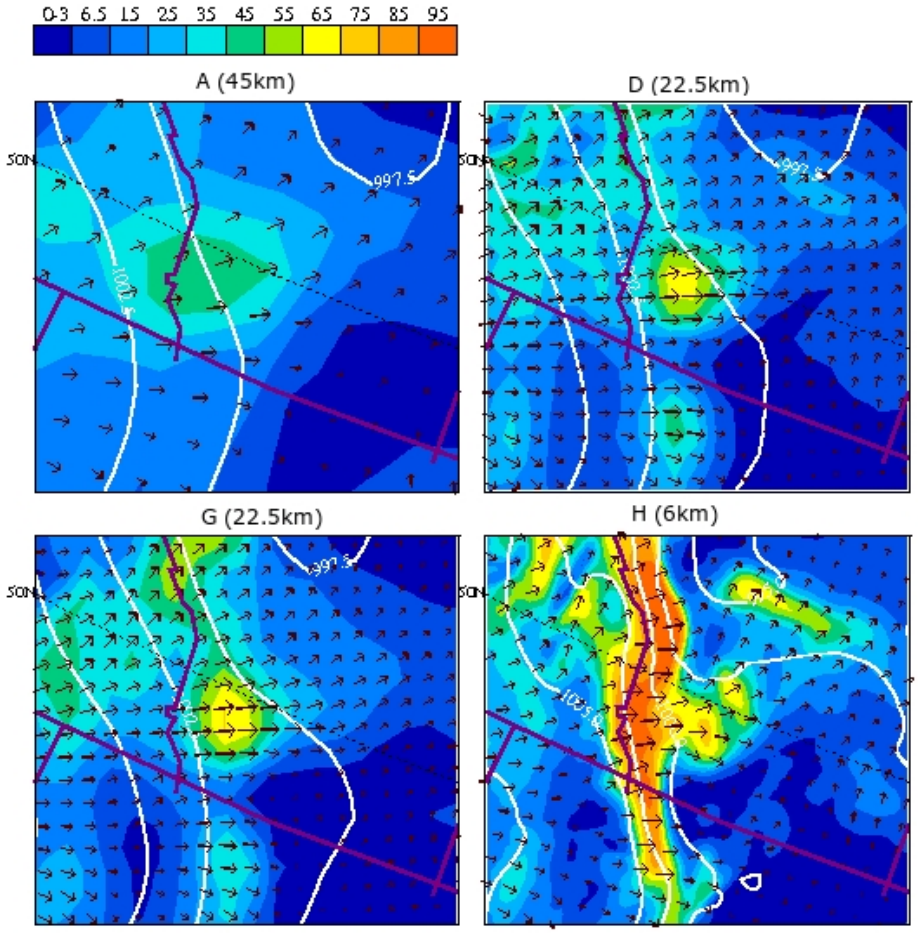

Figure 2.13: Horizontal maps of potential wind power in the model systems A(45km), D(22.5km), G(22.5km) and H(6km).

2.4.3 A mini Cost-Benefit Analysis

Although the 6km ensemble is computationally expensive, it should be noted that the setup can run without any statistical correction. The wind speeds are natively of the correct magnitude. Thus, the power can be computed with a physical method that describes each of the facilities. There will be almost no need for recalibration and reruns during

and in connection with system updates. The forecast for the new facilities will be of high quality from the outset. The maintenance of the system is therefore cheaper. This makes the software around the system simpler and subject for much less updates.
Wind facilities will be described with coordinates, size and a reference power curve. There will from a forecasting perspective be no need of building a large databases of forecasts for the purpose of statistical training and forecasting. The probability density function will give the required statistical information and will allow for smoothing of the final forecast and calculation of required reserve.
A cost benefit analysis of the 8 setups therefore points to the advantage of setup H, if the price is calculated over 3 or more years. It is sometimes tempting to dampen errors with a statistical method, but that is often done equally well by just using the average of a large ensemble, which is much simpler than statistical data handling systems.

2.5 Brief Summary of the Results from the MSEPS Experiments

In the investigated period there were many difficult events, where the 22km forecasts reacted fundamentally wrong and the 6km ensemble at least had the correct structure and where also some members had the right evolution. It is in fact our strongest argument against the 22 and 45 km runs that they are sometimes all wrong. It does not give confidence to the forecasting if the same equations solved with more accuracy can turn out so different as they do. We can hardly say that a given set of equations are good or bad, because the performance is so sensitive to accuracy of the solution method. The more computations, the better the results. At the end of the project we have some uncertainty on how small we we can make the high resolution area. It is not impossible that the area we have used is bigger than required. The resolution of a limited area ensemble will be 10-20 times higher than a global resolution and 3-4 times of the highest possible resolution of a global deterministic model at a given point in time. All can probably scale with the evolution seen from a technical point view. It takes probably of the order 6 years to double the resolution. A global ensemble forecasting system will therefore maybe reach the resolution we managed in year 2008 in about year 2030. The limited area ensemble is therefore a much more cost efficient approach to address the weather forecasting than global system regardless of we think deterministic or probabilistic. A fundamental problem for the global model systems is that they approach resolutions where there is very little energy on the smallest spatial scale. This means this scale is only adding value at places with frequent extreme gradients like in Alberta, but instead the resolution increases the risk of numerical instability. In the bulk of the atmosphere this scale will not give any positive signal on accuracy, because what was initially observed was not accurate and there is no inertia on these scales, so their benefit is only local. So it is questionable if it pays off for the global models to increase the resolution much further than 10km. If this argument is right, then it is also not a solution for Alberta to look into global modelling, because we have shown resolution is significant.

If a given global model system should turn out to be good in year 2008 compared to a similar resolution limited area model system, then care has to be taken in the extrapolation. The global model might never reach the same level as the limited area model system. One could say that it is more certain that a limited area model system is scalable. Short-term forecasts can be improved significantly more with the help of a correctly structured ensemble, which produce correct ensemble spread. The short term forecasting process must be able to pre-screen the measurements for obvious mistakes, but if the error is not representative, then larger acceptance intervals are required. This may lead to misleading forecasts. Successful short-term forecasting requires correlation between error and ensemble spread, because the ensemble spread is the most important term in the short term prediction. In fact this is where the 6km ensemble has shown its major strength relative to the 22km ensemble.

We have experienced that a framework of many model configurations is required to explain model errors. We have seen how small differences between A, B, D, F, G, H, I and K could be explained consistently. This was only possible, because there were in fact 75 x 3 months simulations behind the result. This is equivalent to nearly 19 years of concurrent simulations for each of A-K. Single model configurations would have led to a random signal and consequently low confidence to the conclusions for a 3 month period. Running an ensemble can therefore be considered as a robustification of the methodology and also the validation.

The disappointing part of this result is that we must expect that we can improve significantly on the first 12 hours, but there is a limit for what can be achieved at longer horizons. The construction of the facilities will have some impact on how much the day ahead error can be reduced. However, the wind investors have rather an incentive to keep wind power unpredictable by pure weather forecasts from a market perspective.

Centralized versus Decentralized Forecasting System Integration

The decentral solution discussed in this Section refers to a solution, where each wind facility owner is responsible to contract one or more forecasters and deliver regular forecasts to the AESO.
A centralized system in this Section refers to a system, where all facilities report data to one particular host at which forecasts are either produced or collected. The centralized solution would normally produce wind power and make short-term forecasts 0-5 hours ahead, but may or may not include any in-house NWP modelling.

3.1 Decentralized Forecasting System

The project has demonstrated that the forecast error does diminish with increasing wind power capacity and dispersion level. Although it would in theory be possible to also predict for all wind power in a decentral solution, there could be confidentiality issues on data, which would block for access to the measurements in a decentral solution.
Another aspect is the difficulty in improving the forecasting process based on singular measurements. Every improvement has to be expected to be on the $2^n d$ decimal of a large error on the individual sites, whereas improvements on the aggregated wind power could be on the 1st decimal and on a relatively smaller error. Improvements cause costs that add to the cost of forecasting, which have to be recovered over the operating costs. The more MW are included in a forecasting solution, the more and faster costs can be recovered.
For security reasons, the AESO would also need a forecasting system, because forecasts could be erroneous or missing, or of very bad quality. The penalty for low forecasting quality would also have to be relatively high to prevent that the facility owners would save on the quality of the forecasts. The wind facilities with the lowest capacity would then pay the highest price relative to their delivered wind power. Such a situation would pose an unfair cost factor onto these wind facility owners and may prevent community projects and support.
In general, such a situation will increase the costs of wind power integration and operation unnecessarily and could lead to situations, where large clusters of wind farms are planned to save on the forecasting costs. This is however not desirable, in contrary, it is desirable to disperse new find facilities as much as possible within the area of high wind resources.

A decentral forecasting solution is therefore not recommended for Alberta, as it would make forecasting even more difficult, with lower improvement rates and higher costs.

3.2 Centralized Forecasting System

A centralized solution is more reliable and accurate than a decentral solution. Although the decentral solution may consist of independent sources, it can nevertheless happen that all have the same single point of failure. This would only be realised when it happens as the individual parties may not reveal how or where they purchase their forecasts.
Considering the complexity of forecasting in Alberta and particular grid security issues, a central solution is required, where each supplier of weather forecasts should be required to reveal their data sources for redundancy reasons.

3.2.1 Benefits of a Centralized System

A centralized system can be tendered and functional requirements should be specified as mandatory, desirable and expandable as possible. The system could essentially be managed by a local hosting centre or via a system integrator in AESO.
The system can be built in phases and expanded with the growing requirements, once the basic design has been made including future compatibility.
A centralized solution can be more advanced and fully automatic. It will be capable of producing reliable results also if a large number of online data fall out, as independent data streams would be untiled.
All data are directly available. This gives the operator immediate access to results.
If the operator is equipped with graphical presentation of a probability density function, the actual measurements, a number of forecasts (3-5) derived with known optimisation targets, then the operator can essentially choose the optimisation target by pressing a button. The chosen strategy then goes into a decision process which will derive the required generation mix. It is like an automatic gear, where the user is better off to not change the optimisation target too often. Each change will have a cost associated as the generation mix may have to be re-scheduled.
We consider this as a realistic, economic and future compatible strategy for electrical grid optimisation.

3.2.2 Redundant Systems

There is a tradition for redundant IT systems for critical applications. Some parties prefer to run exact copies and others prefer to run the secondary setup as very simple system in order to save on the maintenance cost and increase the reliability of the secondary system. One can even argue that a cluster has an inherent redundancy as the cluster consist of a number of servers.
WEPROG's strategy is to use two very different systems in order to maximise redundancy and even benefit from both in the daily operation as they give slightly different results.

It should be noted that the complexity and requirements in terms of resources are very different for the weather prediction and the wind to power conversion. The higher the spatial resolution of the weather forecast the more resources are required in that step, but the less in the power prediction step, because the value of the statistical computation lowers with higher spatial resolution of the weather forecast.

The power prediction step will therefore remain a relatively simple process in comparison to the weather forecasting. This means that the costs of running the power prediction as a redundant system is low, while the weather prediction part may require much larger resources.

The cost of running the weather forecasts in the centralized system however still low compared to the total operating costs of the wind power system. Since there are additional benefits particularly on the visualisation, this such a solution is preferable over external weather forecasts in the medium to long-term.

Technical Requirements of a Future Forecasting System

In the pilot project a large amount of online measurements were made available to the forecasters. These included not only power measurements at the wind facilities, but also meteorological variables such as wind speed, wind direction, temperature and pressure including statistical measures of some of these.

What can be stated is that there will be much more value of these measurements than what was achieved in the project in a post-project phase. We will in the following describe what was done to show why we do not feel we took great advantage of the meteorological measurements for the power prediction. Although it may sound like a waste of resources, if this was not done, we believe that it was the right choice for the pilot project's forecasting of wind power to not couple weather and power measurements, because the power measurements had a higher availability in real-time. Evaluation of the forecasting would have been even more complex, if there would have been more irregular data included in the forecasting.

It is rare to be provided with meteorological measurements in real-time forecasting. Therefore, WEPROG does not have tools available to use such measurements for other purpose than predicting wind from wind, temperature from temperature, power from power etc. It was a requirement to deliver such forecasts and it was done, but without any attempt to couple information from variable to variable and thereby gain accuracy. A rather simple and variable independent algorithm was used with out any calibration.

The only assumption in this algorithm was that the ensemble spread is assumed proportional to the actual error at time zero and this relationship decays linearly with time over a period of 5 hours. Obviously, such a relation can be written in a few lines of code and recomputed in one second for many sites once there is a measurement. The important and crucial step in this process is therefore to make a representative ensemble spread, while the rest is kept on a simple and computationally efficient level.

The methodology is very easy to use in industry, because only 3 values per hour are required: the ensemble mean, minimum, and maximum from a short-range ensemble, which are then combines with measurements as described above.

It is worthwhile noting that this is not how we intend to solve this problem in the future. However, we expect that there is a lot more accuracy to be found by combining all information in one set of equations that is solved iteratively. This will identify a state estimate from which we can integrate forward.

Our expectation is therefore that short-term forecasting can be improved significantly

by coupling information and using the ensemble weather forecasts also, because some ensemble members might be nearly correct during the recent hours and their evolution is then likely to be the best predictors for the next hours. Some value of this has been demonstrated already in Ireland and will certainly become more important in the future. The two next sections will describe what measurements we expect to be a requirement for the day-ahead and for 0-4 hour forecasting in a future forecasting solution.

4.1 Measurement Requirements for day-ahead Forecasting

Beyond the 12-hour horizon the important input to any forecasting model are the historical data of total power prediction for each facility. Knowledge about periods with restricted power generation or reduced turbine availability are also of some value. It was found that weekly availability forecasts has improved forecasting in e.g. Australia.
The turbine availability in the historical data is only of real value, if it is possible to get forecasts of turbine availability also in real-time. Without such availability forecasts it is required to assume average availability. This is however inherent in the statistical calibration.
In southern Alberta there are many significant temperature changes around zero degrees Celsius during the winter period. This can cause the snow depth in the model system to be estimated incorrect, which again has impact on the wind profile. Snow depth measurements would therefore add value for forecast horizons up to 3 days. A single measurement is however not of much value for the wind profile, but rather the snow depth over a large area. It is expected that Environment Canada has such measurements available.
Irrigation in spring and early summer may also be difficult to separate from natural precipitation and evaporation in the forecasting models and can cause a positive model bias. However, statistical training with data exclusively from the irrigation period may correct for the forecast errors that arise from irrigation. It is therefore not considered important to obtain irrigation forecasts from the farmers once the system can be trained with data from exclusively May to July .

4.2 Measurement requirements for short-term forecasting

The short-term prediction is essentially a *data assimilation problem*. The target for any forecasting system of that type is to find the best possible state estimate using as many measurements of different type as possible. Some will be rejected, because they are erroneous or the model uncertainty is so wrong that the measurement is rejected. However most measurements will be used to determine the actual state.

It is often very difficult to get an understanding of the weather situation from the power measurements alone, because local downdrafts from the Rockies confuse the overall picture. However, the high-resolution ensemble system can provide a good overview, if it is fed with a sufficient amounts of different measurements across the facilities' area. From one model state, it is impossible to get a good overview. When comparing with several hundred forecasts from the past hours, it will be possible to find a state that fits the real state of the atmosphere very well. The ensemble spread around this state is then used to screen the the measurements in order to prevent that unrealistic measurements can introduce forecast errors.

A good geographical spread of the measurements is very important. It is however also important to have facility representative measurements for control purposes. Wind turbines can fail and independent measurements are therefore of value. The flat regions of the power curves also generate a difficulty in estimating the actual wind speed in the area alone from power measurements. This is especially a problem in the short-term prediction, as it make a difference, whether there is $14m/s$ with a tendency downwards or $23m/s$ with a tendency upwards. Both result in less power, but the cut-off wind speed of $23m/s$ can cause a steeper ramp. All variables help and the price for additional variables such as humidity and temperature is not high, if there is already a supply of some variables from a given site. Nevertheless, wind speed and direction in two altitudes are the two most important variables.

The numerical time-steps of the 6km ensemble is of the order 30-90 seconds depending on the model formulation. This would allow for the possibility of forecasting significant (sometimes) trusted changes in the wind speed in about 10 minute intervals at sharp fronts. Consequently, the measurements should report either every 5 or 10 minutes. Typical events where this would be of value is if a sharp cold front from north would hit a 300MW facility. Then ramp rates would not only reach 80MW per hour but up to 300MW in 10 minutes. It is not unrealistic to warn about such events based on a 6km weather ensemble. Several facilities have been ramping with their full capacity in 5 minutes during such events in the winter 2007-2008. Similar events can also happen in summer but less likely.

As we have pointed out in the industry recommendation, it would not be good practise to have 300MW installed elsewhere than in the middle of the South-Centre region. If a 300MW facility were to be installed on the northern side, then there would be a need to raise at least two masts with independent SCADA systems as well on the northern side, such that the forecasting system can be warned about nearly instant 300MW ramps. The same applies, if such 300MW were installed on the southern part of the South-East region. Here, the masts would have to be well on the southern side of the facility.

The reason for suggesting that a number of independent data sources should be made available by the wind farm operators is because the requirement on availability of the individual site is then lower. This reduces the cost of maintenance for the wind farm operator and gives most likely the best value for money. It should be the wind farm operator that decides which measurements are made availabile. There should however be requirements to the wind farm operators to report on their measurement availability

and announce changes, such that the AESO operator is always able to keep overview, not only in general, but also in the daily operation, e.g. on a Saturday night, if a number of measurements have stopped reporting and are not likely to come online before the following Monday.
We do not mention RADAR measurements, because this measurement type will contain a single point of failure. This should be supplementary information that the operator may take advantage of, if everything else brakes down. RADAR may also be coupled in the future, but single point failure systems should not be given highest priority unless they are better than other sources.

4.3 Additional useful Measurement Types

A small number of LIDAR measurements in the South-West region may also be of value to improve forecasting, because these measurement types are useful to detect the wind profile in a deeper layer of the atmosphere than masts can. A 300m mast will be equivalent, but is less cost effective. A mast will not trace as much as a LIDAR. However, large concurrent ramps will be traced also by a single 300m mast in due time. Again, the mast can be more redundant equipped than the LIDAR. There can be different instruments added. A peculiar measurement from a LIDAR cannot be checked against anything else than the forecasts, whereas different instruments on a mast can be used to check the reported values. The mast is therefore more reliable even though it does not reach as far up. The usefulness of a LIDAR can be detected quite fast and is cheap compared to the cost of a tall mast.
The vertical wind profile is particularly important in the winter, where nearly all models have a trend to mix up the stable planetary boundary layer in daytime, where it does not always happen in this way or as fast as expected.
A centralized system can also benefit from receiving all SYNOP and upper air measurements from Environment Canada. They will complement the measurements within the facility region. Most of them will be reporting at lower frequencies. But again, the aim is to not be dependent on specific data types with one single point of failure, but rather a broad range of independent data sources.

4.4 Overall recommendation on Measurements

Based on the results of the high-resolution ensemble, the data processing of non-stationary measurements have simplified. There are no critical model bias due to topography or roughness differences between model and physical world. There is no need for calibration on the specific measurements. Each ensemble member can be forced towards the measured values without causing model imbalances, if the measurement lies within the ensemble spread at the time of where the measurement is valid.
The overall recommendation is therefore that the requirements for measurements should be kept flexible and the cost of these should maybe be paid over MW-based tariffs. The

responsible consortium of wind power should then use the budget strategically for what can be made available and is most beneficial for the forecasting process. The project has assembled enough measurements and experiences to quantify the value of the individual data sources.

Therefore, it is worth the while to keep the two years of data as a reference to estimate the value of future measurements. Improvements to the forecasting system should therefore be validated on this period until a better and more representative period is available in a future operational environment.

Methodology and Approach

If the wind generation profile and demand would have a high correlation every day, it would not be difficult to integrate large amounts of wind power in the grid. This is unfortunately not the case in Alberta. Alberta's wind power facilities are located within a small geographical area and variable wind speed. The variability of the wind speed in this area has a rather low correlation to the demand of the entire province. It is not only the correlation between wind and demand that matters for the integration of wind power. Additionally, each change in the magnitude of the forecast error has a negative impact on the correlation between the planned scheduled generation.
Although more MW of wind power will increase the inertia of wind power, increases of the relative forecast accuracy and consequently also improvements of the correlation to the demand are moderate. The more dispersion of the installed capacity, the lower the cost of integration.
For a given distribution of MW it is possible to compute the resulting error quite accurate hour by hour by using a spatial correlation approach measured in ensemble spread. If the error is known at a few points (e.g. existing facilities), then the error can be computed and extrapolated within the ensemble spread band to any point. WEPROG has developed an algorithm for such correlation tasks based on the ensemble technique ([8]), which will be able to assist in making use of such information for future planning.
Nevertheless, it is unlikely that it is possible to find a distribution of facilities that reduces the forecast error in MW and increases the correlation to the demand significantly. Without introduction of new methodologies to take hand about the error, we must therefore expect significant additional costs by integrating more wind power.
We describe in the following how forecasting can best be used to facilitate wind power integration. Focus will be on how integration can take place on market terms.

5.1 Basic Analysis of Wind Integration Costs

Every significant change in the difference between demand and wind power generation has to be balanced either via reserve or the market. Preferably it should be the cheapest option, while in reality it is the most secure option that will often have to be chosen. The wind integration cost increases not only with the magnitude of each forecast error but also with the variability of the forecast error. Every significant change in error result in a loss. There is no *large volume discount* on the market, if there is reduced competition in supply because of urgency. However, the prices will not suffer from volatility as long as

the market considers that the AESO has enough reserve available, because the operator would be expected to choose the cheapest option to get the required power.
The pre-allocated reserve is therefore a threshold value at which price volatility increases. Once the operator is as many MW out of balance as his reserve, there is reason to expect significantly higher dispatch prices.
The sign of the imbalance also matters. It is not critical for the grid, if there is more wind power than expected. Excess energy just means that the value of the wind power is not as high as it should have been. Lack of energy is more critical. The market knows that the operator cannot use pre-allocated reserve for very long as a scheduled unit could fall out any time.
A dynamic wind reserve allocation according to a predicted uncertainty would prevent volatility on prices. The market may then not be able to estimate how much reserve the operator may have at hand and the competition remains to the benefit of the operator. The cost of permanent reserve is high and would be more expensive than what the 2-hour market would provide. Also, the market could learn how much reserve is allocated over time, even if the numbers would not be published.
This discussion shows the conflict in being the market administrator and balance responsible. How much information should the market be given ? The market participants can use any forecast optimisation to optimise their marginal costs, but they can also use the same information to optimise their bids in moments where they know that there is a party in need of power.

5.2 When Forecasting is Critical

Traditionally wind power forecasting has been seen as required for the market or for the safe operation of the grid. This pattern developed over the years while wind power was regarded as a burden in the energy system more or less everywhere in the world.
Today wind power is accepted, not because it is convenient, but because it is a clean renewable energy source and it increases the competition on the generation price of primary power. There is no longer doubt on that there will be built as much wind power over time as is technically feasible. Therefore, time has come to look upon wind power forecasting as a value added service to wind power rather than just a mandatory requirement.
The design of the Alberta wind power management system should therefore take its focus on how forecasting can be used to optimise the overall operating costs. This is achieved, if all forecast information is used and forecasts are continuously refreshed based on recent measurements. The detailed information about the state should be accessible to the AESO operator in an interface such as WEPROG's PLATON graphical user interface (see Appendix C).
It can be summarised that forecasting is most critical in the following events:

- Ramps of wind power in the summer, where the demand is low. The grid is then relatively more sensitive to deficit or surplus in power generation. The general pattern is that there is more risk of excessive power generation in the summer,

because the periods without wind is twice as long as those with a lot of wind. The base generation from wind is therefore low, but every peak will require that scheduled units ramp down. Forecasting must warn about such periods, so the scheduled units are prepared to ramp in the two hour market.

- Before a sudden ramp down in the winter after a long period of full generation from wind. There will be a tendency towards a deficit of energy, because several scheduled units may have been switch off during the windy period. For this purpose probabilistic forecasts of the demand and wind generation are essential for the AESO to ensure that sufficient additional capacity is available once the wind disappears.

Both events will be critical once there is enough wind power installed. During the pilot project, there were around 70 critical ramps in which probabilistic forecasts of ramping and primary power generation gave useful information about the uncertainty. This is more than one event per week. However, some days have 2-3 ramps. Therefore, in average, it can be expected that there will be one critical event per week, where scheduled generation has to take over for all wind in a period of 2-3 hours or wind takes over a part of the scheduled generation. The power generation from wind can be quite constant in between with 1-3 days of approximately the same output value. This means that there are times where scheduled units could be switched off completely to lower the average generation costs. The consequence is a significant cost each time the units restart, which is undesirable, as reserve is not free of costs.
Additional reserve might therefore be more expensive than using the primary power to balance the wind power as long as there is a enough scheduled capacity to match the ramp rate from wind power. The project has shown that the ramp rates are difficult to forecast, but that the ensemble spread on ramp rate is useful to evaluate, whether additional reserve has to be online during periods with risk of strong ramps. The obvious strategy is to define a probability threshold value depending on the nominated ramping capability of the active primary power assuming that the strongest unit is unavailable and the reserve has kicked in. A future 2GW ramp down from wind power during 3 hours after a long period with strong wind will be the challenge, if this takes place under very cold conditions where there is little excess capacity. There might be no alternative to curtail some wind power in advance to reduce the effective ramp rate and associated costs.

5.3 Probabilistic versus Deterministic Forecasting

The pilot project has demonstrated that an operator cannot rely on one forecast. A security margin would always have to be added. As the cost of these securities are significant, there would be reason to use an objective quantification of the security requirement.
In the deterministic model approach this security has to be predicted from historic information. That means, the accuracy of the past forecasts determines tomorrows security requirement. The disadvantage is that such an approach obviously can not cover extreme events, but rather gives a conservative best guess based on average conditions.

What is required from a security perspective can on the other hand be computed alone from ensemble forecasting data valid in the moment of interest. All ensemble forecasts are therefore valid for the extreme event, which might have a return period higher than the age of wind power.

The actual uncertainty state of the weather situation is used instead of a climatic value in the deterministic case. About 15 years of operational ensemble forecasting in the world's large meteorological services has resulted in this as being the recommended approach for extreme event forecasting, even though this conclusion was achieved with data from modest resolution ensembles. With the data generated in the pilot project we can demonstrate a correlation of at least 0.4 between the predicted and measured forecast error. Although 0.4 seems to be a low correlation, it should be noted that there is a continuous numerically smaller error that anti-correlates. This error is not important and could essentially be reduced by smoothing the forecast, but it does limit the achievable correlation by definition.

Ramp Rate Forecasting

Ramp rate forecasts also benefit from being probabilistic, because the ramps are short lasting. However, it is not enough that they are probabilistic. Each of the ensemble members behind the probabilistic forecast has to be configured to be able to produce the actual ramp. That means, if all ensemble members are tuned with a least square methodology, which smooths the forecasts, especially in extreme (ramping) events, then the ensemble cannot produce or warn of an actual ramp either. Ensemble forecasts should therefore be configured to produce a correct probability density of the requested parameter for any forecast horizon. This is opposite to how deterministic models are trained. They are trained separately on each forecast horizon and the forecast is then by definition not good at producing extreme values, because the single forecast is a best fit of all correct and failed forecasts in the past. The difficulty therefore increases with the prediction horizon.

For an operator looking at a ramp rate forecast this is confusing, because the ramp starts weak on the first forecast, where it is visible and becomes stronger and stronger the closer the forecast approaches the event. The operator may think that the weather forecast amplifies the event, but in fact it is the statistical dampening, which explains part of the change in ramp rate.

To summarise, the ensemble forecast methodology provides more possibilities in the optimisation process as well as in the risk evaluation.

5.4 A Note on Area Forecasting

It has become state-of-the-art to use inherent dispersion in deterministic wind power predictions by only training and predicting on aggregated generation instead of individual sites.

This approach may give better statistical scores and it may also reduce the damping that would otherwise be at longer forecast horizons. Obviously, this methodology cannot give correct forecasts at the sites. As an example, an operator can not check the forecast on the individual site and verify, whether the forecast is in contradiction or accordance with the measurement. A simple subjective extrapolation from the area to the site would not include a consistent direction dependency check. When extreme events take place, then they can mostly be detected at one facility some time prior to the others in the area, and hence prior to the event being visible in the entire area. Thus, it is important for an operator to be able to look at individual forecasts and measurements in extreme events. The operator will need the consistent forecasts for the individual site to know, whether there is agreement and reason to believe in the forecast for all sites when he has to extrapolate into the future and act upon forecasts. Some of the questions an operator may ask are:

- Will the local event stay as a local event or will it spread to areas
- Does the measurement give a signal for the same reason as the forecast or is the forecast ramping due to a signal from another site ?

Considering the learning process of cut-offs in a model, it is rather seldom that all facilities in a region have concurrent cut-offs. The learning process is therefore based on very few events and hence not very accurate. Thus, the likelihood that cutoffs are predicted correct when they finally take place is not high.
It should also be noted that even if the weather forecast is right, then the aggregated result might not be correct, because the learning took place without using the knowledge of where the wind power was produced within the area.
A plain methodology applied on the individual sites is therefore recommended, also because these can be verified by the operator and will give highest confidence in the operation and consistency in the generation. Forecasts should therefore be on site basis also if there only 5km between.

5.5 Wind power's Position in the Market

The way forecasting should be implemented depends on how liberal the market should be. Forecasting methodology and the level of publication should also be chosen in light of how wind power is paid by the consumers. If the consumer pay a high fixed price, then more openness would be expected, but if there is pure liberal conditions, then there would be reason to let wind power act on the same terms and conditions as the scheduled market participants without any price regulation.
Wind power has been operating on market terms in Denmark for a while. However, the payback rate is very low and hence no new capacity was added for a period of 5 years, because the pay back was better elsewhere. One reason why the transition to market terms was unsuccessful was that wind was disallowed to update the generation profile

after gate closure, which is 12 hours before midnight. All other generation types could update without penalties, whereas wind power got a penalty for the forecast from the day before. The rules effectively forbid pooling and secure that the market can speculate in forecast error and thereby increased prices. The rules will be updated in 2008 to allow wind power to operate in pools and with intra-day generation updates. This will most likely increase the market value of wind power again. Wind power on market terms received approximately 5 cents per KWh in year 2007 and an additional production incentive of about 2 cents. The average price per produced KWh is then half of what the scheduled units get, because they produce more while there is no wind and high prices. The Danish solution has effectively been optimised to increase volatility on prices while preserving low average prices. It was not the target for the wind power integration, but the consequence of how it was done. We mention the danish case, because the western Denmark is the area in Europe with the highest concentration of wind power and the only market with a relatively large amount of wind power on market terms. Close to 40% is on market terms in year 2008, while all wind power will be on market terms in year 2012.
The TSO and Energy regulator have now started to suggest that the consumers should use more electricity, while it is cheap and save on electricity when its expensive. This is not very successful, because more than 2/3 of the energy price is a fixed tax independent of the energy price. More flexible consumption would reduce the volatility of the energy price, but there is too little incentive to change habits, if tax is the bulk of the energy price. Especially, if it requires investments to adopt electrical devices to predefined or dynamic start and stop.

5.6 Forecast Publication

One approach could be to publish all forecast information as soon as it is available. This will allow all market participants to optimise their bids according to the wind generation. Obviously the wind generation would need something in return for being exposed to the market. This would normally be a regulated price.
Some countries have chosen this path with success, because it changes the once expensive electricity with the fixed tariff to be the cheapest electricity. Germany had initially a high level of support on renewable energy. With the dramatic market price increases the fixed price are now approximately on the same level. The consumers will in these cases benefit from lower prices after some time. The previously paid higher price can be regarded as an insurance against increasing prices.
The other approach is to let wind power operate on market terms and allow pooling with scheduled units for the purpose of reducing balance costs. This path secures efficiency and leaves the control of the electricity prices to the competition on the market.

This approach does not exclude adjustments of production incentives to ensure that targets for renewable energy are met over time, but it requires that a regulator monitors the real costs of generation, because market participants could essentially bid in under their

marginal cost to make wind power trading inefficient and costly and eventually stop any development. The TSO's in Germany have reported about such attempts, where some market participants have offered their energy on prices far under their marginal costs to ensure that the TSO's would not be able to sell the wind power. This is especially destructive, if the "market clearing price" (MCP) is not determined by a "price-cross" or "pay-at-MCP", but instead as "pay-as-bid" price, as it is e.g. practised on the European Energy Exchange (EEX), the energy market place in Germany.
The EEX has announced to even change rules to allow for negative prices. Only a fraction of the German wind power can be curtailed by the balance responsible parties. This would under the new bidding scheme lead to that it can cost money to get rid of wind power, if there is two little competition, too little demand and too much nuclear power.

Wind power cannot grow sustainable to high penetration levels on market terms with one daily bid based on a single forecast. Low prices and high wind power will always correlate in such a scenario. Wind power and scheduled units should instead work together in contracts for a longer time window. There are different ways of achieving the same goal:

1. Wind power forecasts for the next week are refreshed and published as frequent as possible. All wind power is grouped into one pool, which AESO is balance responsible for. Forecasting is financed by consumer tariffs and is targeted towards making an open market, where all parties (also consumers) have access to all reasonable information about the aggregated wind power production.

2. Wind power is allowed to pool with scheduled units on a monthly basis and pays for forecasting and publishes what is strategically important for the market. The AESO should have real time access to forecast data, but probably restricted to what is of benefit for operating the grid secure in real-time.

What is technically done in both cases might be the same, but the second approach would require some settlement considerations. And, if the second approach would have the investors in wind power in the management board, then the optimisation would clearly focus on achieving a maximum price.

5.7 A possible Future Wind Power Management

Following the discussions above and in order to fit the wind generation well into the Alberta power market we therefore recommend to introduce a "wind generation pool" (WGP) including one or more "balance providing parties" (BPP) with scheduled units.

We further recommend that a consortium is build from the wind stakeholders including the AESO or alternatively with the AESO as sort of regulator. The WGP should ideally be hosted in the AESO for efficiency and grid security reasons. The WGP will also need to use fast and preferably even automatic updates of the BPP's generation.

The purpose of the WGP is to:

- bid into the market with a smooth generation profile of the pool
- prevent that the market can predict WGP's bid on the market from weather forecasts
- maximise the reliability of the WGP power generation
- keep balance costs at a minimum using BPP.

The pool will act in competition will scheduled units. This is only possible if WGP can stay in balance just as well as a scheduled unit and prevent that competitors can predict the bid into the market from weather forecasts.
It is fundamental for the future efficiency of wind power that measurements are not published in a detail that helps the market to predict the WGP's generation pattern. One cannot predict what a coal based unit will produce tomorrow from historical measurements in the same ways as it is possible for wind power. The same applies for hydro power. Therefore, individual wind power measurements should not be published, but instead the pool generation. It is sufficient that AESO knows how much wind power is produced at the individual facilities. The WGP can operate on market terms by using:

1. A weather forecast component with capabilities of producing a large number of different high resolution NWP forecasts as a representative ensemble to estimate the forecast uncertainty.
2. A wind power module computing wind power (P) and ramp rate ($RR = \frac{dP}{dt}$)
3. A demand forecast using available capacity and wind power to compute prices
4. An optimisation module computing power requirements from BPP considering prices and the expected balancing power, which is $(P_{max}) - P_{min}$.
5. A security module that compares the ramping capabilities of BPP against the maximum ramp rate forecast (RR_{max}).
6. A smoothing module checking, if more capacity from BPP should be included
7. A cost and security module testing, if BPP can deliver balance for each ensemble member in case the truth should turn out as that member
8. A reliable measurement network including power and met measurements
9. A 0-4h short-term forecasting tool using all measurements and weather forecasts
10. An optimisation modelling distributing the likely imbalance most cost efficient to the different BPP
11. An efficient scheduling system of BPP to optimise security and costs

The pooling concept for wind power will therefore suppress price volatility and maximise grid security. The WGP will have to deal with the real generation, where a smooth forecast with a good error statistics has no value for the WGP. What counts is to stay in balance.

This does not mean that the wind generator will get the same payment as the scheduled unit. There is a price to pay to BPP for the flexibility.

However, this price will always be lower than the market price for primary reserve, because it is a dynamic reserve bought in advance that enables the WGP to deliver the maximum possible power production from wind. This balancing power is required much more regular than primary reserve. Thus, the BPP will be able to generate nearly as much energy as normally anticipated for a unit outside the pool, and most likely even more than that in the future. The BPP may in the future get a lower average price per produced MWh, but the BPP will also have more production hours than an average scheduled unit. There is a price for generating at maximum and this price is a lower average generation price for the wind generator's equivalent to the balance costs, but the WGP will follow the market price and pay a little less to wind power and somewhat more to the BPP. Alternatively, wind power has to provide some of the reserve for itself.

5.7.1 Selection of the "Balance Providing Party"

The WGP will select balance providing parties (BPP's) in a daily, weekly or monthly tender. A monthly selection of the BPP's is easy to manage in the settlement system, although it might be that lower prices could be achieved in the future with weekly tenders. With today's forecasting technology, it is possible to quantify how many MW of scheduled capacity would be required for a weekly tender. However, the amount of MW is not so critical as long as there is enough capacity. WGP can bid all unused capacity from the BPP into the market according to the price specified by the BPP.

The WGP will comply to the standard gate closures on the market. However, the WGP's bid will contain the pool's output, rather than the aggregated wind power only and the WGP will be allowed to adjust the BPP according to the wind power forecast down to 15 minutes ahead such that the WGP stays in balance. Only the AESO will get to know how much power is due to wind.

The forecast that the WGP delivers to BPP is not a wind power forecast but a forecast of the amount of MW the BPP has to generate. This ensure that no party can learn how much wind power is actually produced by the WGP. And, in that way, the WGP will be able and is recommended to use multiple BPPs to ensure that none of them can derive the actual wind generation. The only requirement would be that the published forecast would only denote the total WGP output and the wind power generation alone is only published as e.g. averages per day.

The WGP concept enables wind power to act as a scheduled unit. It hides the complexity of wind power on the market and will allow the BPP to schedule plants according to wind without being responsible for forecasting or measurement handling.

It can be considered a central system, which implies high grid security. There is an

exchange of information from the WGP to the BPP, which no other market participants will have access to.

5.7.2 Further Development of the "Wind Generation Pool"

The WGP could also develop as a renewable energy pool, as well as benefit from a storage unit in order to stay in balance after WGP's gate closure. This will depend on the price development of the BPP. The most important development for the WGP is however to have a mixture of generation to always have the lowest possible marginal costs.
The WGP has to serve the interests of those who have the financial risk. Therefore, this approach fits into traditional market rules, except for the special internal 15-minute pool gate closure. The main difficulty of the wind power management is in that way removed from the AESO operator, because the WGP will not create significant imbalances.
The WGP will over time even become a *price maker* with increasing amount of MW rather than a *price taker*. Consequently the WGP must forecast demand and also pool prices outside Alberta in order to determine the best possible price for the bid.
The future application of the WGP will be greatly enhanced, if a TSO is closely involved in building up the system. The collaborative work carried out by the work group around the pilot project can therefore be seen as a preparation towards such a "wind generation pool". Much of the knowhow that has been achieved during the pilot project will be hosted and can be exploited within the WGP consortium. The WGP consortium is therefore the natural extension of the initial work.
Such a WGP will combine system security and the economic interest of the investors and hence facilitate the growth of renewable energy in Alberta. Finally, it also creates the opportunity to export a technology, because there are more and more market administrators around the world that struggle with volatility in the balance costs.
The WGP concept is compliant with production incentives, CO_2 certificates and support systems in use around the world. The WGP may receive incentives, but the more incentive, the higher the price to be paid to the BPP, because the BPP wants to have a share of the incentive for the uncertain part of the wind power generation.

5.7.3 Reserve dedicated for Wind Power

There are examples around the world where the TSO invent alternatives to normal reserve for imbalances on the primary power (e.g. the Wind reserve in the German RWE and Vattenfall areas). None of these try to explore nor exploit the value of forecasting. They rather withhold forecast information, because it is felt that the market can abuse the information. The power system is therefore not optimised unless the generators themselves try to acquire forecasts for the wind generation.
Denmark is using a day-ahead market for 100-200MW on up-regulation capacity, while down-regulation is seldom expensive on this market.

5.8 The effect of Measurements

The short term forecast accuracy is the most important for a "wind generation pool" (WGP) approach with an effective gate closure of 15 minutes. It will be the error of the 0-15 minute forecast that will be visible in the settlement. The forecasts determine what the BPP should deliver to match the planned generation for the WGP. A WGP will probably also participate on the 2-hour market. However, we must assume that this market will be driven by errors in the demand forecast.
The forecasts in the 0-4 hour horizon are however always important, because they determine what kind of mixture of power to use from the BPP. The WGP will most likely have a mixture of generation with different costs. The WGP does not know the marginal cost of the BPP, but what has been offered in the monthly tender. Therefore, it is now the WGP's task to make use of this capacity as much as possible by scheduling according to the short-term forecast.
There is a fundamental difference in the 0-4 hour forecast compared to the day-ahead forecast. A WGP will in the day-ahead bid only use:

- The maximum potential production given by the ensemble maximum, which is mostly used to define how much the WGP should produce in each of the settlement intervals.
- The difference between minimum and maximum, which is checked against the BPP's capacity in the pool. This may be used to constraint the WGP's output. Thus, the WGP may not be able to use the ensemble maximum forecast, but the minimum plus the capacity of BPP. It may even be an option to curtail, if the time interval of exceedance will be short.

In the day-ahead market, there will be no utilisation of a so-called best forecast, while the short-term will focus on the best forecast to actually schedule the BPP according to the pool costs.
The AESO operator should be able to monitor the performance of the WGP via graphical access to the relevant output of the WGP. It is not likely that the AESO operator will need to interfere in the process in more than a few events per year. This would be, if the BPP fails to deliver. Otherwise the WGP is expected to stay in near balance. The balancing costs will be a matter between the BPP and the WGP, while production incentives and balancing costs may need to be adjusted by authorities on a yearly basis, if wind power should remain feasible.

5.9 Human Assessment and Automatic Forecasting

It is our opinion, that one of the key points for operational wind power forecasting is that any operational solution needs to be able to run fully automatically. A responsible officer for the operations in Denmark once said "the decision has to be the best of knowledge

also when the operator has pain in his toe". In other words, the decision process has to be objective, no matter in which condition and mood an operator is. A forecasting system should therefore be set up in a way that the operator or forecasters in the TSO can change the automatic solution according to empirical cost/benefit functions from their experience though.

While human assessment and interpretation of weather forecast results is an imperative topic for our societies and human life, it is not suitable in the same way for wind power forecasting, especially not outside the control rooms of the system operators.

Meteorological forecasters prepare in general one or maybe two forecasts per day and try to modify this with modest update for consistency.

The dynamic uncertainty varies sometimes faster than one forecaster can analyse and at other times there are no uncertainties for days. The amount of information that a forecaster would have to process to generate a qualified add-on to the automatic process is at certain times more than possible and at other times there is nothing to do than confirm what the automatic solution provides. Therefore, it is not likely that a forecaster can add much value over longer terms.

What human intervention can do is to couple information from automatic systems that do not communicate as they should. This could be demand forecast issues, limited generation capacity and their impact on the wind power forecast optimisation target. This requires personal that understands the power system more than meteorological aspects.

A model system can well be optimised to access exactly the information that a forecaster would take into consideration. If there would be an added value from a forecaster, then the procedures should be identified frequently and programmed so it can also be performed by an automatic approach. This is the safer and cheaper way for a system that should be up and performing 7x24hours.

Where Human Assessment is useful

The recommended implementation of human assessment on the meteorological site would be to use offline mode assessment, where a forecaster writes daily reports of how the objective forecasts could have been enhanced. One day per fortnight or month could be allocated for this or as long time as the forecaster desires. Most important is that there is no time pressure, that all information is available and the forecaster can also verify, if the correction went in the right direction for the correct reason.

All forecast data and measurements should therefore be made available to the forecaster. After a while, the forecasters assessment is coded as a module and tried out for a reference period. If the module can pass this period with improvements on a predefined cost of security measure, then it should be accepted for operation. Following this strategy there will be no direct meteorological forecaster decision required on the operation. The effect will be indirect and objective once the theory has been validated. This procedure secures the forecaster and the grid.

Considering the complexity of the forecasting process and that very few people understand the full production chain, it should be noted that subjective information added by a

party who has responsibility and access to only a fraction of the production chain should be chosen with care. This applies also to wind power forecasters, because they are still dependent on information that they often cannot access. For some forecasters it is limited to the state estimate process of third parties and for others it is even including the weather forecast stage. There is however only one possibility to assess uncertainty and that is to hold the result up against other independent forecasts. This is exactly what an ensemble forecasting system is designed for and is doing by itself all the time with an objective method. The objective uncertainty given with the forecast should include all uncertainty and the single forecast provided should be what is most clever to believe in within the uncertainty interval. If it is not so, then the programming should be updated so the uncertainty and best value is consistent.
The value of human assessment for wind power forecasting therefore has to be evaluated carefully and the risk and consequences for wrong advice from a commercial party should also be considered, when adding such services to the daily operations. This is especially dangerous, if advice may be seriously wrong and the consultant would have been expected to know better. This can well happen in a critical event. Therefore, instruction within the company of the consultant would normally always be to give out advice very carefully and cautiously.
The potential value of human assessment is however most significant in extreme cases, where there is a risk of a concurrent cutoff. Such situations are exactly, where the advice is most critical and where it is most important that the TSO takes full responsibility for the grid security.

5.10 Forecasting at the different Forecast Horizons

The purpose of the "wind generation pool" (WGP) is to increase the value of wind power in the market. The WGP will therefore have to forecast all horizons that are technically feasible, exchange information with a "balance providing party" (BPP) and deploy the output where ever possible.
The following list the different forecast horizons and the recommended usage of forecasting.

- The *s*hort-term forecasts up to 4 hours in which measurements and ensemble forecasts are combined. The forecasts should be in 5 or 10 minute resolution. These should be targeted on accuracy of wind power alone and they should be used to compute how much a BPP should deliver. These forecasts should be created approximately every 10 or 15 minutes.

- The *s*hort-term and medium-term forecasts starting at 4 hours to approximately a week, where Ensemble forecasts for particular events are used. They must be in hourly resolution and be targeted to forecast the minimum and maximum expected generation. The generation profile of the WGP will be determined from those alone. Also a price forecast should be made taking the demand into consideration. The

outcome will be one very important forecast per day to the market, whereas the BPP should receive updated forecasts with regular intervals such as every hour or every 3 hours.

- The *l*ong-term forecasts cover the horizons longer than 7 days and will be based on statistical processing of historic ensemble forecasts valid in the same season. The WGP will in its monthly tender need some estimate of how many GWh's a BPP will be expected to provide and also the peak capacity and frequency distribution of the primary power within the number of hours with different levels of ramping. This information will be important for computing the offer for the BPP to the WGP. The forecast process will run once per month and generate statistical forecasts for every week and month of the year for the current facilities.

The 3 forecasting methodologies should enable the WGP to manage wind power efficient on the market, give the AESO operator maximum overview and enable the market participants to give cost efficient bids in monthly tendering processes.

5.11 Why become a "Balance Providing Party" ?

The economist will say that anything can be bought for money, if the price is right. It may not guarantee an optimal contract for the WGP. However, for the following discussion we assume this situation as being applicable.
The WGP will have the most detailed knowledge of the market and the best forecast, such that the "Balance Providing Party" (BPP) can consequently benefit from more optimal scheduling.
The WGP will over time become *price maker* with increasing capacity and at times exclusive generator. To be part of WGP's generation will secure many generation hours and save labour costs, because the scheduling is defined by the WGP. The net result is that it will be more attractive to be BPP than competing on the remaining generation.
Wind generators inside the WGP must expect to get less payment per energy unit than scheduled units, but on the other hand also generate as much as possible. Part of the payment goes to the BPP. There is no way around price increases on scheduled energy. It is a required service, but there be less and less required, but usage still have to cover the marginal cost of hours with no generation.
The marginal costs of the WGP will over time and with increasing energy prices become more and more competitive, because wind power is likely to develop to be the cheapest power generation form.
A last question may be what would happen if the WGP does not get competitive offers ? Then, the WGP can use its forecast capabilities to forecast traditionally with minimum error for a base load and use the ensemble for the more uncertain production. The WGP will be able to bid into the market with a range of different prices. The certain energy (Ensemble minimum) will get a low price, and the price will thereafter increase in intervals up to the Ensemble maximum. The 2-hour market can then be used to get in balance.

Next Steps in the Integration of Wind Power Forecasting

The project has shown that improvements are required, but Alberta's wind conditions are not typical, which means that it would be difficult to go to a public service organisation outside Alberta and ask for optimal services unless there would be a cost recovery.
There would on the other hand be scientific interest in Alberta's wind conditions, because the conditions are extreme and difficult.
It would be possible to isolate a couple of events to conduct research. Obvious candidates for such work would be the cold events were nearly all forecasts ramped up, but where it did not happen. There were two periods where this error was dominant. The opposite error was rather rare. The unforecasted power seem to rather be related to events where the wind turns in a northerly directions from a westerly direction. A handful of these events also exist with the 6th of September 2007 as the most critical.
Single events may give a misleading impression. The better way is to select 5-10 periods of 7 days in which the above mentioned errors were frequent. This would be sufficiently long to be robust and sufficiently short to allow for many simulations. The aim of these would be to enhance the vertical momentum transport for extreme stable and unstable conditions.
The forecasts that have been generated in the project can be used for improving the short-term forecasts, but also to develop better day-ahead probabilistic ramp forecasts.
Probability distributions of long-term forecasts could be generated by using day 2 and 3 of all ensemble forecasts. They are not bound tightly to the actual initial conditions. They can therefore be used to rather give seasonal forecast.
In addition there will be work required on improving the roughness and height of the mountains in the high resolution 6km grid. There will also be reason to work on how to analyse the surface better. As mentioned in the Section 4, the focus should be on snow cover characteristics in the winter and vegetation in summer and perhaps also consider irrigation.
Investments in additional measurements should probably not have the highest priority until there is more profound knowledge on the impact of the amount of measurements that were available to the project. It should first be investigated if the dry air downdrafts are resolvable in the required detail. If this fails then the next step could be to consider one or two LIDAR.
The high resolution 6km ensemble has shown how the atmospheric dynamics work. The down draft in the models correspond with a very good approximation to the observed

structures. The 6km ensemble can therefore be used to estimate the future wind power generation with past weather data. The 6km ensemble can in addition show some valuable information regarding the wind resource.
To conclude, it is worth the while to keep the two years of data as a reference to estimate the value of future measurements and forecasts. Improvements to the forecasting system should therefore always be validated on this period until a better and more representative period is available in a future operational environment.

6.1 Positioning of future wind facilities

The results of the 6km ensemble suggests that the forecast error and uncertainty can be reduced, if future facilities take predictability into account. This means that the positioning of facilities should take the characteristics of the regions and the flow over the mountains into account.
The South-West (SW) region is more directly dependent on the mountains than South-East (SE) and South-Centre (SC), because the facilities in SW rarely ramp fast, when a cold front moves southward. The facilities in the SE and SC area ramp in that case nearly concurrent. On the other hand, these areas will be easier to predict for westerly flow than SW, because SW is mostly the first area to ramp. If predictability is taken into account in the positioning of future wind facilities, the following rules are recommended to follow:

- SW facilities should be built rectangular with the longest dimension in direction west-east. This will suppress concurrent cutoffs for two reasons. The first is that cutoff take place in narrow south-north bands and the second is that turbines on the lee side of the mountains are rarely cutting off, but if they do, then it is at least not concurrent with the westerly turbine rows due the wake effects.
- SC and SE facilities should be built rectangular with the longest dimension in the south-north direction
- The installed capacity per area unit should be highest in the centre region and reduce gradually in all directions as smooth as possible. Forecasting will not capture details, but this matters less, if the installed capacity varies smooth per area unit.
- The facilities should preferably provide SCADA data on all corners of the facilities. There is no requirement to report a specific selection of variables.
- Larger distances between turbines give lower forecast error.

These guidelines will help to reduce ramp rates without compromising on the generation and reduce the sensitivity to forecast errors, because forecast errors within the facility area will partially cancel each other out. They are not going to bring a “jump” in the forecasting quality, but they will reduce the balancing costs and thereby also increase the market value of wind power.

Project Summary and Conclusions

The Alberta Wind Power Forecasting Pilot project confirmed what was a discussion topic between the AESO and the various stakeholders. The wind was as variable as discussed and even more difficult to predict. The project was fortunately a pilot project, but forecasters around the world would probably prior to the forecast have claimed that the evolution on forecasting would be one or two steps further than the pilot stage.

The project developed to a challenge on the meteorological accuracy. As more and more difficult events arose in the forecasting period, it seemed like the only possibility to carry out a large test of numerous model configurations in order to identify, if better weather model configurations exists.

A total of 600 independent weather forecasts were run in a 6-hourly cycle for a period of 3 months. They were verified on the existing facilities using the same conversion from wind to power methodology. The top 10 ranking was highly dominated by members from the highest resolution and one member was even better than any combination of the others. Nevertheless, the error of that member is significantly higher than in other areas of the world. The configuration that seemed to be best will not be accepted as neither state-of-the-art in NWP modelling nor the way forward, because it has amplified damping characteristics. We can characterise this model configuration as specifically tuned towards very difficult, windy and variable conditions, but not for general purpose weather forecasting.

Based on that experience we can conclude that it is not likely that a model configuration can be found, which will give reasonable low balancing costs, if traditional forecasting is used to bid the wind generation profile into the market. The forecast error is about twice the error of other similar markets measured in terms of installed capacity and rather 3 times the error on markets with dispersed wind power. On this basis we must foresee very high balancing costs also because of modest interconnection.

We therefore recommend a forecasting methodology that bypasses the day-ahead forecast error with our "wind generation pool" scheme. The scheduled units will be asked to deliver a certain amount of energy over a period to a pool. They will be advised on the expected variance and other statistical measures and will receive frequent short-term forecasts for what they should deliver 0-6 hours ahead.

This forecasting methodology will result in very few disturbances from wind on the two hour market and the reserve compared to what a further increase in the wind generation capacity would normally cause. A pool also exploits that there is volume discount in a fair and timely market whereas the two hour market cannot give the same sustained volume discount, because of the high inertia level of most generating capacity.

What this means for the wind power forecasting is that the target is to generate many correctly structured forecasts from which realistic probabilistic ramp rates can be derived. This information is required on the market in order to calculate the cost of the balancing power.

The proposed methodology gives time to gradually improve the weather forecasts to handle the specific conditions in Alberta. As improvements are demonstrated they can be used to reduce the daily balancing requirements. In the unlikely event that the sustained perfect forecast should ever be developed, then the pool will reduce to a wind only pool, which could bid in one forecast in the day-ahead market.

The solution can be managed in a pool of mixed scheduled and non-scheduled generation on market terms, while it would also be possible to let the AESO be responsible alone. The choice depends on how the market and wind power stakeholders would consider the AESO both as a generator and market administrator. The solution should in both cases be hosted centralized for performance, cost, technical and security reasons.

It will be exiting to follow and maybe also participate in the integration of wind power in Alberta in the next years. We have gained significant understanding thanks to our many experimental ensemble setup and especially the high resolution 6km ensemble and we believe this is an important step in explaining the dynamics of the Chinook. We got a proof of our theory and that was important for being able to write this report and provide well founded recommendations.

Bibliography

[1] Dupont, N. *Isslaget, der kostede en hel måneds afregning*, Naturlig Energi, 7, pp. 16-17, 2006.

[2] Ernst, B., Oakleaf,B., Ahlstrom, M.L., Lange,M., Moehrlen, C., Lange, B., Focken, U., Rohrig, K., *Predicting the Wind , Models and Methods of Wind Forecasting for Utility Operations Planning*, IEEE Power & Energy, Vol.5 No.6, November/December 2007.

[3] Giebel,G. (ed.), Jake Badger, Lars Landberg, Henrik Aalborg Nielsen, Torben Skov Nielsen, Henrik Madsen, Kai Sattler, Henrik Feddersen, Henrik Vedel, John Tøfting, Lars Kruse, Lars Voulund, *Wind Power Prediction using Ensembles.* Risø-R-1527, September 2005.

[4] Goeber,M., Wilson, C., *W*hy, when a model resolution is improved, do the forecasts often verify worse?, WWRP/WGNE Joint Working Group on Verification, See also http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/ Goeber_Wilson/Double_penalty_combined.html,

[5] Jørgensen, J., *High Resolution Numerical Wind Energy Model For On and Offshore Forecasting Using Ensemble Predictions*, Honeymoon Final Report, EU ENERGIE Contract ENK5-CT-2002-00606, European Commission, 2005a.

[6] Jørgensen, J.U., Moehrlen, C., McKeogh, E.J., *A* New Generation Operational On- and Offshore Numerical Prediction System, Proc. World Wind Energy Conference "Clean Power for the World", Berlin, Germany, (2002).

[7] McCrank, D., *Wind Power Forecasting Pilot Project Progress Report*, http://www.aeso.ca/downloads/Wind_Forecasting_Progress_Review_27_Feb_08.pdf, Feb. (2008)

[8] Möhrlen, C., Jørgensen, J.U., Pinson, P., Madsen, H., Runge Kristoffersen, J., *H*RENSEMBLEHR - HIGH RESOLUTION ENSEMBLE FOR HORNS REV, Proc. European Offshore Wind Energy Conference, Berlin, (2007).

[9] C. Möhrlen and J. U. Jørgensen, "Forecasting wind power in high wind penetration markets using multi-scheme ensemble prediction methods," in *Proc. German Wind Energy Conference DEWEK*, Bremen, Nov. (2006).

[10] Moehrlen, C. *Uncertainty in Wind Energy Foercasting*, Thesis (PhD) Deptartment of Civil and Environmental Engineering, University College Cork, Ireland, DP2004MOHR, May (2004).

[11] Moehrlen, C., Jørgensen J.U., McKeogh, E.J., *P*ower predictions in Complex Terrain with an operational Numerical Weather Prediction Model in Ireland including Ensemble Prediction, Proc. World Wind Energy Conference "Clean Power for the World", Berlin, Germany, (2002).

[12] Möhrlen, C. and J. U. Jørgensen, *Verification of Ensemble Prediction Systems for a new market: Wind Energy*, ECMWF Special Project Interim Report 3 (2005). [1]

[13] Murphy, A.H. and Epstein, E.S., *S*kill scores and correlation coefficients in model verification. Mon. Wea. Rev., 117, 572-581, (1989).

[14] M. Pahlow, L.-E. Langhans, C. Möhrlen, and J. U. Jørgensen, "On the potential of coupling renewables into energy pools," *Z. Energiewirtschaft*, vol. 31, pp. 35-46, (2007).

[15] Pahlow, M., Möhrlen, C., Jørgensen, J.U. *Application of cost functions for large-scale integration of wind power using a multi-scheme ensemble prediction technique*, accepted in *Optimization Advances in Electric Power Systems*, Ed. Edgardo D. Castronuovo, NOVA Publisher NY, (2008).

[16] Stephenson, D.B. and F. J. Doblas-Reyes, *S*tatistical methods for interpreting Monte Carlo ensemble forecasts. Tellus, 52A, 300-322,(2000).

[17] Lang, S., C. Möhrlen, J.Jørgensen, Ó Gallachóir B. P., E. McKeogh, E., *Application of a Multi-Scheme Ensemble Prediction System for wind power forecasting in Ireland and comparison with validation results from Denmark and Germany*, Scientific Proceedings, European Wind Energy Conference, March, Athens, Greece, 2006.

[18] S. Lang, C. Möhrlen, J.Jørgensen, Ó Gallachóir B. P., E. McKeogh, *Aggregate forecasting of wind generation on the Irish grid using a Multi-Scheme Ensemble Prediction System*, Proc. Conference on Renewable Energy in Maritime Climates, April, Dublin, 2006.

[19] Taylor, K.E., *S*ummarizing multiple aspects of model performance in a single diagram. J. Geophys. Res., D7, 106, 7183-7192, (2001).

[1] ECMWF Reports avaialble at: http://www.ecmwf.int/about/special_projects/finished_projects or http://www.honeymoon-windpower.net

Glossary

Abbreviations

Abbreviation	Explanation
4DVAR	Four-Dimensional Variational Data Assimilation
BPP	Balance Providing Party
BRP	Balance Responsible Party
CMC	Canadian Meteorological Center
ECMWF	European Center for Medium range Weather Forecasting
EPS	Ensemble Prediction System
EEX	European Energy Exchange, the European energy market place in Leipzig, Germany
GTS	Global Telecommunication System, a global onetwork for atmspheric data
MAE	Mean absolut error
MCP	Market Clearing Price
MSLP	Mean sea level pressure
MST	Mountain Standard Time
NCEP	National Center for Environmental Prediction
NOAA	National Oceanic And Atmospheric Administration
NWP	Numerical Weather Prediciton
PDF	Probability density function
PMT	Probabilistic multi-trend filter
POD	Probability of Determination
POFD	Probability of false determination
OR	Odds ratio
RMS	Root mean square error
SS	Skill scores - verification method used for ensemble predictions
TSS	True skill scores, also “Hanssen and Kuipers discriminant”
STDV	Strandard Deviation
TSO	Transmission System Operator
UTC	Universal Time Coordinated
VAR	Variance
WGP	Wind Generation Pool

Glossary of Meteorological Terms

Most of the following terms are from the electronic version of the American Meteorological Society's Meteorological Glossary, which is a copy of the second edition of the Glossary (http://amsglossary.allenpress.com/glossary).

Analysis: is the production of an accurate image of the true state of the atmosphere at a given time, represented by a collection of numbers, usually on regular model grids. Objective analysis is an automated procedure for performing such analysis versus subjective, hand analysis. See also Data Assimilation.

Cyclonic circulation: Atmospheric circulation associated with a cyclone (depression, low pressure area). It is counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere.

Data Assimilation: Data assimilation is an analysis technique in which the observed information is accumulated into the model state by taking advantage of consistency constraints with laws of time evolution and physical properties. It is the process of combining observations and short- range forecasts to obtain an initial condition for NWP. The purpose of data assimilation is to determine as accurately as possible the state of the atmospheric flow by using all the available information.

Ensemble forecast: A set of different forecasts all valid at the same forecast time(s). The differences between the forecasts can provide information on the probability distribution of the predicted variables. The forecasts in the ensemble may have different initial conditions, boundary conditions, parameter settings, or may even be from entirely independent NWP models.

Envelope orography: A method used in numerical models for weather forecasting in which it is assumed that mountain passes and valleys are filled mostly with stagnant air, thus increasing the average height of the model mountains and enhancing the blocking effect.

First Guess: The use of short-range forecasts as a first guess has been universally adopted in operational systems into what is called an "analysis cycle". Initially climatology, or a combination of climatology and a short forecast were used as a first guess. The first guess or background field is our best estimate of the state of the atmosphere prior to the use of the observations.

Forecast: is a scientific predictions about future states of the atmpsphere made with a numerical model or method. A forecast incorporates meteorological, oceanographic, and/or river flow rate forecasts; makes predictions for locations where observational data will not be available; and is usually initialized by the results of a nowcast. see also *Numerical Forecasting*.

Front: In meteorology, generally, the interface or transition zone between two air masses of different density.

Frontal System: The orientation and nature of the fronts within the circulation of a frontal cyclone (cyclonic circulation).

Gravity wave drag: A generally zonal acceleration produced by upward propagating gravity waves at levels where the waves break. Gravity wave drag plays an important role in explaining the zonal mean flow and thermal structure at higher atmospheric levels, particularly in the mesosphere.

Gravity wave drag parameterization: A parameterization designed to approximate the effect on the resolved flow of the gravity wave drag that would be generated by unresolved subgrid-scale topography in an atmospheric model.

Initial Conditions: Initial conditions in a global model are prepared by making a synthesis of observed values of atmospheric fields taken over a for example 24 hour period and short-range forecasts provided by the global model itself. This synthesis is a process of assimilating observed values into a model. The use of both observations and model forecasts in the construction of initial values is required. High quality data are sparsely and irregularly distributed over the globe. Short-range model forecasts carry knowledge forward in time of earlier observations and also provide a crucial background for extracting useful information from expensive satellite observations.

Isobars: A line of equal or constant pressure; an isopleth of pressure. In meteorology, it most often refers to a line drawn through all points of equal atmospheric pressure along a given reference surface, such as a constant-height surface (notably mean sea level on surface charts), an isentropic surface, the vertical plane of a synoptic cross section, etc. The pattern of isobars has always been a main feature of surface-chart analysis. Isobars are usually drawn at intervals of one millibar or more, depending on the scale needed to identify or illustrate a specific meteorological pattern.

Lateral Boundary Condition: A set of mathematical conditions to be satisfied, in the solution of a differential equation, at the edges or physical boundaries (including fluid boundaries) of the region in which the solution is sought. The nature of these conditions is usually determined by the physical nature of the problem, and is a necessary part of the problem's complete formulation. Common boundary conditions for the atmosphere are that the velocity component normal to the earth's surface vanish, and that the individual derivative of pressure vanish at the upper surface.

Numerical Weather Prediction: NWP is an initial- boundary value problem: given an estimate of the present state of the atmosphere (initial conditions), and appropriate surface and lateral boundary conditions, the model simulates (forecasts) the atmospheric evolution.

Numerical Integration: A solution of the governing equations of hydrodynamics by numerical methods. The numerical solutions are carried out with the aid of computers ranging from desktop workstations to the most powerful computers available.

Numerical Forecasting: (Also called mathematical forecasting, dynamical forecasting, physical forecasting, numerical weather prediction.) The integration of the governing equations of hydrodynamics by numerical methods subject to specified initial conditions. Numerical approximations are fundamental to almost all dynamical weather prediction schemes since the complexity and nonlinearity of the hydrodynamic equations do not allow exact solutions of the continuous equations.

Mesosphere: The region of the atmosphere lying above the stratosphere and extending from the stratopause at about 50 km height to the mesopause at 85-95 km. The mesosphere is characterized by decreasing temperature with increasing height, reflecting the decreasing absorption of solar ultraviolet radiation by ozone.

Mesoscale: Pertaining to atmospheric phenomena having horizontal scales ranging from a few to several hundred kilometers. From a dynamical perspective, this term pertains to processes encompassing deep moist convection and the full spectrum of inertio-gravity waves but stopping short of synoptic-scale phenomena, which have Rossby numbers less than 1.

Parameterisation: The representation of physical effects in a dynamic model in terms of simplified parameters, which represent a series of simplifications of the full turbulence model to remove complex terms and form a closed set of equations that lead to a hierarchy of so-called closure models of decreasing complexity.

State Estimate: Is the outcome of combining observations and short- range forecasts to obtain an initial condition for NWP models. It is the most accurate possible state of the atmospheric flow from available observational and NWP information at a given time. See also Analysis and Data Assimilation.

Stratosphere: The stratosphere starts just above the troposphere and extends to 50 kilometers (31 miles) high. Compared to the troposphere, this part of the atmosphere is dry and less dense. The stratopause separates the stratosphere from the next layer.

Synoptic Scale: Used with respect to weather systems ranging in size from several hundred kilometers to several thousand kilometers, the scale of migratory high and low pressure systems (frontal cyclones) of the lower troposphere.

SYNOP (surface synoptic observations): This is a numerical code (called FM-12 by WMO) used for reporting weather observations made by manned and automated weather stations. SYNOP reports are typically sent every six hours on shortwave using RTTY. A report consists of groups of numbers (and slashes where data is

not available) describing general weather information, such as the temperature, barometric pressure and visibility at a weather station.

Thermosphere: The thermosphere starts just above the mesosphere and extends to 600 kilometers (372 miles) height. The temperatures go up with increasing altitude due to the Sun's energy. This layer is known as the upper atmosphere.

Troposphere: The troposphere starts at the Earth's surface and extends 8 to 14.5 kilometers high (5 to 9 miles). This part of the atmosphere is the most dense. Almost all weather is in this region. The tropopause separates the troposphere from the next layer. The tropopause and the troposphere are known as the lower atmosphere.

Glossary of Energy Terms

Ancillary Services: ANCILLARY SERVICES are Interconnected Operations Services identified as necessary to effect a transfer of electricity between purchasing and selling entities (TRANSMISSION) and which a provider of TRANSMISSION services must include in an open access transmission tariff.

Market Clearing Price: The Market Clearing Price (MCP) describes the equilibrium price determined in the hourly auction of the electricity spot market, which is calculated based on the bids entered into the trading system.

Equilibrium-price: Defines the price at intersection of the buy and sale curves within the price range defined by the energy market.

Price-Cross scheme or Pay-at-MCP scheme: The System price is for each hour determined by the intersection of the aggregated supply and demand curves representing all bids and offers. In other words, selected supply bidders are paid and selected demand bidders are charged at the MCP price regardless of their bid price.

Pay-as-bid scheme: The "pay-as-bid" price scheme, the payments for selected bids are determined as their bid costs. In other words, selected supply bidders are paid the price of their bid and the selected demand bidders are charged at their bid.

Appendix A

This appendix contains error statistics in table format for the real time project period for ensemble A, B and D. The two first are in 45km resolution and D is in 22km equivalent to the delivered forecast. All results are generated with the same power prediction version. What differs between the 3 systems is only the weather forecasting setup. All 75 ensemble members within A, B and D runs in the same model grid and resolution. A fixed member in A, B and D differ only in their model area, orthography and resolution.

There are 225 weather forecasts generated every 6 hours, but there has been no attempt made to use them as a 225 member ensemble. The reason is the quality difference which will be obvious later in this appendix.

Each ensemble is running every 6 hours and use only large scale global state estimate/forecasts as input. The hourly forecasting procedure, where on-line measurements is used, is not included, because the target is to measure the performance of the weather forecasts.

All numbers are given in % of installed capacity. Each number is averaged over a 6 hour interval. A number at 9 hours is the average of the 6-11 hour forecasts. The forecast length (fcl) given here is relative to the initial time. In a real time environment the effective forecast horizon would be considered 5 hour shorter than the one presented here. If we were to measure the error in a running look ahead time manor we would add another 3 hours (half of the ensemble run frequency). Thus, the 9 hour horizon corresponds to the average forecast horizon which is the basis for the ultra short term forecast, where the on-line measurements go in.

The statistical training procedure for the wind to power conversion used in the results in this section allows for 8 direction bins and 15 wind speed bins. All 3 systems were trained with 2 years data.

Training and forecasting took place on each of the existing facilities. The forecast is the aggregated of each individual facility. Transmission constraints have been ignored.

Validation was done on hourly mean values centered at full hours.

The optimization target for the forecast used in the statistical tests, where one forecast is required, is minimum Mean Absolute Error achieved by combining ensemble members.

There is no attempt in these results to use seasonal training nor time of the day. The sensitivity of these effects are expected to be handled in the weather forecasts.

A.1 Verification of RMS error for different Regions

fcl	A	B	D	SA	IR
3	17.32	17.63	15.69	10.87	6.75
9	19.18	18.87	15.88	10.87	6.75
15	21.40	21.28	17.19	11.06	7.21
21	21.78	22.38	17.94	11.43	7.90
27	22.51	23.38	18.63	12.21	8.86
33	22.58	23.65	18.92	12.66	9.77
39	22.93	23.96	19.25	13.11	10.83
45	23.46	24.39	19.57	13.51	11.65

Table A.1: RMS error for Existing Facilities. Alberta data includes the project period.

The first comparison shows how much more difficult wind power prediction is in Alberta than in other areas. South Australia (SA) and Ireland (IRL) are chosen for comparison, because they have larger wind farms like Alberta, with a similar average production as in Alberta. The periods are not concurrent. A particular windy autumn-winter period was chosen in Ireland end of 2006 to show the error growth under extreme conditions. The South Australia number is an average over almost 2 years. The distance between the wind farms in SA is larger than in Alberta. Irelands area is also larger than Alberta, if we consider the location of wind power.
System B is the same methodology as the two shown for SA and IRL, but the error growth and error level is approximately a factor of two higher in Alberta than in IRL and SA.
If we compare to system D, the result looks a little better for Alberta, because the error growth does no longer exceed IRL.
From that one could conclude that the low forecast quality in Alberta is therefore not connected to that the weather forecast reduce in quality, but rather that the forecast is never really good even in the shorter forecast horizons.
Either the error is related to very high variability of the wind or generally poor weather forecast performance in the Alberta error. It is difficult to separate these two with an objective method, because some of the variability is not timely, but rather spatial and a consequence of abnormal strong horizontal gradients. The industry would probably not classify this as variable conditions, but the meteorological error level is the same for strong gradients and high frequent variability.

A.2 Verification of Normalized Mean Absolute Error

fcl	A	B	D
3	12.36	12.55	11.05
9	13.56	13.07	11.04
15	15.42	14.89	12.05
21	15.82	15.68	12.62
27	16.39	16.52	13.14
33	16.47	16.88	13.46
39	16.81	17.18	13.69
45	17.33	17.54	13.98

Table A.2: Normalized Mean Absolute Error for Existing Facilities in the project period

The absolute error follows the RMS error pattern. System A seems however to be relatively better than system B on RMS and MAE. The taller envelope orography of system A therefore reduces the numerically largest errors most. It is obviously better to be able to produce a stronger downdraft, although this is equivalent to more variability and the potential for higher error. That is an indication of that even the 45km resolution setup somehow produces and block downdrafts at about the correct time.

A.3 Verification of Bias

fcl	A	B	D
3	1.19	1.55	1.24
9	-0.06	0.93	0.75
15	-1.18	-0.03	0.32
21	-0.94	-0.65	-0.27
27	-1.42	-1.46	-0.53
33	-1.00	-1.18	-0.42
39	-0.62	-1.28	-0.55
45	-0.89	-1.63	-0.66

Table A.3: Bias for Existing Facilities in the project period

The bias drift is systematically less for system D. The steepness of the mountains in A and D are almost similar. It can therefore be concluded that the truncation error from the poor resolution in A has negative effect on the model bias drift. One can say that if resolution helps on the model bias, then there is something fundamentally wrong in using moderate resolutions.

A.4 Verification of correlation

fcl	A	B	D
3	0.85	0.85	0.88
9	0.81	0.83	0.88
15	0.76	0.77	0.85
21	0.75	0.74	0.84
27	0.73	0.72	0.82
33	0.73	0.71	0.82
39	0.72	0.69	0.81
45	0.70	0.68	0.80

Table A.4: Correlation between forecast and measurement for Existing Facilities in the project period

The correlation gives the cleanest signal on the relative forecast quality between the systems. The correlation of system D of 0.8 should be seen in light of 0.9 in Ireland at a horizon of 45 hours and for dispersed wind power in Denmark of almost 0.95, but still D is clearly better than A and B.

A.5 Verification of correlation between error and ensemble spread

fcl	A	B	D
3	0.34	0.27	0.36
9	0.28	0.28	0.36
15	0.29	0.25	0.34
21	0.29	0.23	0.34
27	0.27	0.21	0.33
33	0.29	0.21	0.34
39	0.30	0.21	0.34
45	0.30	0.20	0.34

Table A.5: Correlation between ensemble spread and forecast error for Existing Facilities in the project period

The ability to predict errors can be measured with the correlation between forecast spread of each ensemble and the actual error of the forecast verified above. The higher resolution in D is again best. The correlation hardly reduces with forecast horizon. This is important

for allocation of reserve. The correlation is still low, but there is considerable daily high frequent moderate amplitude variability, which drags down this correlation and a background error, which is not dependent on the mean wind speed, but a variety of meteorological and non-meteorological factors. Moderate variability is not as important as the larger variability over hours. The result also suggests ithat *reserve can just as well be bought day-ahead* as on a shorter notice in more frequent intervals, because the accuracy is independent of the forecast length. This shall be understood as follows. If forecasts have the correct ramping structure, then the difference between the members represent the actual uncertainty, because most error is due to phase errors. This will be shown in the next appendix.

A.6 Verification of Ramp rate RMS

fcl	A	B	D
3	6.08	6.00	6.12
9	6.11	6.03	6.06
15	6.14	6.05	6.16
21	6.20	6.10	6.18
27	6.15	6.09	6.14
33	6.29	6.25	6.17
39	6.17	6.15	6.18
45	6.14	6.21	6.22

Table A.6: Ramp rate RMS for Existing Facilities in the project period

It is disappointing to see that ramp rate errors can be predicted equally good or bad with the cheapest and simplest approach as D. Additionally, experiments showed that the plain average was better than trying to use the best members. Obviously, the ramp rate is a on/off field that is very difficult to hit correct. The dispersion term is responsible for the bulk of the RMS term. It is not shown, but it varies between 5.5% to 6.1%. The major fraction of the RMS error of ramps is therefore phase error related. The reason why otherwise better models have a similar score than "bad" models is probably that ramps are short lasting and all models hit them randomly. All forecasts are tuned to the same structure, but if the error is 2 or 4 hours wrong it may not have much impact on the magnitude of the error. In fact it is possible to argue that if forecasts only have phase errors and the frequency distribution is correct, the MAE of the ramp rate forecast is invariant to the forecast length, but the RMS error reduces, if ramps are softer. The A and B models have a bias drift, so this is not exactly the case. Therefore, we can state it is disappointing to see that D is not better than A and B on this error parameter.

A.7 Verification of Member Ranking

order	1	2	3	4	5	6	7	8	9	10
3	D4101	D4121	D4001	D4021	D4102	D4122	D3121	D4004	D3101	D3122
9	D4004	D3220	D1122	D1220	D3122	D3200	D1120	D4220	D4200	D3120
15	D4004	D1122	D1020	D1120	D3122	D3120	D1220	D3220	D3020	D2001
21	D4004	D1020	D2001	D2021	D3020	D1202	D3021	D1021	D1122	D3202
27	D4004	D1202	D2001	D3202	D2021	D1020	D3021	D3020	D1021	D3001
33	D4004	D1202	D2001	D3202	D1000	D2021	D1001	D3001	D3300	D1102
39	D4004	D1202	D3001	D3300	D3202	D1001	D1000	D3000	D2001	D1011
45	D4004	D1202	D3001	D1001	D1000	D3202	D3300	D1011	D3000	D1010

Table A.7: Member Ranking for Existing Facilities in the project period

The member ranking gives detailed information of which members are best on various horizons. The winner is very clearly the model named **D2004**, which is the best combination of all members in ensemble D. The member named **D1202** seems to be a quite good configuration as it is best individual member for all the longer horizons, but is not visible in top 10 on the shorter horizons. We are quite sure that some capability of this member causes that it is smooting more in this resolution. It is normally not a model configuration that is seen in top 10 elsewhere. It can only compete with the ensemble mean, if the member has a certain level of smoothing built in. That all top 10 members are from system D, is a very clear indication of that spatial resolution pays off.

A.8 Verification of Frequency Distribution of generation

generation	5	15	25	35	45	55	65	75	85	95
measurement	29.3	10.4	7.2	7.1	6.5	7.0	7.4	8.3	10.3	6.5
9H-forecast	27.9	10.1	8.8	7.9	7.0	6.5	7.2	9.6	11.3	3.6
45H-forecast	25.9	11.4	10.1	8.3	7.5	6.3	7.3	9.5	11.6	2.2

Table A.8: Frequency Distribution of generation for Existing Facilities in the project period

The frequency distribution of the output of the single forecasts should be close to the measurement frequency in order to achieve reasonable ramp rates. This the assumption behind WEPROG's optimization. The frequency distribution is slightly worse for the longer horizons than for the shorter horizon. With high uncertainty we must expect that

the combined forecast will underpredict high generation and overpredict low generation. This is also what we see. The general pattern is that the system was capable of producing full generation and no generation at longer forecast horizons and therefore also ramped reasonable. Without such ramping capabilities, it would not be possible to also predict the error of the ramp rate forecast with a correlation of 0.4.

A.9 Verification of Existing Facility Wind Power at 0-4hours

fcl	Raw Forecast	MSEPS+Online data	Persistence
1	12.3	5.1	5.1
2	12.5	7.9	8.3
3	12.8	10.0	10.9
4	13.1	11.4	13.0

Table A.9: Comparison of MAE error of wind power for the ultra short-term forecasting at existing facilities using the *WEPROG ensemble algorithm for short-term predictions* for the forecasting horizon 0-4 hours ahead delivered to the AESO pilot project in the period May 2007 - April 2008. The Raw forecast is based on historic data and does not know the recent measurement. The MSEPS+Online data combines the Raw forecast and the recent measurement used in the Persistence forecast.

The 0-4h short-term verification for the existing facility power shows that there is room for improvement. The algorithm was the simplest possible and it is at least not worse than persistence. An increase in inertia from a larger capacity will improve the result. A large fraction is small amplitude variability with low predictability, which will gradually disappear with increasing capacity. It will also help to get more represenative ensemble spread. It is very difficult to estimate what the error could be in the future, because the inertia, the ensemble and the methodology will be different. Improvements will however only appear really signifcant at higher levels of inertia.

A.10 Verification of Ramp rate Error's correlation with Ramp rate Spread

fcl	A	B	D
3	0.28	0.37	0.38
9	0.35	0.32	0.39
15	0.32	0.31	0.38
21	0.29	0.30	0.36
27	0.28	0.26	0.34
33	0.28	0.25	0.35
39	0.28	0.24	0.34
45	0.29	0.22	0.34

Table A.10: Ramp rate Error's correlation with Ramp rate Spread for Existing Facilities in the project period

The prediction of ramp rate error is forecasted very good by system D. A correlation of 0.4 between ensemble spread and actual error on the ramping shows that the systems knows when there is a risk of a ramp very well. While the RMS requires exact timing and shows the same skill of A, B and D, we observe that when we look at the collective result of all members there is more skill in D. A parameter can obviously be so difficult to predict that a single value hits random, but it is obviously still possible to get it's variability under control. Ramps do appear as spikes with varying sign and this is the worst type of field to verify with RMS. It is on the other hand not surpricing that an ensemble tends to cause spread exactly around the spikes.

A.11 Extreme Ramp Event Analysis

To analyse the ramping capability of the MSEPS probabilistic forecasts that were not delivered to the project, we extracted those cases, where the measurements showed a change in power of equal or more than 80MW per hour during the year. Instead of only looking at the minimum and maximum, we now verified, whether any of the ensemble members ramped correct. There were 69 events found, which were studied in detail and the statistical results are presented hereafter .
The contingency table is a useful way to see what types of errors have ben made. We generated such contingency tables for 4 different categories of pase errors (0h,1h,2h,3,4h), where we allowed a magnitude error of up to 5% (all observed ramp rates were multiplied by 0.95).
The contingency table categorises hits, misses, false alarm and correct negatives. A perfect forecast system would produce only hits and correct negatives, (i.e. no ramp observed

and no ramp forecasted) and no misses or false alarms. The contingency table provides the possibility to classify how far the current forecast model is away from being "perfect".
Contingency Table for allowed phase errors of 0H

		OBS		
		yes	no	Total
fore-	yes	44	1530	1574
cast	no	25	775	800
		69	2305	

Table A.11: Contingency Table for Extreme Events with allowed phase errors of 0 hours.

Contingency Table for allowed phase errors of 1H

		OBS		
		yes	no	Total
fore-	yes	56	1131	1187
cast	no	13	1174	1187
		69	2305	

Table A.12: Contingency Table for Extreme Events with allowed phase errors of 1 hours.

Contingency Table for allowed phase errors of 2H

		OBS		
		yes	no	Total
fore-	yes	60	903	963
cast	no	9	1402	1411
		69	2305	

Table A.13: Contingency Table for Extreme Events with allowed phase errors of 2 hours.

Contingency Table for allowed phase errors of 3H

		OBS		
		yes	no	Total
fore-	yes	62	761	823
cast	no	7	1544	1551
		69	2305	

Table A.14: Contingency Table for Extreme Events with allowed phase errors of 3 hours.

Contingency Table for allowed phase errors of 4H

		OBS		
		yes	no	Total
fore-	yes	62	632	694
cast	no	7	1673	1680
		69	2305	

Table A.15: Contingency Table for Extreme Events with allowed phase errors of 4 hours.

A number of categorical statistics have been computed from the elements in the contingency table to describe particular aspects of forecast performance. These are:

1 Accuracy (ACC)
2 Probability of detection (POD)
3 Probability of false detection (POFD)
4 Odds Ratio (OR)
5 True Skill Statistic (TSS)

These categoral statistics will answer the following questions:
1 What fraction of the forecasts was correct?
2 What fraction of the observed *yes* events were correctly forecast?
3 What fraction of the observed *no* events were forecasted as *yes*?
4 What is the ratio of the odds of a forecast being correct or being wrong?
5 How well did the forecast separate the events from the non-events?

Accuracy

The Accuracy measure is simple and intuitive statistics. It can however be misleading, because it is heavily influenced by the most common category. The range is from 0 to 1.

$$Accuracy\ (fraction\ correct) = \frac{(hits + correct\ negatives)}{total} \tag{A.1}$$

It can be seen that for the accuracy measure, i.e. the fraction of forecasts that was correct is increasing with increasing allowance of a phase error. This is not surprising, as the simple test only measures the ratio of hits and correct forecasted "non-events" to total events.

Name	ID	allowed phase err	Value	No skill	Perfect skill
Accuracy	ACC	0H	0.322	0	1
Accuracy	ACC	1H	0.483	0	1
Accuracy	ACC	2H	0.574	0	1
Accuracy	ACC	3H	0.631	0	1
Accuracy	ACC	4H	0.681	0	1

Table A.16: Probabilistic verification of *Accuracy* on the 70 Extreme Ramp Events in the period May 2007-April 2008, answering the question *What fraction of the forecasts was correct?*.

Probability of detection

The Probability of detection (POD) is sensitive to hits, but ignores false alarms. It is said to be very sensitive to the climatological frequency of the event and hence a suitable measure for rare events.
It is defined as:

$$POD = \frac{hits}{(hits + misses)} \tag{A.2}$$

The most interesting aspect of this result is that the probability of detecting an event seems to have a threshold value at 3 hours. Thereafter, there is no improvement of the skill by allowing a higher phase error. This is consistent with the 6 events that were found "non-predictable" within the 69 extreme ramp events (see Table 2.2 in Section 2.4.1).

Name	ID	allowed phase err	Value	No skill	Perfect skill
Probability of detection	POD	0H	0.638	0	1
Probability of detection	POD	1H	0.812	0	1
Probability of detection	POD	2H	0.870	0	1
Probability of detection	POD	3H	0.899	0	1
Probability of detection	POD	4H	0.899	0	1

Table A.17: Probabilistic verification of *Probability of detection* on the 70 Extreme Ramp Events in the period May 2007-April 2008, answering the question *What fraction of the observed "yes" events were correctly forecast?*.

Probability Of False Detection

The probability of false detection (POFD) is sensitive to false alarms, but ignores misses. It is an important component of the Relative Operating Characteristic (ROC) used widely for probabilistic forecasts.

It is defined as:

$$POFD = \frac{false\ alarm}{(correct\ negatives + false\ alarm)} \tag{A.3}$$

From the results in Table A.18 it can be seen that the threshold at 3 hours is also visible when looking at the probability of false detection. While there the longer phase errors still reduces the false alarms, it is much less after 3 hours than from 0-3h.

Name	ID	allowed phase err	Value	No skill	Perfect skill
Probability of false detection	POFD	0H	0.664	1	0
Probability of false detection	POFD	1H	0.491	1	0
Probability of false detection	POFD	2H	0.392	1	0
Probability of false detection	POFD	3H	0.330	1	0
Probability of false detection	POFD	4H	0.274	1	0

Table A.18: Probabilistic verification of *Probability of false detection* on the 70 Extreme Ramp Events in the period May 2007-April 2008, answering the question *What fraction of the observed "no" events were forecasted as "yes"?.*

Odds Ratio

The odds ratio (OR) measures the ratio of the odds of making a hit to the odds of making a false alarm. It is a suitable measure for rarer events and is less sensitive to hedging. Note, the odds ratio is not the same as the ratio of the probability of making a hit (hits / # forecasts) to the probability of making a false alarm (false alarms / # forecasts), since both of those can depend on the frequency (i.e., the prior probability) of the event.
It is defined to:

$$OR = \frac{(hits + correct\ neg)}{(misses + false\ alarms)} = \frac{\frac{POD}{(1-POD)}}{\frac{POFD}{(1-POFD)}} \tag{A.4}$$

The odds ratio shown in Table A.19 behaves similar to the probability of falls detection. While it increaes most from allowances of 0-3h of phase errors, the ratio of succeeding against failing to forecast an extreme ramp event is less increasing for more than 3 hours phase error allowance. This confirms the fact that the odds of predicting an event after 4h or not are not increasing, but rather deteriorating. The numbers are relatively small, i.e. at 0h phase error allowance, the odds of predicting the ramp and not predicting the ramp are less than 1 time higher. When allowing 2 hours, the odds however already increase by almost 4, i.e. the likelyhood of predicting a ramp correct than failing to predict it is 4.5 times higher. The reason for the small numbers is due to the fact that the numbers of ramping events are relatively small in comparison to the numers of hours of the test (69 versus 2305).

Name	ID	allowed phase err	Value	No skill	Perfect skill
Odds Ratio	OR	0H	0.892	1	∞
Odds Ratio	OR	1H	4.471	1	
Odds Ratio	OR	2H	10.351	1	
Odds Ratio	OR	3H	17.970	1	
Odds Ratio	OR	4H	23.446	1	

Table A.19: Probabilistic verification of *Odds Ratio* on the 70 Extreme Ramp Events in the period May 2007-April 2008, answering the question *What is the ratio of the odds of a forecast being correct or being wrong?*.

True Skill Statistic

The true skill score TSS, also known as "Hanssen and Kuipers discriminant" or "Peirces's skill score", uses all elements in the contingency table and does not depend on the frequency of the events. The true skill statistics is identical to TSS = POD - POFD, whereas the TSS can also be interpreted as (accuracy of events) + (accuracy of non-events) - 1. For rare events TSS is weighted towards the first term (POD). It ranges from -1 to 1.
It is defined to:

$$TSS = \frac{hits}{(hits + misses)} - \frac{false\ alarm}{(false\ alarm + correct\ neg)} \tag{A.5}$$

Table A.20 confirms the findings from the probability of detection (POD) scores and demonstrates the weight towards the first term in the equation, the POD. In this test the improvement seems to decrease even earlier. Already after 2 hours phase error allowance the skill shows less than 10% and less than 5% improvements in skills for the allowance of 3h and 4h phase errors, respectively.

Name	ID	allowed phase err	Value	No skill	Perfect skill
True Skill Statistic	TSS	0H	-0.026	0	1
True Skill Statistic	TSS	1H	0.321	0	1
True Skill Statistic	TSS	2H	0.478	0	1
True Skill Statistic	TSS	3H	0.568	0	1
True Skill Statistic	TSS	4H	0.624	0	1

Table A.20: Probabilistic verification of *True Skill Statistic* on the 70 Extreme Ramp Events in the period May 2007-April 2008, answering the question *How well did the forecast separate the events from the non-events?*.

Appendix B

The results in Appendix B are created under the same assumptions as the results in Appendix A with the following exceptions:

Only Q4 of the project period is included starting 2nd of February and ending 30th of April 2008.

The statistical training procedure used in the results in this section allows for 1 direction bin and 10 wind speed bins.

All systems are calibrated towards no bias, if possible with their own forecasts for Q4.

The super ensemble (J) is defined as a 75 member ensemble where each member is an average of A, B, D, F, G, H and I. Each member is therefore the ensemble mean of 7 members with the same model equations.

These assumptions prevent that the statistical training can learn the data set, thus the error measures are reasonable and sufficiently accurate to evaluate the various ensembles against each other, because the target is not to evaluate the wind to power conversion process, but if there exist weather forecast methods that are better than others.
The detailed system description of the 9 configurations is given in Section 2. Here we only note that A, B and K refer to 45/60km spatial resolution, while the configurations D, F, G, I are the 22km resolution, H is the 6km resolution and J is the super ensemble of all excluding K.

B.1 Verification of normalised Mean Absolute Error for Q4

fcl	A	B	D	F	G	H	I	J	K
3	14.70	15.94	13.65	12.01	13.04	11.99	12.84	12.62	16.86
9	15.04	14.45	12.06	11.85	11.94	12.24	12.25	11.89	17.22
15	16.96	16.73	13.31	12.79	12.90	13.35	13.14	13.26	19.23
21	17.22	17.72	14.18	13.87	13.70	13.61	13.59	13.97	19.79
27	17.57	18.78	15.04	14.79	14.41	13.73	14.26	14.66	20.92
33	17.53	19.33	15.46	15.57	15.41	14.42	15.14	15.32	21.19
39	17.50	19.34	16.04	16.54	16.12	15.18	16.10	16.02	na
45	18.31	20.09	16.71	17.22	16.97	16.34	16.98	17.01	na

Table B.1: Normalised Mean Absolute Error for Existing Facilities in Q4

The MAE error analysis shows that system H has a very low error growth until 39 hours, but that there is a jump at 45 hours. There is a risk that this can be triggered by technical problems in the beginning of the test period. We tried to predict 72 hours ahead, but there was not enough CPU time on the system, if the weather was very bad (typically in snow storms). Such events may increase the required CPU time with 10-15%, which was enough to cause problems in keeping the schedule. There were incomplete forecasts until it was decided to reduce the forecast horizon to 48 hours. Since then there were no problems in getting the members to complete, but we lost the possibility to carry this analysis on to 72 hours. We are therefore almost sure that the error growth at the end is linked to this issue, because there would not be a physical reason why the error pattern should change at that horizon.

The error at the shorter horizon of system H is certainly not impressing. There is an obvious task in improving the initial conditions. The problem is most likely the interpolation from a near 100km resolution to 6km resolution. This brings some dynamical imbalances as it takes quite many hours (27) before the 6km ensemble is better than the others. System I is doing good considering that Environment Canada's global model is only available twice per day. System A has a very low error growth in this period showing the importance of a tall orography and probably also a single large area. System B and K show that, if care is not taken, then the error would have been even higher.

B.2 Verification of RMS error for Q4

fcl	A	B	D	F	G	H	I	J	K
3	20.13	21.64	18.50	16.82	17.81	16.39	17.44	16.93	22.23
9	20.56	19.71	16.77	16.47	16.65	16.61	17.16	16.39	22.42
15	22.71	22.25	18.31	17.69	17.93	18.05	18.20	17.98	24.56
21	22.88	23.40	19.51	19.06	18.94	18.55	18.60	18.82	25.24
27	23.44	24.52	20.52	20.01	19.49	18.64	19.29	19.50	26.57
33	23.56	25.41	21.04	20.78	20.51	19.45	20.33	20.28	26.78
39	23.53	25.31	21.76	21.70	21.27	20.01	21.18	21.03	n/a
45	24.41	26.24	22.71	22.44	21.98	21.21	22.00	22.15	n/a

Table B.2: RMS error for Existing Facilities in Q4

System D degrades with forecast length on the RMS error compared to the other 22km ensembles. This shows that having 75 members and a "wilder" spread from the lateral boundary conditions compared to only 8 conservative lateral boundaries is critical for the RMS error. It shows that the uncertainty is not caused on the large scale, but is generated within the 22km model area. This in fact shows that it is not unreasonable to run a smaller large scale ensemble as the various global forecasting services do. But the local scale is more complicated and requires more ensemble members. This is a strong argument for using limited area models for ensemble forecasting, because they can be tuned to make the ensemble spread, where it is needed. It can be concluded that system D obviously makes larger mistakes due to the 75 different boundary conditions than F, G and I. Some of the mistakes can be triggered by the poor performance of B, because it is B that provided the lateral boundaries to D.
System H lies considerably under system D and also system G. System G has the same boundaries as H, but runs in the same resolution as D. System F, which has 45km boundaries, but only 8 conservative members lies in the middle between D and G. That shows that reducing the number of different lateral boundaries and increasing the spatial resolution of the boundaries both gives an add on. Further improvement is then achieved by also increasing the resolution of the main inner 75 member ensemble since H is better than G.
The result is as expected, but it is still not satisfactory that the initial error of H is not reduced considering that there is so much more detail.
The "super ensemble" has a good performance although system A and B are included with the same weight as all others. This means that a very smooth forecast can have a quite good RMS score.

B.3 Verification of Bias for Q4

fcl	A	B	D	F	G	H	I	J	K
3	2.16	0.77	0.78	1.17	0.96	1.27	0.69	0.57	0.10
9	1.53	1.21	0.33	0.98	0.87	0.54	0.76	0.35	-1.21
15	-0.06	0.32	-0.78	-0.47	-0.50	-0.06	0.46	-0.68	-1.53
21	0.39	0.49	-1.30	-0.29	-0.25	-0.28	0.15	-0.84	-1.38
27	0.06	-0.56	-2.12	-0.72	-0.00	-0.23	0.03	-1.15	-3.31
33	0.01	-1.05	-2.00	-0.67	-0.61	-0.59	-0.14	-1.29	-2.92
39	0.14	-0.90	-2.16	-0.74	-0.75	-0.80	-0.45	-1.29	n/a
45	-0.92	-1.45	-2.36	-0.88	-0.44	-0.55	-0.39	-1.54	n/a

Table B.3: Bias for Existing Facilities in Q4

All setups with envelope orography have a low bias, but B and K drift away. The statistical training can not even correct for this. All setups have some problems in the very short term with apparent noise.

B.4 Verification of Dispersion for Q4

fcl	A	B	D	F	G	H	I	J	K
3	19.83	21.42	18.20	16.63	17.55	15.90	17.18	16.49	21.98
9	20.30	19.61	16.57	16.27	16.45	15.98	16.86	16.02	22.21
15	22.42	22.09	18.01	17.43	17.64	17.14	17.85	17.46	24.08
21	22.64	23.20	19.16	18.80	18.69	17.70	18.25	18.31	24.60
27	23.07	24.13	19.90	19.52	19.04	17.70	18.80	18.76	25.52
33	23.22	24.92	20.38	20.16	19.80	18.44	19.73	19.43	25.67
39	23.13	24.62	21.00	20.86	20.45	18.74	20.32	20.04	-
45	23.77	25.44	21.86	21.38	21.02	19.71	20.98	21.00	-

Table B.4: Dispersion for Existing Facilities in Q4

The dispersion is one of the parameters that shows the skill of H best. It is the phase errors that reduce most with spatial resolution. The error growth of the dispersion is low, but there is not much doubt that there is much more to gain on that term with better initial conditions, bigger model area and more accurate orography.

B.5 Verification of SDBIAS for Q4

fcl	A	B	D	F	G	H	I	J	K
3	-2.73	-2.96	-3.26	-2.21	-2.90	-3.77	-2.93	-3.79	-3.33
9	-2.90	-1.57	-2.59	-2.38	-2.46	-4.53	-3.05	-3.46	-2.86
15	-3.61	-2.65	-3.21	-2.99	-3.18	-5.67	-3.56	-4.25	-4.58
21	-3.29	-3.08	-3.45	-3.09	-3.08	-5.56	-3.62	-4.25	-5.47
27	-4.16	-4.32	-4.50	-4.37	-4.12	-5.86	-4.32	-5.21	-6.59
33	-4.01	-4.85	-4.83	-4.99	-5.33	-6.14	-4.92	-5.65	-7.04
39	-4.29	-5.77	-5.29	-5.91	-5.81	-6.96	-5.95	-6.25	na
45	-5.47	-6.23	-5.66	-6.75	-6.41	-7.84	-6.62	-6.88	na

Table B.5: SDBIAS for Existing Facilities in Q4

The results of the SDBIAS are somewhat surprising, because it is H that has the lowest variability of all. This shows the value of running an ensemble instead of only a single model. A single model is normally showing much stronger variability and increases the RMS error, but what we achieved was the opposite by increasing the resolution. The combination of many high resolution forecasts is smoother than the combination of many models in less extreme resolution.
There are several explanations. One is that the model error reduces and another is that the high resolution model allows for the gradients that natively exists, whereas the low resolution model truncates the gradients. Also, that the model system has much more degree of freedom means that ensemble members become more uncorrelated, if there is no clear trend. Thus the ensemble is smoothing automatically while the wind is scattered, but rather making the gradient more detailed in periods with little variability.

B.6 Verification of Correlation for Q4

fcl	A	B	D	F	G	H	I	J	K
3	0.80	0.77	0.83	0.86	0.84	0.87	0.85	0.86	0.75
9	0.79	0.81	0.86	0.87	0.86	0.86	0.85	0.87	0.75
15	0.74	0.75	0.83	0.84	0.84	0.83	0.83	0.84	0.68
21	0.74	0.72	0.81	0.82	0.82	0.83	0.83	0.82	0.66
27	0.72	0.69	0.79	0.80	0.81	0.82	0.81	0.81	0.62
33	0.71	0.66	0.77	0.78	0.78	0.81	0.79	0.79	0.61
39	0.71	0.66	0.75	0.75	0.76	0.79	0.76	0.77	na
45	0.68	0.62	0.73	0.73	0.74	0.76	0.74	0.74	na

Table B.6: Correlation between forecast and measurement for Existing Facilities in Q4

The correlation is not so much in favour of system H as one would expect from the dispersion term. We can from this conclude that system H wins on small details required to predict the phase of events, rather than a general ability to predict high and low generation periods. There, all setups seem more similar in performance. The correlation counts in the hours, where the production is extreme high or low, while the dispersion counts more during the ramping hours. Obviously all low resolution setups have poor correlation.
What is probably characteristic for D, F, G and I is that they produce power at the right time, but not the correct amount, because they tend of have the same wind everywhere in the facility area. Thus, all facilities have some power, while system H rather give full generation at some facilities and no generation at other facilities, which is obviously less than full generation. This is more representative for the actual generation pattern.

B.7 Verification of Predictability of Error for Q4

fcl	A	B	D	F	G	H	I	J
3	0.14	0.10	0.31	0.29	0.27	0.28	0.21	0.32
9	0.19	0.22	0.33	0.32	0.32	0.42	0.27	0.34
15	0.12	0.21	0.20	0.23	0.23	0.37	0.26	0.23
21	0.18	0.24	0.18	0.20	0.22	0.38	0.28	0.22
27	0.17	0.19	0.19	0.20	0.27	0.41	0.29	0.23
33	0.19	0.18	0.19	0.17	0.28	0.37	0.29	0.23
39	0.19	0.18	0.20	0.17	0.25	0.37	0.28	0.22
45	0.18	0.12	0.21	0.18	0.26	0.42	0.27	0.22

Table B.7: Correlation between ensemble spread and error for Existing Facilities in Q4

The ability to predict errors with system H is outstanding. It shows the ensemble is representing the actual error sources. This is, the strong gradients of the downdrafts. The correlation hardly deteriorates with forecast horizon, so there is reason to expect that reserve can be computed quite well with system H, most likely also 72 hours in advance.
The result also shows that the ensemble members in H have the correct structure, but that it is the timing and position that is the main problem. This is consistent with previous conclusions on the details from the correlation and the dispersion term. It is exactly what we hoped to generate, when the plan of system H was made. We did not believe in the individual high resolution forecasts, but rather that it would be possible to get insight in when there is error and when there isn't. Knowing this result, then it is a good idea to have a central system with this type of forecasting to secure a fair market for reserve in order to reduce volatility on pricing, because this computation can be done without any data from the facilities.

The geographical location is sufficient. This result expresses the core of high resolution ensemble forecasting and why it can add value in large-scale integration of wind power, if the market for reserve and primary power is integrated in some way. Then day-ahead forecasting is not to be considered as a bottleneck for the wind power integration.

B.8 Verification of Ramp rate RMS for Q4

fcl	A	B	D	F	G	H	I	J
3	6.43	6.60	6.70	6.70	6.68	6.68	6.54	6.50
9	6.58	6.57	6.78	6.71	6.57	6.94	6.44	6.45
15	6.45	6.26	6.62	6.49	6.43	6.43	6.41	6.38
21	6.64	6.43	6.64	6.59	6.73	6.75	6.53	6.49
27	6.31	6.32	6.69	6.75	6.58	6.58	6.44	6.48
33	6.64	6.70	7.04	6.60	6.85	6.68	6.88	6.70
39	6.41	6.56	6.62	7.01	6.89	6.50	6.62	6.47
45	6.48	6.65	6.57	6.65	6.49	6.70	6.35	6.47

Table B.8: Ramp rate RMS for Existing Facilities in Q4

There is a tendency to that the "super ensemble" is doing best on ramp rate RMS verification, but this is marginal. It appears like smooth forecasts are doing best on this parameter. As previously pointed out, the ramps are short like spikes with changing sign and the likelihood of short significant errors is high in all configurations.
It is unfortunate that this parameter is not better predicted, because it shows something about the required amount of generation units during a ramp, which can be different than after the ramp, if the the ramp is so steep that no single plant can manage to compensate for the drop in the wind generation. Such ramps may be nearly double expensive and where curtailment would be an option to reduce the cost of wind generation integration.

B.9 Verification of Ramp rate Error's correlation with Ramp rate Spread for Q4

fcl	A	B	D	F	G	H	I	J
3	0.26	0.40	0.43	0.33	0.38	0.35	0.42	0.46
9	0.26	0.19	0.33	0.34	0.39	0.36	0.27	0.36
15	0.22	0.21	0.28	0.25	0.24	0.30	0.28	0.24
21	0.24	0.34	0.30	0.27	0.28	0.33	0.24	0.30
27	0.27	0.21	0.36	0.33	0.30	0.34	0.33	0.29
33	0.20	0.22	0.29	0.30	0.33	0.21	0.36	0.31
39	0.23	0.19	0.28	0.33	0.34	0.23	0.33	0.29
45	0.25	0.21	0.35	0.32	0.27	0.28	0.28	0.26

Table B.9: Ramp rate Error's correlation with Ramp rate Spread for Existing Facilities in Q4

The prediction of the ramp rate error follows the pattern for the ramp rate RMS, but the results does not seem statistically significant. Nearly all columns have one forecast horizon that deviate from the other. We should therefore conclude that the results in this table should be regarded as uncertain.
There is not much to gain, but we observe that system H is doing quite well on the shorter horizons. We dare to say that this could be an indicator of that this parameter will improve once the initial conditions have been improved. This would mean that we have some hours before the ramp, where better probabilistic forecasts of the actual ramp rate could be expected. This result is therefore more encouraging than it looks like at first glance. That is, the ramp rate forecast will improve, if we can reduce numerical noise and increase the sharpness of gradients of the down draft zone.

B.10 Verification of Member Ranking for Q4

order	1	2	3	4	5	6	7	8	9	10
3	H2111	H2122	H3111	H0121	H1023	H2112	H4004	H1121	F2102	H3121
9	J2123	J4123	J3123	J1123	J2103	J1223	J3223	J2203	J2223	F2123
15	J1223	J2003	J2123	F2003	J2203	F3023	J3223	F2122	F2102	J3023
21	H1313	J2003	I1003	I3003	I1000	H4004	I2003	J3003	H0313	H0102
27	H1313	H0122	H1311	H1312	H0102	H4004	H4122	H0112	H2313	H2122
33	H1313	H1312	H1311	H4004	H2313	H2311	H2312	H2122	H2320	H1310
39	H1313	H1311	H4004	H1312	H2311	H4320	H2313	H2320	H2310	H3004
45	H4004	H1311	H1313	H1312	H1310	J3202	H2004	H2320	H1004	H0320

Table B.10: Member Ranking for Existing Facilities in Q4

The member ranking gives the impression that only the J and H were included, but all models were included. System J is doing well in the short-term, where there is noise in system H, again the problem with the initial conditions. There is a member **H1313** which is outstanding good, but not at 45 hours. Here the combined ensemble forecast (**H4004**) is better, but it may be that this is due to that some technical problem during the attempt to run 72 hours ahead caused trouble for **H1313**. We can however see that **H4004** is improving relatively with increasing horizon as it is getting further and further up in top 10. The member **H1313** has the wind diagnosed in a way as if the surface is ice-covered. We use this assumption, because the periods with sudden warming during the down draft may lead to a heavy ice like consistency of the snow cover, when it is freezing again. The roughness of wet frozen snow is somewhat higher than roughness average ice conditions, but the temperature changes in that surface is as ice, which is fairly invariant to sun light compared to what the underlying surface would have. The member **H1313** is doing very good and there is reason to ask why it can beat others so consistently on the medium horizons. We suspect that this member is better to dampen noise than the other members, so the initial noise disappears faster and the forecast is then sustained better than the others. What is characteristic about this member is that it has a damping advection scheme and additional damping, if there is roughness and strong wind correlates in the upstream direction, which is exactly the case in the southern Alberta at westerly wind. That this member is best again points back to the noise problem in the initial conditions and it shows what should be worked on.

The members named H1311 and H1312 are also in top 5 and they originate from the same weather ensemble member, but refer to the wind at 30 and 100m altitude. They are likely to be better, once the snow cover is gone.
The many results that we have shown are in fact consistent and give very strong evidence that the 6km ensemble or more generally a very high resolution ensemble is the way forward. Although we have seen that one particular model configuration outperforms the

other in this test, we expect that this will not be the case once work has been done on the initial conditions. The other ensemble members will improve relatively, so the member **H131[123]** will get more competition.

Considering that we have by no means an optimal roughness and orography and that the initial conditions can be improved, then we are convinced that the results that the project has demonstrated that improvements can be made. 6 months dedicated work on the problem will not eliminate the error, but reduce the error to maybe tolerable levels well below what the project has seen.
It goes beyond the scope of this report to explain the approximations that have been made in system H, because of the limited time, but they are so significant that we are sure they have been a limiting factor. It would however have been too much of a risk to try to improve on these during Q4. This would have led to results that could not have been compared directly. We wanted to see the effect of spatial resolution and increasing mountain height and the result is positive. There is for sure more to gain following this path.

In the pilot project we have set up the model system with approximately 1-2 days of work per model area. The weather forecasts from the operational systems in NCEP, Environment Canada and ECMWF are much closer to their optimal performance, because their scientists have worked for years on the mountain blocking problem. It would therefore be against all common sense, if we, with 1-2 days work, should have already reached the ultimate accuracy in the weather forecasts.
We should also acknowledge that we had no expectation that setup B would not have worked in Alberta based on experience from Europe, Australia and the Eastern part of the USA. But we encountered within half a year 3 areas where increased spatial resolutions were required. The 3 areas were Japan, Egypt and Alberta. They all benefit significantly from spatial model resolution. This is due to the trend that wind power is built, where the conditions are extreme windy which correlate with variable weather. Therefore, we must expect that this trend also requires higher spatial resolution at other places.
The equations in member **H131*** have previously been found to give the best result in Denmark for high resolution forecasting in 5km. However, the member was not performing better than others in Egypt during the summer. Therefore, such results cannot be generalised and this is why an ensemble is always worth the effort, because common sense and theory does not always give the correct answer to what is best in weather forecasting.

Verification of Accumulated Frequency Distribution of the Error for Q4

ERROR	A	B	D	F	G	H	I	J
100	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
81	98.90	98.40	98.00	98.70	99.40	99.80	99.60	99.50
64	92.90	91.50	95.40	95.90	96.30	96.90	96.60	96.10
49	85.80	84.60	90.50	90.30	90.00	92.10	89.50	90.80
36	75.90	72.30	80.20	79.40	78.80	80.80	78.70	79.80
25	59.90	55.20	63.80	59.40	61.60	62.50	61.90	61.00
16	39.40	33.20	41.90	36.30	38.20	41.70	38.20	38.20
9	16.20	11.90	14.10	14.60	15.90	18.80	15.80	14.10
4	3.40	2.60	3.70	4.20	3.60	4.80	3.70	3.80
1	0.00	-0.00	0.20	0.10	0.00	0.10	0.00	0.10

Table B.11: Accumulated Frequency Distribution of the Error for Existing Facilities in Q4

The RMS score penalises large errors most, but it is instructive to see the frequency distribution of errors for one horizon (39 hours). The first column shows the error level and the number in the following columns show the percentage of forecasts that are better than the error level given in column 1. The difference between the two first columns is the percentage of hours with more than 81% error. We observe that system H has most hours with less error already starting from the highest error level between 81 and 100%. Setup D has been the worst of all setups in the very top, but A and B show in general the worst result, if we disregard D at the highest error level.

We have demonstrated consistency in the results between a number of different statistical parameters. There is no doubt that system H provides the most useful results.

A frame work of many model configurations are required to explain model errors. We have seen how modest differences between A, B, D, F, G, H, I and K could be explained consistently. That was only possible, because there were in fact 75 x 3 months simulations behind the result. This is equivalent to nearly 19 years of simulations. Single model configurations can led to a random signal and consequently low confidence to the conclusions. Therefore, we can in general not be sure about the ranking tables, but the remainder of the results are very robust. Running an ensemble can therefore be considered as a robustification of the methodology.

Appendix C

C.1 PLATON - Graphical User Interface for the MSEPS Online Forecasts

In the pilot project, the forecasters were only asked to provide data. To demonstrate some of the advantages of the MSEPS 75 member ensemble prediction system, we granted AESO access to our "PLATON" graphical user interface for plotting of real-time data showing the probability forecasts and also horizontal maps of weather variables and wind power, as well as horizontal maps of wind power including the online power measurements.

When setting up the *AESO Testbed*, where 9 different ensemble systems were running in real-time, AESO was also able to follow the different model system's forecasts online.

The following graphs show examples from the PLATON version dedicated to the AESO pilot project. The menu to the left (orange background) is where the user can select different presentation forms, model systems, forecast areas or sites, dates and forecast lengths and variables or fields. There are time series provided for wind power and ramp rates, and horizontal fields for the entire ensemble and the mean, minimum and maximum of the ensemble for standard meteorological variable (e.g. wind, mslp, cloud cover, temperature, precipitation) and wind power. For these horizontal weather maps, a *validation feature* provides the possibility to compare previous forecasts valid at the same time to gain overview of the development and changes in the initial conditions. While the menu is designed for intuitive use, an online user-guide is available where the possibilities and features are explained. The user-guide can also be downloaded as pdf file.

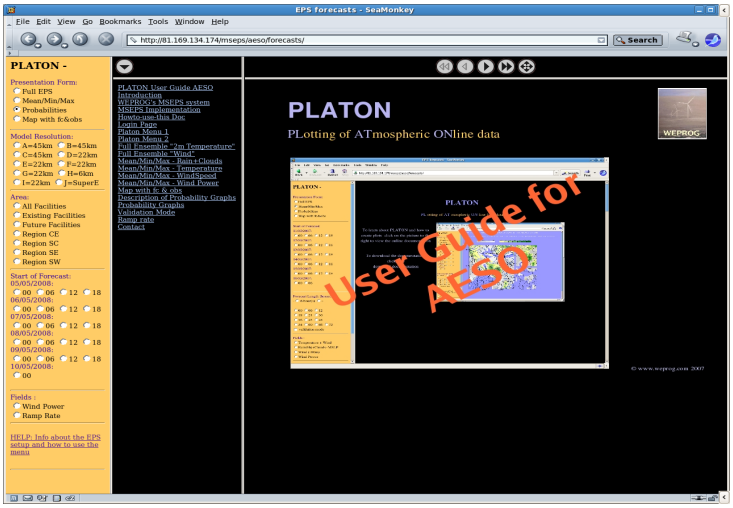

Figure C.1: Graphical User Interface PLATON: Frontpage and User-guide start page.

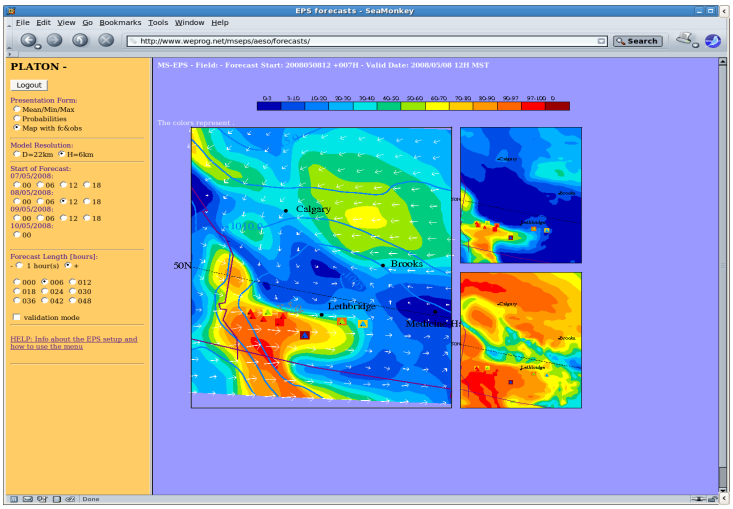

Figure C.2: PLATON: *Map with pwrobs*. This presentation form shows a horizontal map of potential wind power zoomed in to the region of interest together with site specific (trained) wind power forecasts as squares (forecasts) and triangles (measurements). The example shows a typical situation of a low pressure system built by the lee effect of the Rockies and extending out to the wind facility region, usually creating large uncertainties whether and when production starts and ends. The bigger figure shows the ensemble mean and the two small the minimum and maximum at each grid point. If a facility is in dark blue and red background on the min/max figures, then it means that the there could be no or full generation. A cutoff event shows up as a dark blue area enclosed in dark red.

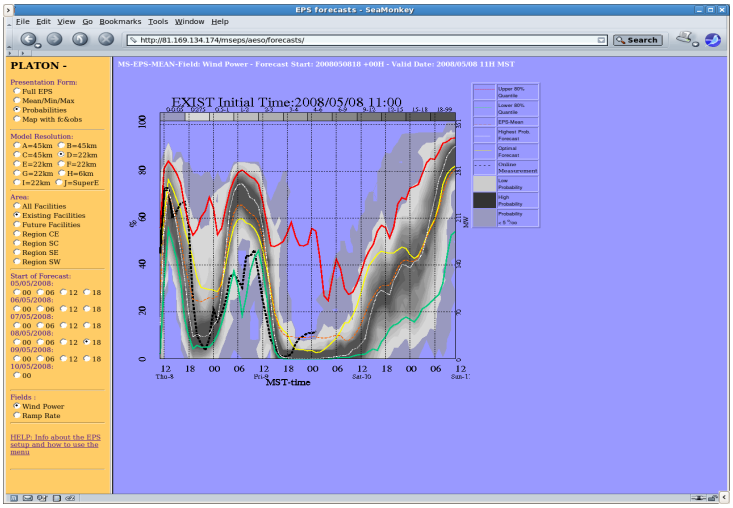

Figure C.3: PLATON: *Wind power probability.* This example shows the probability density of the forecast for the model system D, which delivered hourly 48h forecasts to the project. Instead of lines for each ensemble member, gray shading is used to display probabilities. The outer purple areas denote the 95% quantiles showing the high uncertainty that prevails many hours of the day.

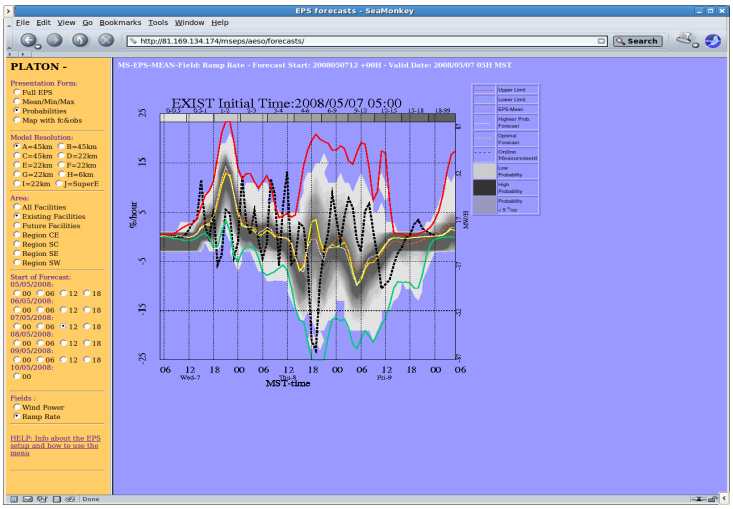

Figure C.4: PLATON: *Ramp rate* forecast. The ramp rate forecast is the change in power production over 1 hour and in conjunction with the probability graphs is providing the end-user with an overview of the situation at hand. In extreme cases, where there is a large spread, the ramp rate forecast shows the likelihood of a steep ramp, where the EPS mean or optimal forecast (yellow line) might be too conservative.

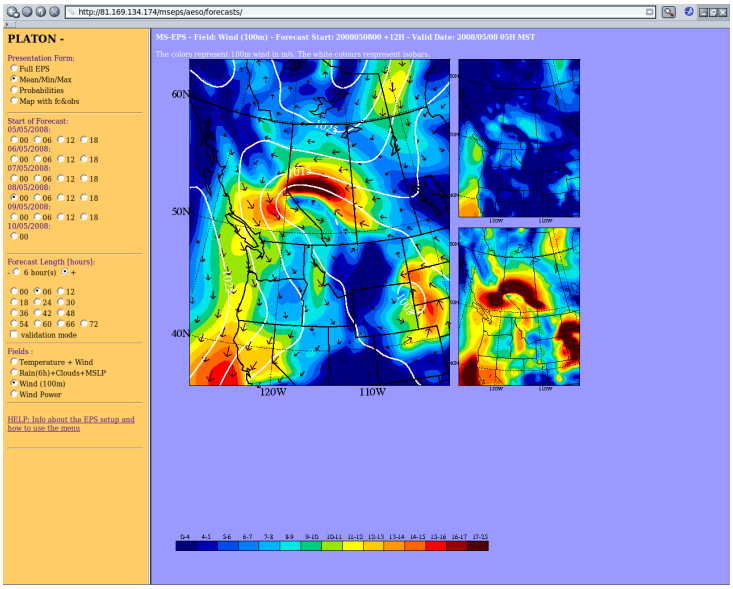

Figure C.5: PLATON: *Mean/Min/Max maps.* This presentations gives an overview of the large-scale weather over the area of interest. The EPS mean is the left larger figure, the EPS min is the upper right and the EPS maximum is the lower right figure. The example shows wind speed with direction and pressure isobars. The map is also available for Temperature, Precipitation and precipitation, and wind power potential.

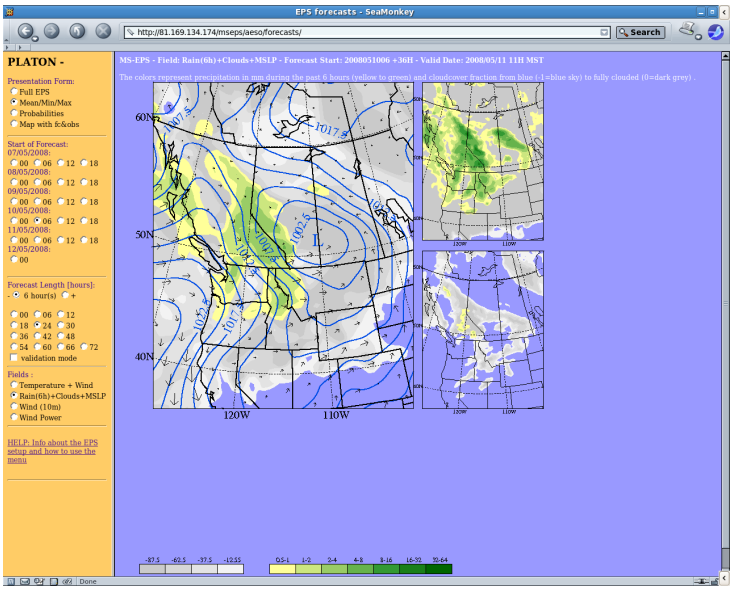

Figure C.6: PLATON:*Mean/Min/Max maps.* Example showing precipitation, cloud cover and isobars of mean sea level pressure. The maps can be used to forecast guaranteed sun/rain on region basis.

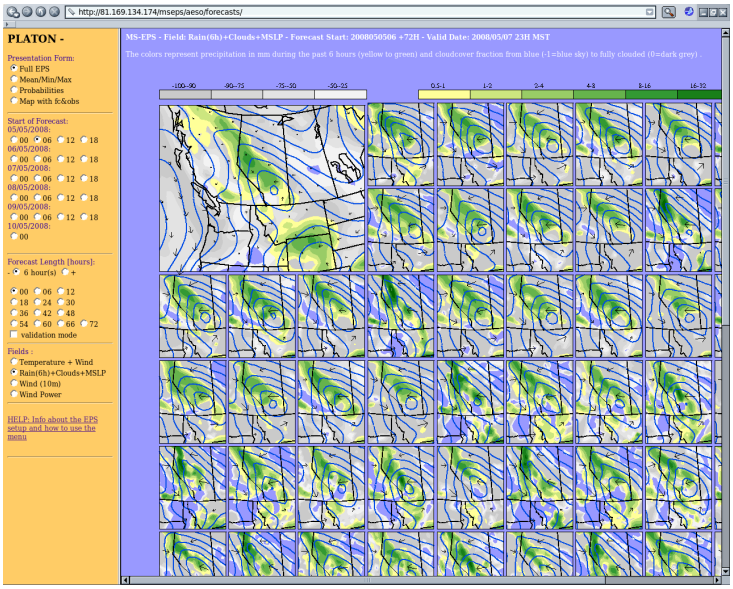

Figure C.7: PLATON: *Full Ensemble*. The full ensemble presentation shows horizontal graphs of all 75 ensemble members plus the EPS mean, which is the larger left upper figure. Although there is a lot of information in these maps, they provide the user with the full picture of the uncertainty of the chosen weather parameter in the region. The map is also available for wind speed, temperature, and wind power potential.

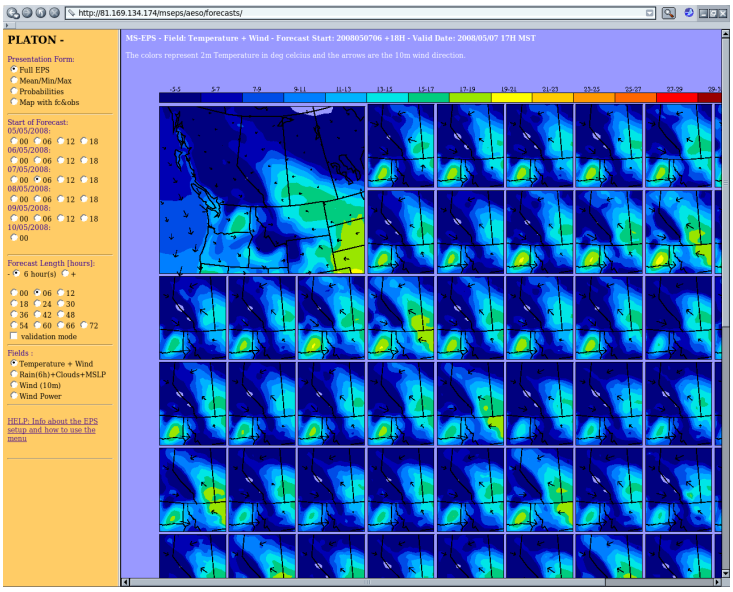

Figure C.8: PLATON: *Full Ensemble*. Example showing temperature and wind speed arrows.