

**Final Technical Report
Cover Page**

Federal Agency to which Report is submitted: DOE EERE – Wind & Water Power Program

Recipient: NOAA Earth Systems Research Laboratory

Award Number: DE-EE0003080

Project Title: The Wind Forecast Improvement Project (WFIP): A Public/Private Partnership for Improving Short Term Wind Energy Forecasts and Quantifying the Benefits of Utility Operations.

Project Period: February 1, 2011 - October 31, 2013

Principle Investigator: James Wilczak, NOAA/ESRL Team Lead- Boundary Layer Processes and Applications, james.m.wilczak@noaa.gov, 303-497-6245

Report Submitted by: Melinda Marquis, NOAA/ESRL Renewable Energy Manager, melinda.marquis@noaa.gov, 303-497-4487

Date of Report: April 30, 2014

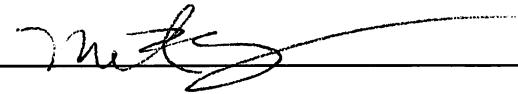
Covering Period: February 1, 2011 – October 31, 2013

Working Partners: WindLogics Inc., (Cathy Finley, Senior Scientist, cfinley@windlogics.com, 651-556-4283)

AWS Truepower LLC, (Jeffrey Freedman, currently Research Associate Professor, SUNY Albany, jfreedman@albany.edu, 518-437-8737)

Cost-Sharing Partners: None

DOE Project Team: DOE HQ Program Manager – Jose Zayas
DOE Field Contract Officer – Pamela Brodie
DOE Field Grants Management Specialist – Jane Sanders
DOE Field Project Officer – Brad Ring
DOE/CNVJ Project Monitor – Yelena Onnen

Signature of Submitting Official: 

This report is based upon work supported by the U. S. Department of Energy under Award No. DE-EE0003080. Any findings, opinions, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the Department of Energy.

TABLE OF CONTENTS

TABLE OF CONTENTS	i
<i>Executive Summary.....</i>	1
1. Project Overview.....	4
1.1 Goals and Key Tasks	4
1.2 Team Partners, Two Study Areas.....	8
2. WFIP Observations	11
2.1 Instrumentation.....	11
2.2 Site Selection and Preparation, Leases, Data Transmission and Handling	21
2.3 Data Quality Control and Instrument Performance	22
2.3 Instrument Inter-comparisons	31
3. NOAA Models	34
3.1 Rapid Update Cycle (RUC)	35
3.2 Rapid Refresh (RAP).....	36
3.3 High Resolution Rapid Refresh (HRRR).....	37
3.4 NAM and NAM CONUSNEST	37
3.5 RAP and HRRR improvements.....	40
3.6 HPC & Data Storage Requirements.....	40
4. Data Assimilation	41
4.1 RAP/HRRR	43
4.2 NAM/NDAS and CONUSnest.....	44
4.3 Additional Observational Data Processing	45
4.4 GSI 3DVar parameter settings.....	46
5. Evaluation of Real-Time Forecasts	48
5.1. Real-time model evaluation web site	48
5.2. Conversion of wind speed to power.....	49
5.3. Bulk error statistics: RAP and RUC models	50
5.4. Bulk Error Statistics: ERSL RAP and HRRR.....	54
6. Data Denial Simulations	56
6.1. Observations assimilated	56
6.2. Data denial simulation dates	57
6.3. Model bias estimation	57
6.4. Wind profiler evaluation	67
6.5. Sodar evaluation.....	71
6.6. Tall tower evaluation	73

6.6.1. Bias correction sensitivity	73
6.6.2. NSA/SSA & Forecast length	74
6.6.3. Seasonal variation	77
6.6.4. Validation hour sensitivity	79
6.6.5. Observed power dependence	83
6.6.6. Large forecast errors	83
6.6.7. Effects of spatial averaging	86
6.6.8. Geographic outlier sensitivity analysis	90
6.7. NAM results.....	92
6.7.1 Wind Profiler and sodar verification	92
6.7.1 NAM/NDAS Conventional verification over the Plains	97
6.7.2 Tall tower and nacelle verification	100
7. Ramp Tool and Metric	105
7.1 Background	105
7.2 Ramp definition and identification	108
7.2.1 Fixed Time Interval Method	109
7.2.2 Min-Max Method	110
7.2.3 Explicit Derivative Method	111
7.3 Matching of forecast and observed ramps	112
7.4 Forecast skill scoring methodology using single ramp definition	113
7.5 Forecast skill scoring: Matrix of skill values.....	118
7.6 Results from the WFIP data denial experiments.....	119
8. Surface flux and wind profile observations	129
9. Summary and Conclusion	135
10. References	140
Appendices	146
List of Figures	150
List of Tables	157
List of Acronyms	158

Executive Summary

The Wind Forecast Improvement Project (WFIP) is a U. S. Department of Energy (DOE) sponsored research project whose overarching goals are to improve the accuracy of short-term wind energy forecasts, and to demonstrate the economic value of these improvements. WFIP participants included DOE and National Oceanic and Atmospheric Administration (NOAA) laboratories; the NOAA National Weather Service (NWS); and two teams of partners from the private sector and university communities, led by AWS Truepower and WindLogics.

WFIP considered two avenues for improving wind energy forecasts. The first was through the assimilation of new meteorological observations into numerical weather prediction (NWP) models. New instrumentation was deployed or acquired during concurrent year-long field campaigns in two high wind energy resource areas of the U.S. The first was in the upper Great Plains, where DOE and NOAA partnered with the WindLogics team. The second field campaign was centered in west Texas, where DOE and NOAA partnered with the AWS Truepower team. The WFIP observing systems included 12 wind profiling radars, 12 sodars, and several lidars. In addition, WFIP allowed for NOAA to collect and assimilate for the first time proprietary tall tower (184 sites) and wind turbine nacelle anemometer (411 sites) meteorological observations from the wind energy industry. A necessary key component of WFIP was to develop improved quality control (QC) procedures to ensure that the assimilated observations were as accurate as possible, as a few erroneous observations can easily negate the positive impact of many accurate observations when assimilated into a NWP model. With proper data QC algorithms applied, good agreement was found between the co-located sodar, wind profiling radar, and lidar observed wind speeds.

The second avenue for improving wind energy forecasts was to improve the NWP models directly. Midway through the WFIP field program, NOAA/NWS upgraded its operational hourly-updated NWP forecast model from the Rapid Update Cycle (RUC) model to the Rapid Refresh (RAP) model, and the impacts of this upgrade have been evaluated using WFIP observations. During the course of WFIP NOAA/ESRL made further improvements to the research version of the RAP, and to the research High Resolution Rapid Refresh (HRRR) model, incorporating more advanced model physics and numerics, new data types assimilated, and better data assimilation procedures. Also, with WFIP funding NOAA was able to obtain the computer infrastructure to make the massive amounts of raw model output from the HRRR model available in real-time to both the two private sector teams, as well as to the entire wind energy industry.

Pseudo-power forecasts were evaluated by converting tall tower (mostly 60-80m) and model wind speeds to equivalent power using a standard International Electrotechnical Commission Class 2 (IEC2) power curve. Percent mean absolute error (MAE) power improvements between the NWS RUC operational hourly-updated forecast model and the real-time research NOAA/Earth System Research Laboratory (ESRL) RAP hourly-updated forecast model, calculated over the first 6 months of the WFIP field campaign, were significant. In the Northern Study Area (NSA) a 13% power improvement at

forecast hour 01 was found, decreasing to a 6-7% improvement at forecast hour 15. In the Southern Study Area (SSA) a 15% power improvement at forecast hour 01 was observed, decreasing to 5% improvement for a 15 h forecast. This improvement reflects the combined effects of the better RAP model versus the RUC model, as well as the contribution from assimilation of the WFIP observations into the research RAP model.

To quantify the impact of assimilation of the additional WFIP observations only, data denial (DD) experiments were run with the RAP and the NWS/North American Mesoscale (NAM) models. Six DD episodes were run with the RAP, each from 7-12 days long, spanning all four seasons of the year. Using conventional statistical analysis with the tall tower data sets for verification, the experimental simulations were found to improve the average MAE power forecast skill at the 95% confidence level for the first 7 forecast hours in the NSA, and through forecast hour 03 in the SSA. MAE power forecast skill improvement in the first 6 forecast hours ranged from 8% to 3% in the NSA, and from 6% to 1% in the SSA. Although the NAM DD simulations were only run for two episodes (December and January) the results are fully consistent with the findings from the RAP model over the larger data set. The forecast skill improvement due to assimilation of the new WFIP observations was also found to be dependent on the location of the verifying site. Verifying tower sites that were on the periphery of the NSA and SSA domains had smaller improvements than those located within the core observing network area, demonstrating the increased benefit of having more observations spread over a larger geographic area.

The degree of spatial averaging of the forecasts and observations before they are compared is found to have a profound impact on the skill of the forecast, with the power MAE decreasing by more than a factor of 2 as the spatial averaging extends to the full study area domain. This demonstrates the advantage to utilities and grid operators of having spatially distributed generation, not only because it provides less variability in generation, but also because the generation that is produced can be better forecast. Surprisingly, the impact of assimilation of the new WFIP observations measured as a percent improvement stays constant or even increases with the degree of spatial averaging, up to domains on the order of 400 km x 600km, indicating that even from a balancing authority's point of view, there is significant value to be gained from deploying and assimilating new observations.

A wind ramp tool and metric was developed for WFIP, and used to evaluate the skill of the RAP model at forecasting ramp events. Assimilation of the WFIP observations was found to improve the ramp forecast skill, averaged over the first 9 forecast hours, by more than 10% in the NSA, and by 3.5% in the SSA.

Reasons for the greater impact of the special WFIP observations in the NSA than in the SSA are, first, the NSA had more tall tower observations, more wind profiler observations, and the addition of nacelle anemometer observations; the greater numbers of observations is likely to have contributed to the greater improvement in both conventional MAE and ramp forecast skill. Second, the new observations were spread over a wider geographic area in the NSA than in the SSA, allowing for the model initial field improvements to be more robust and affect a wider area, thereby having a more lasting positive impact before advecting out of the study area.

Among the key successes of WFIP is that it has demonstrated that even in this era when large quantities of satellite and other data are routinely incorporated into operational weather forecast models, wind power forecasts still can be improved substantially through the assimilation of additional new observations focused within the atmospheric boundary layer. WFIP has also shown that the magnitude of the improvement increases with the number of observations, as well as the area that they are spread over. Further, the impact of the new observations is even larger for wind ramp events, which are important for grid operators. The improvements in forecast skill found in WFIP are significant compared to the year-to-year improvements of a few percent that research and operational forecasting centers typically find for low-level wind forecasts. Also as a result of WFIP, large quantities of proprietary hub-height wind speed observations were made available to NOAA, a type of observation that NOAA historically has had very limited access to. One of the legacies of WFIP is that those observations will continue to be sent to NOAA indefinitely, assimilated into NOAA weather forecasting models, and used to evaluate NOAA models. Finally, significant improvements in wind power forecasting were also found during WFIP by using improved forecasting models. Since prior to WFIP improving hub-height winds had not been a focal point for NOAA forecasting research, this suggests that we may have just begun to scratch the surface, and further large improvements are yet likely to occur as a result of future wind energy-focused research programs.

1. Project Overview

Wind power is a variable power source, dependent on weather conditions. Electric grid operators keep the grid stable by balancing the variable amount of power produced from wind plants by increasing or decreasing power production from conventional generation stations, including coal and natural gas. Having accurate advance knowledge of when wind power will ramp up or down through accurate weather forecasts can lead to improvements in the efficiency of operation of these fossil fuel plants, as well as the entire electrical grid system, resulting in lower costs as well as lower CO₂ emissions. Lowering the costs of integrating wind energy onto the grid can accelerate the development of wind energy as a growing component of the nation's energy portfolio.

Private sector forecasting companies rely on NOAA's operational weather forecasting models to provide the foundational wind and temperature forecasts that they use to make power forecasts for the energy industry. In some cases these company's products consist of statistical post-processing techniques applied to remove biases and reduce errors in NOAA's wind forecasts, and in other cases the companies use NOAA's forecasts to provide the initial and boundary conditions for computer forecast models that the companies themselves run over smaller regional domains. In either case, improvements in the accuracy of NOAA's wind forecasts will result in more skillful power prediction products that private forecasting companies provide to the energy industry.

1.1 Goals and Key Tasks

WFIP was a DOE sponsored research project whose overarching goals were to improve the accuracy of wind energy forecasts, and to demonstrate the economic value of these improvements. WFIP participants included several DOE national laboratories (National Renewable Energy Laboratory/NREL, Argonne National Laboratory/ANL, Pacific Northwest National Laboratory/PNNL, and Lawrence Livermore National Laboratory/LLNL); two NOAA research laboratories (Earth Systems Research Laboratory/ESRL and the Air Resources Laboratory/ARL); the NOAA National Weather Service/NWS; and two teams of partners from the private sector and university communities, led by AWS Truepower and WindLogics.

Prior to WFIP, NOAA did not have a focused program to improve its foundational wind forecasts for the wind energy industry. WFIP offered the opportunity for NOAA to jump-start its efforts at improving forecast model skill for this industry, as well as the opportunity to work directly with experts in wind energy, thereby allowing NOAA to gain insights into the ways that NOAA models are used and a better understanding of the wind energy-specific problems that exist in NOAA models. It also offered the WFIP private sector partners (WindLogics inc., and AWS Truepower) the opportunity to advance their own forecasting capabilities either through use of the improved NOAA forecasts or through their own forecasting systems.

Three final reports have been written on WFIP. This report provides an overview of the entire project, including the roles and tasks of the two private sector partners, and then focuses on the research done within NOAA. The other two reports, written by teams led by WindLogics and AWS Truepower, provide detailed analyses of the impact of WFIP from the perspective of private forecasting companies and electric grid balancing authorities.

WFIP considered two avenues for improving wind energy forecasts. The first was to deploy networks of mainly remote sensing observations while also acquiring proprietary meteorological observations from the wind energy industry, and for the first time assimilating these data into numerical weather prediction (NWP) models. Additional observations allow for a more precise depiction of the model's initial state of the atmosphere, potentially resulting in more accurate forecasts. The intent of the WFIP instrumentation networks were to provide observations focused on the atmospheric boundary layer and above, and over a sufficiently broad area, to influence NWP forecasts out to at least 6 hours lead time. These observations were collected over a full year, to allow for an evaluation of seasonal differences in the skill of the models and the impact of the observations.

A necessary key component of WFIP was to develop improved quality control procedures to ensure that the assimilated observations were as accurate as possible, as a few erroneous observations can easily negate the positive impact of many accurate observations when assimilated into a NWP model. Instrumentation and data quality control are discussed in detail in Section 2. NOAA was responsible for managing the integration of the observational data (most of it arriving and used in real-time) from NOAA, the DOE labs, and the industry/university partners, and was responsible for data archival.

The second avenue for improving wind energy forecasts was to improve the NWP models directly. Midway through the WFIP field program, NOAA/NWS upgraded its operational hourly-updated NWP forecast model from the Rapid Update Cycle (RUC) model to the Rapid Refresh (RAP) model. The WFIP observations allowed for a quantitative determination of the improvement gained in hub-height wind speed forecasts with this model upgrade. Also, NOAA/ESRL was (and continues to be) in the process of improving the RAP model, and also developing a higher resolution version of the RAP model, called the High Resolution Rapid Refresh (HRRR) model. The WFIP observations allowed for a determination of the skill of the HRRR at forecasting hub-height winds through broad regions of the Midwest, and allowed the NWS to evaluate its NAM model skill at forecasting hub-height winds for the first time. Due to the great volumes of data generated by the HRRR, raw model output from it was not publically available prior to WFIP. One of the goals of WFIP was to obtain the computer infrastructure to make this data available in real-time to both the two private sector teams, as well as to any other party on the wind energy industry

Beyond evaluating the skill of the NOAA models at forecasting hub-height winds, the WFIP observations also made it possible to determine shortcomings in the model's physical parameterization schemes (e.g., turbulence mixing), thereby making possible fundamental improvements in the model physics. The NOAA models and their improvements are discussed in Section 3, and their data assimilation

systems are described in Section 4. Evaluation of the real-time model forecasting models is discussed in Section 5.

A motivation for the instrumentation deployments and data impact study were the investigations of Benjamin et al (2004a, 2010), which analyzed forecast improvements to wind speed and other parameters through data denial experiments using NOAA operational observing systems such as radiosondes, aircraft, radar wind profilers, and surface mesonet. The precise determination of the impact of assimilation of the special WFIP observations came from similar carefully controlled data denial experiments, in which identical versions of the RAP model were run with and without the fully quality controlled observations. Six data denial episodes were run, each 7 to 12 days long, that spanned all four seasons of the year. In addition to the RAP, data denial experiments were also conducted with the NAM for the two winter episodes. The data denial simulations, as well as model biases, are discussed in Section 6.

The bulk of the statistical analysis performed for WFIP was done comparing forecasts to observations at individual observation locations, and then averaging the statistics from the individual locations into an overall statistic. These statistics are appropriate if one is interested in the skill of making a point forecast, for example the skill in forecasting for an individual wind plant that fits within a single model grid cell. For some applications one would instead be interested in comparing spatially averaged power generation with spatially averaged forecast power; for example if a number of dispersed wind plants were feeding power into a transmission line, and the overall power flowing through that transmission line is the quantity of interest. Spatially averaged forecast skill can differ from the average skill of individual point locations if the point locations have compensating errors, where an over-forecast at one point balances an under-forecast at another point. For this reason in Section 6 we also investigate how model forecast skill varies with geographic spatial averaging.

One of the features of energy production from a wind turbine is that the power output typically has long periods of time with either zero power production (for speeds below the turbine's cut-in speed) or near 100% of its maximum capacity production for high speeds. The wind power production frequently jumps rapidly between these two extremes of near zero or near 100% power, and these jumps, referred to as ramp events, can be very rapid due to the wind power increasing approximately as the cube of the wind speed in the middle portion of the turbine's power curve (an example of a power curve can be found in Fig. 5.2). Recognizing the importance of these ramps events for grid operation, and that standard statistical metrics may not adequately measure the skill of NWP models at forecasting these important events, one of the goals of WFIP was to develop a new metric for ramp events. Section 7 describes a ramp tool and metric that was developed for WFIP and applied to the WFIP forecasting results.

As part of WFIP a physical process study was carried out to investigate the relationship of hub-height winds on surface heat and momentum fluxes, and to evaluate the applicability of flux-dependent wind

profile laws at replicating the wind profiler through the wind turbine rotor layer. Results of this analysis are presented in Section 8.

The final major component of the WFIP analysis is an evaluation of the economic benefits that would have accrued from the improved accuracy of the WFIP wind power forecasts had they been used by grid operators. Initial analyses were done by the two private sector partners and their collaborators, using models of the electric grid system. DOE has decided to undertake additional studies to explore the complex interactions between wind forecasting and power system operations prior to publication of these results. The initial work performed by the WFIP teams provided important insight into the benefits and shortcomings of various power system assumptions, market designs, and modeling tools in identifying costs and savings. The desire to explore these important issues in more detail is the impetus for the new analysis. Over the next year (2014-2015), DOE plans to engage with industry experts, grid operators and economic modelers to accurately define methodologies that provide quantification of total financial savings and other ancillary benefits of improved short-term wind power production forecasts.

In summary, the core tasks of WFIP are to:

- Disseminate the HRRR model output to the wind energy industry, including WFIP private sector partners.
- Determine NOAA and private sector model skill at forecasting hub heights winds in diverse regions of the U.S. Midwest.
- Improve the foundational operational and research NOAA forecast models.
- Increase the number of atmospheric observations in the two study area domains.
- Develop new quality control algorithms for radar wind profiler and tall tower observations.
- Develop the capability to ingest and assimilate industry-provided tall tower and nacelle mounted anemometer observations into NOAA models.
- Improve the initialization of NOAA and industry atmospheric mesoscale models.
- Increase the accuracy of predicted wind speed and direction changes in short-term (0-6 hr) forecasts.
- Determine the impact of assimilation of new WFIP observations (tall towers, nacelle anemometer winds, sodars, and wind profiling radars).

- Develop a ramp tool that can be used to quantify model skill at forecasting wind ramp events.
- Create working relationships between NOAA and the wind energy industry that provide a two-way flow of information, thereby accelerating improvements in wind energy forecasting.
- Inform NOAA on the value of networks of boundary layer wind profiling instrumentation.
- Provide critical analyses of the strengths and weaknesses of NOAA and private sector forecasting models, potentially leading to improved model physical parameterizations.
- Investigate the ability of standard flux-profile relationships at characterizing the wind profile through the turbine rotor layer.
- Quantify the economic impact of improved wind power forecasts.
- Disseminate project results to the wind energy community, contributing to a continuous improvement in state-of-the-art of short-term forecasting methods.

1.2 Team Partners, Two Study Areas

DOE selected two teams from the private sector to collaborate with DOE and NOAA. The first team was led by AWS Truepower, and included MESO Inc., Texas Tech University, the University of Oklahoma, North Carolina State University, ICF Inc., DOE/NREL, and the Energy Reliability Council of Texas (ERCOT). The second team was led by WindLogics and included South Dakota State University, DOE/NREL, and the Midwest Independent System Operator (MISO). One of the differences between the two study partners is that WindLogics relies solely on NOAA forecast models to make its forecasts, applying machine-learning post-processing algorithms to improve upon the raw forecasts, while AWS Truepower runs local region mesoscale NWP models that are initialized from NOAA forecast models.

The geographical study areas proposed by these two teams were the upper Midwest (WindLogics) and western Texas (AWS Truepower). The locations of the two study areas and the instrumentation deployed or made available in each area are shown in Fig. 1.1 and Fig. 1.2, and the types and numbers of meteorological observing instruments in each area are also listed in Table 1.1. Comparing the two model domains, the Northern Study Area (NSA) domain is larger, and has a more even distribution of wind plants, whereas the Southern Study Area (SSA) domain is smaller, and has a much more concentrated distribution of wind plants near the center of the domain. Also, greater topographic variation exists in the SSA than the NSA, which can affect the relative skill of NWP forecasts in the two areas. The field campaign portion of WFIP ran from August 2011 until early September 2012, although the tall tower network data in the SSA did not become available until November 29, 2011, which resulted in a slightly shorter analysis period for hub-height winds in the SSA.

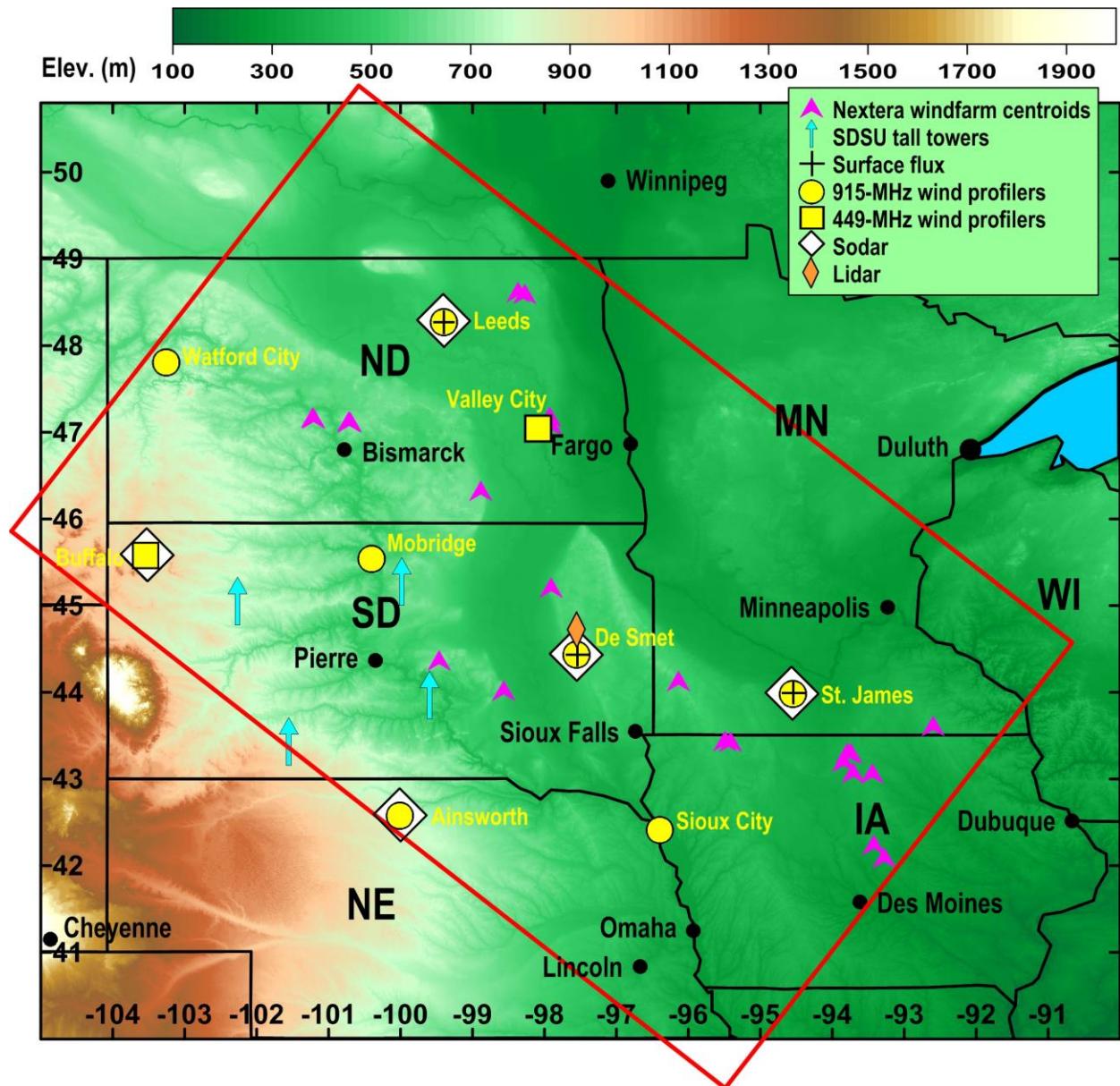


Figure 1.1. Geographic domain of the Northern Study Area. Surface elevation is shown by color shading. Instrument types and locations are shown, as well as the locations of the Next Era wind farms.

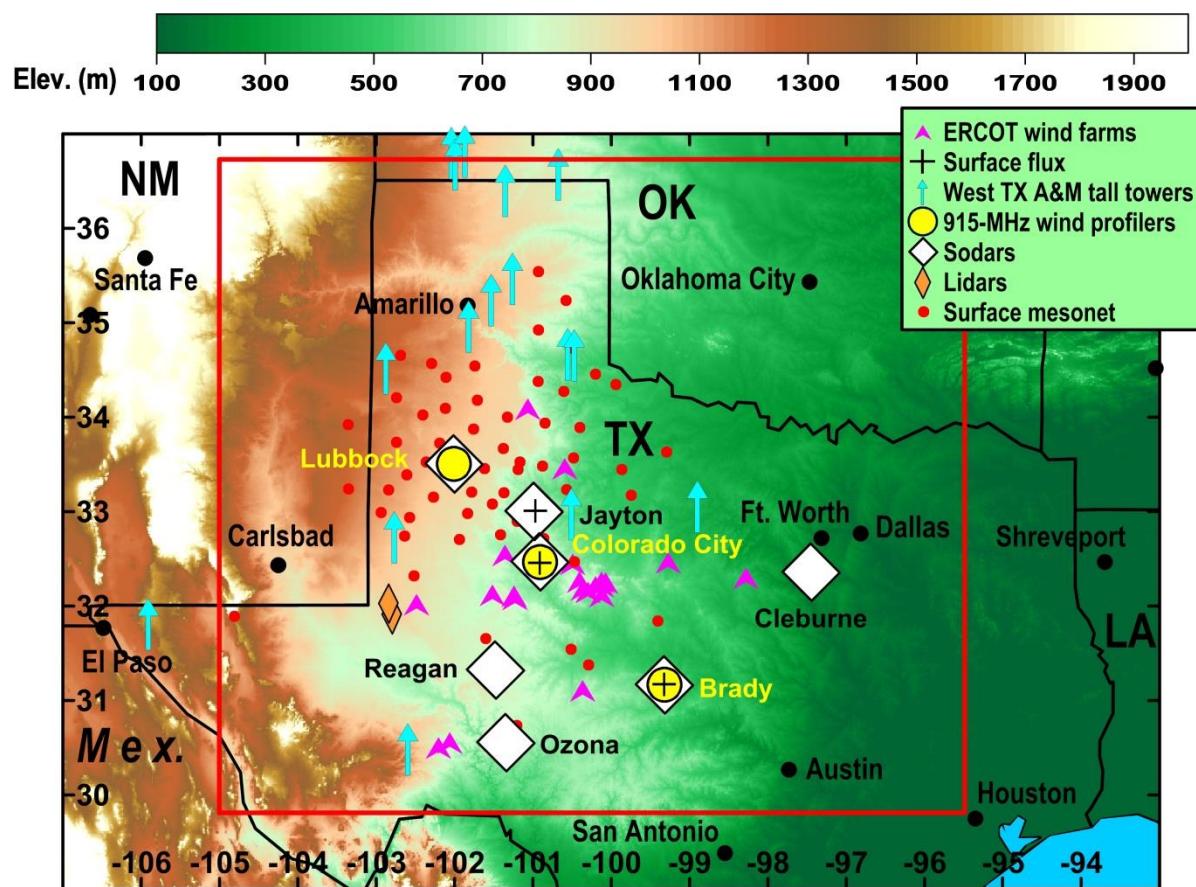


Figure 1.2. Geographic domain of the Southern Study Area. Surface elevation is shown by color shading. Instrument types and locations are shown, as well as the locations of wind farms providing power to ERCOT.

Instrument	NSA	SSA
915 MHz W-P Radar	7	3
449 MHz W-P Radar	2	
Doppler Sodar	5	7
W-P Lidar	1	2 (short term)
Surface Flux Station	3	3
Surface Met Station	8	63
Tall Towers	133	51
Nacelle winds	411	

Table 1.1 The types and numbers of meteorological observing instruments deployed in the two study domains. W-P indicates a vertical wind profiling capability.

2. WFIP Observations

2.1 Instrumentation

A suite of different type of atmospheric observing systems from NOAA, DOE national laboratories, and the private sector was assembled for use in WFIP, listed in Table 2.1. These observing systems served two purposes. First, they provided observations on the current state of the atmosphere (i.e., weather conditions) that were assimilated into the NWP models used to make wind power forecasts. Forecasting wind power is an initial value problem, and the better that one can specify the initial state of the atmosphere the more accurate of a forecast can be made. The second purpose of the observations was to validate the NWP models. These validations answer the question that if new observations were collected near a wind farm, would the forecasts be more accurate in the vicinity of those same wind farms? NOAA has restricted its validation analysis to use of these atmospheric observing systems. The two WFIP private sector partners, WindLogics and AWS Truepower, also performed validation studies using actual wind plant power production data. In this section we describe the instruments, data QC, geographical deployment of the instruments, data transmission protocols, and inter-comparisons of the different instrument types.

Instrument	NOAA/ ESRL	NOAA/ ARL	DOE/ PNNL	DOE/ ANL	DOE/ LLNL	WindLogics	AWS Truepower	NRG- Leosphere	West Texas A&M	Iberdrola
915 MHz W-P Radar	6	1		2 (1-STI)			1 (TTU)			
449 MHz W-P Radar	2									
Doppler Sodar		3		3		2	4			
W-P Lidar					1			2 (short term)		
Surface Flux Station		3		3						
Surface Met Station	8	1	6						56	
Tall Towers						118	35		15	15
Nacelle winds						411				

Table 2.1. List of instrument types, numbers, and providers, deployed or made available for WFIP. One wind profiling (W-P) radar deployed by DOE/ANL was leased from Sonoma Technology Inc. (STI). The AWS Truepower wind profiling radar was owned and operated by Texas Tech University.

Wind Profiling Radars with RASS

A network of 12 wind profiling radars (WPR's) was assembled for WFIP. These also included Radio Acoustic Sounding Systems (RASS) components for measuring temperature profiles. Two different types of WPR's were used. The first uses 915 MHz frequency microwaves (33 cm wavelength) while the second uses 449 MHz (67 cm wavelength) microwaves. The 915 MHz systems (Fig. 2.1) are frequently referred to as "boundary layer profilers" and have a typical lowest range gate near 100m, with a maximum detectable signal that varies with atmospheric conditions (higher in a moist atmosphere) but that typically ranges from 1.5 to 4 km above ground level. Typically two vertical sampling modes are interlaced in time, a 60 m high resolution mode and a coarser resolution 100m mode. Figure 2.2 displays a 24 hour time-height cross section of data from the Brady Texas 915 MHz WPR, showing the sudden onset (at 01 UTC; 18 CST) and cessation (between 14-16 UTC; 08-10 CST) of a low-level jet and its vertical structure, where UTC is the Universal Time Coordinate, or Greenwich Mean Time, and CST is Central Standard Time. The depth of the atmosphere that the WPR was able to observe on this day was approximately 3km above ground level (AGL).



Figure 2.1. The WFIP 915 MHz wind profiling radar at Saint James, MN. The center enclosure contains the transmitter, a phased-array antennae, and a clutter-suppression screen. The four white rectangular boxes are the Radio Acoustic Sounding System (RASS) loud-speakers. The gray enclosure to the right of the wind profiler is a co-located sodar system, and a 10m mast with surface met instrumentation is visible between the profiler and sodar. The portable trailer contains the computer data acquisition system and communication equipment.

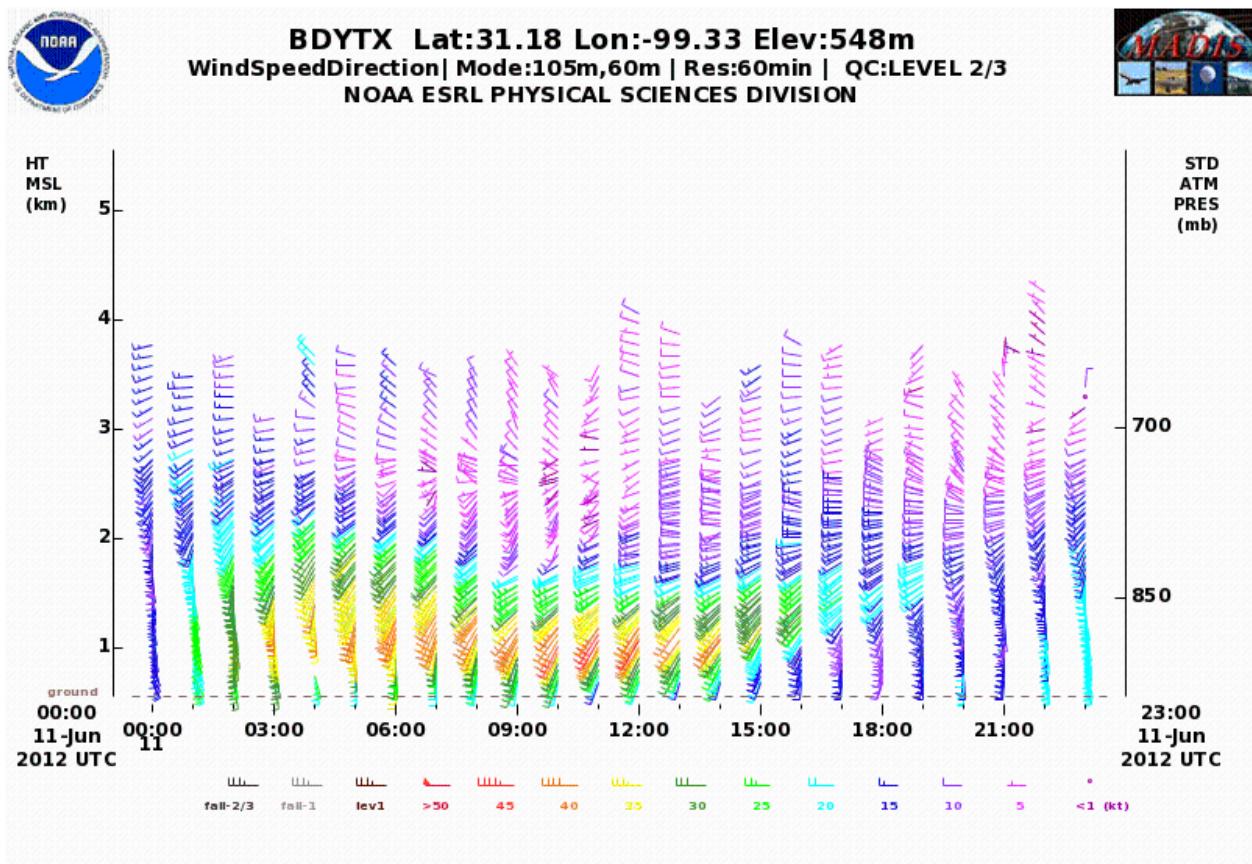


Figure 2.2. 24 hour time-height cross-section of hourly averaged winds from the 915 MHz Brady Texas Wind Profiling Radar. The onset of a nocturnal low-level jet occurs at 01 UTC (18 CST), and ends between 14-16 UTC (08-10 CST) the next morning.

The second type of WPR's are the 449 MHz systems (Fig. 2.3), which not only have a lower frequency, but also more powerful transmitters. Both of these facts allow the 449 MHz WPR to observe a deeper layer of the atmosphere, often to 7 km AGL. Figure 2.4 displays a 24 h time-height cross-section from the Buffalo ND, 449 MHz WPR. The radar observes winds through the lowest approximately 6 km of the atmosphere on this day, and indicates two upper level wind maxima (red wind bars) near 6 km above mean sea level (MSL) at the beginning and end of the day that would not have been observed with the 915 MHz systems.



Figure 2.3. The WFIP 449 MHz wind profiling radar located at Buffalo, ND. The four white enclosures house the RASS loudspeakers, and the small portable building contains the computer data acquisition system and communication equipment.

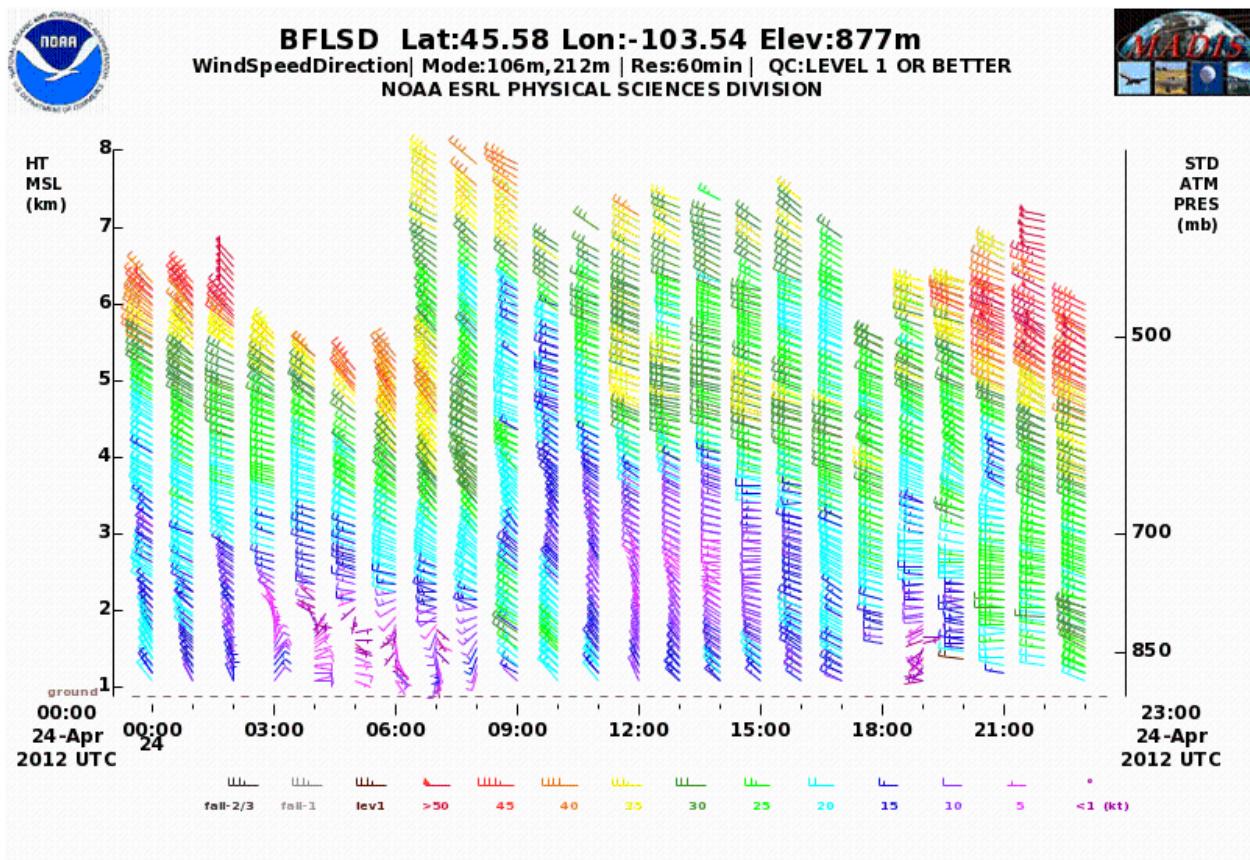


Figure 2.4. 24 hour time-height cross-section of hourly averaged winds from the 449 MHz Buffalo ND Wind Profiling Radar. Two upper-level wind maxima are observed, one between 00-07 UTC (18-01 CST), and another between 20-23 UTC (14-17 CST).

Both the 915 and 449 MHz wind profiling radars generally came equipped with RASS. RASS measures the virtual temperature (the temperature that a completely dry parcel of air would have if it had the same density and pressure as a parcel of moist air) by emitting a vertically propagating acoustic signal from a loudspeaker near the side of the radar antennae, and tracking the speed of the acoustic signal with the Doppler radar beam. Since the speed of sound depends on the temperature of the air, the vertical profile of virtual temperature can be measured. A time-height profile of the virtual temperature from the Buffalo WPR is shown in Fig. 2.5. The height coverage of RASS for the 449 MHz systems was typically 1.0 km, and 0.6 km for the 915 MHz systems. RASS temperatures were measured and averaged over the last 5 minute period of each hour.

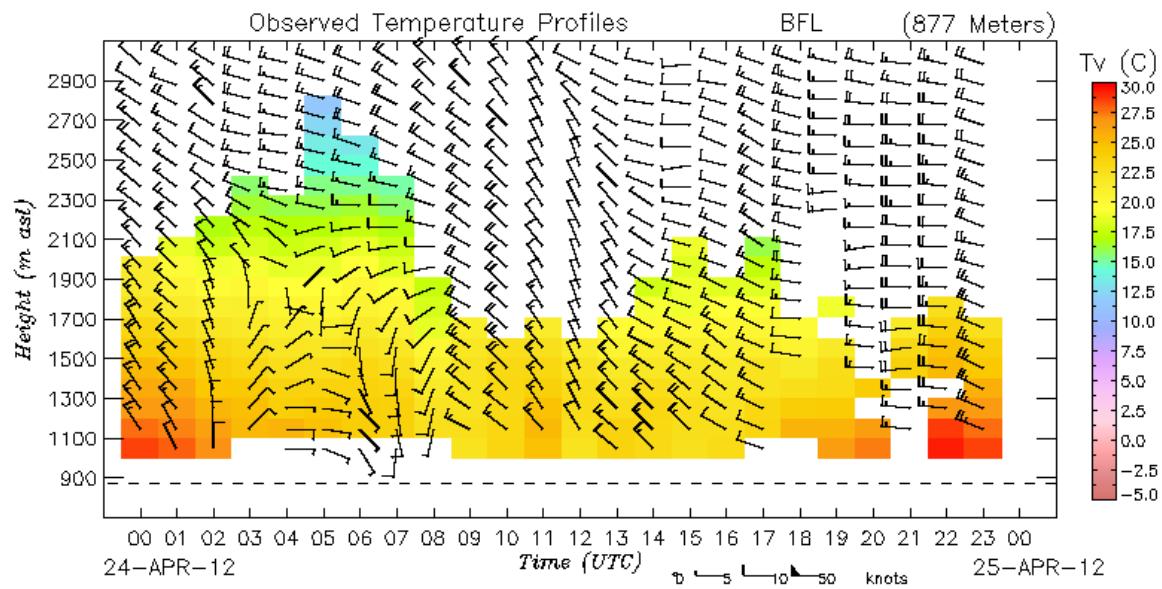


Figure 2.5. 24 hour of hourly-sampled RASS virtual temperature (color contours) and wind barb time-height cross-section from the 449 MHz Buffalo ND Wind Profiling Radar.

Sodars

A network of 12 Doppler sodars was also assembled for WFIP. These sodars, although of different ages and manufacturers, all had similar performance characteristics, providing wind speeds to a maximum height of 200 m AGL with either 5 or 10 m vertical resolution. Fig. 2.6 displays a 24-hour time-height cross section of winds from the Reagan TX sodar. This data shows the development of a low-level jet during hours 01-09 UTC (19-03 CST), and a strong wind ramp event between hours 16-17 UTC (10-11 CST).

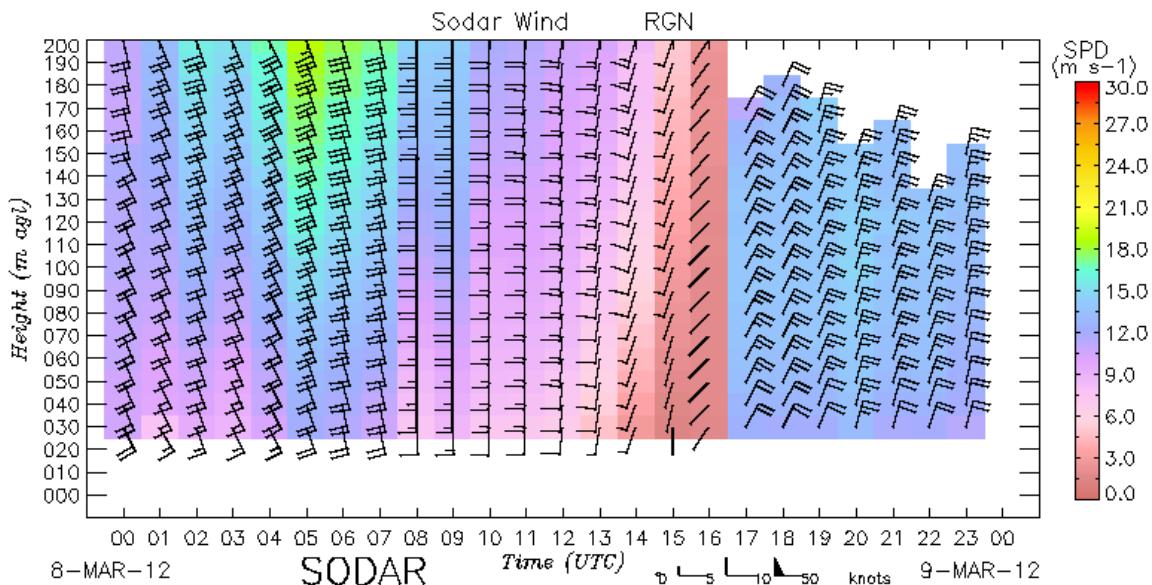


Figure 2.6. 24 hour wind time-height cross-section from the Reagan TX sodar. Colors indicate the wind speed, bars the vector wind.

Lidars

Three lidar systems were available during at least parts of the WFIP field campaign. The first lidar system was a Leosphere WindCube7 system provided by DOE/PNNL, which was intended to be deployed for the entire year-long field campaign. During the later stages of the field campaign, NRG-Leosphere also offered to donate two other lidar systems for a shorter duration campaign in the last several months of WFIP. The DOE/PNNL lidar and one of the donated Leosphere lidars had similar performance characteristics, providing hourly averaged winds from 40 to a maximum of 200m with 20 m vertical resolution (Fig. 2.7). The remaining donated lidar system was a WindCube8 system, which provided 10-min averaged winds from 40m to a maximum of 460m with 20 m vertical resolution (Fig. 2.8).

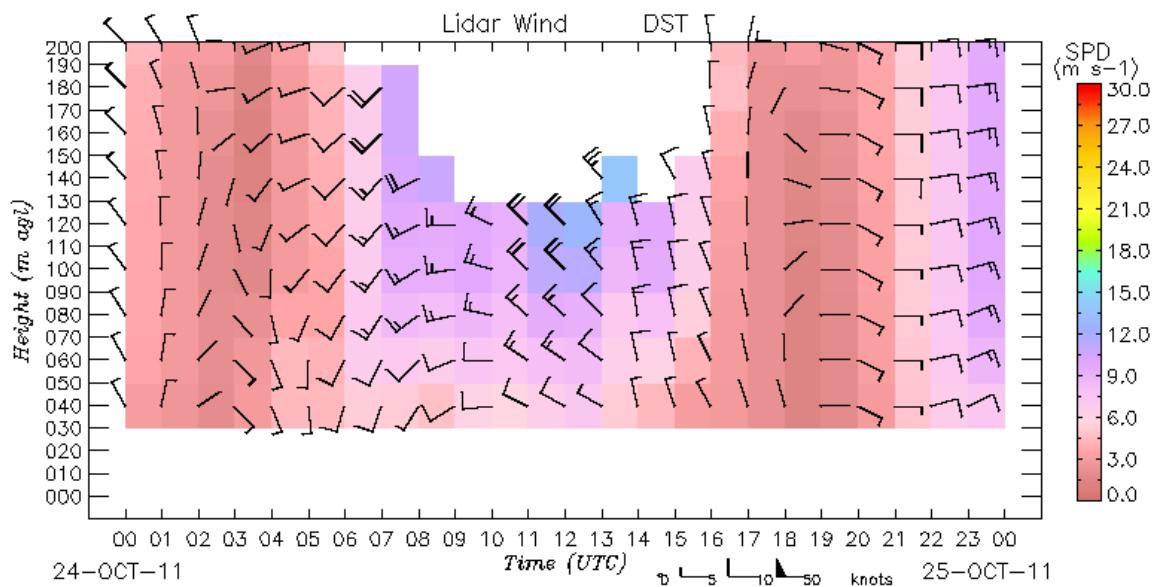


Figure 2.7. 24 hour wind time-height cross-section from the DOE/PNNL lidar deployed at DeSmet SD. Colors indicate the wind speed, barbs the vector wind.

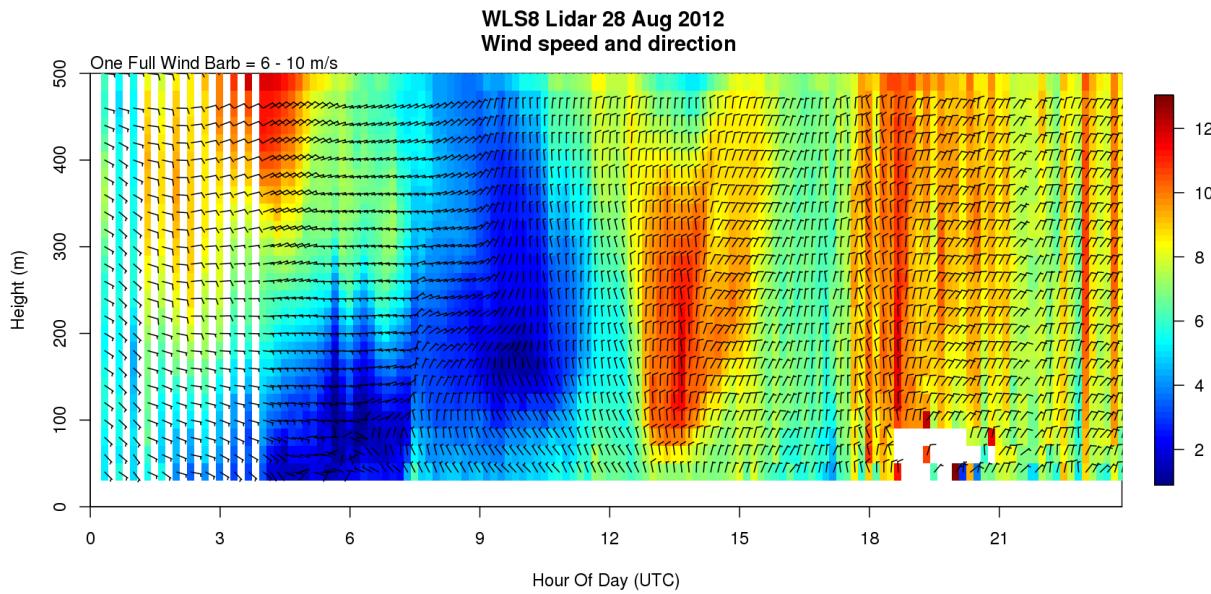


Figure 2.8. 24 hour wind time-height cross-section from the Leosphere WindCube8 lidar deployed in the southwestern portion of the SSA domain. Colors indicate the wind speed, barbs the vector wind.

Tall Tower Winds

One of the key instrumentation systems made available for WFIP were networks of anemometers mounted on tall towers. The tall towers were mostly deployed by private wind industry companies, and provided to NOAA as part of WFIP under Non-Disclosure Agreements. WindLogics in the NSA domain provided data from the greatest number of towers. Using WFIP funding they upgraded communications on 35 towers which then were able to provide data to NOAA in real-time for data assimilation. They also provided data from another 79 towers, but these data arrived one to two days late, and were used for assimilation only in the retrospective data denial experiments. Since the locations of these towers as well as the data from them are proprietary, we do not provide a map of tower locations. However, the geographic spread of the towers largely follows the distribution of NextEra wind farms shown in the basemap for the NSA (Fig. 1.1). WindLogics also contracted with South Dakota State University (SDSU) to provide real-time data from four tall towers that they operate in South Dakota, whose locations are shown in Fig. 1.1.

In the Southern Study Area, ERCOT provided reliable real-time data from 34 tall towers, which became available for use only on November 29, 2011. Again because the locations of these towers as well as the data from them are proprietary, we do not provide a map of tower locations. The geographic spread of the towers is much more concentrated than for the NSA, reflecting the fact that much of ERCOT's wind energy generation also comes from a very concentrated region in West Texas. In addition, Texas Tech University (TTU) provided observations from their 200m tower located near Lubbock, Texas. Also,

observations from a network of 15 tall towers operated by the West Texas A&M University were used, but were only available for the retrospective data denial simulations and not for the real-time forecasts.

A final set of tall tower observations was provided by Iberdrola USA, independent of WFIP, and part of a longer term data sharing agreement with NOAA. Data from 15 towers in the Midwestern U.S. were made available, and were used in the retrospective data denial simulations but not in real-time forecasts. Of these, 14 were located near the NSA domain, with the remaining one near the SSA domain.

Most of the tall tower data (with the exception of the TTU 200m tower) provided one or more levels of observations between 40 and 60m, with a few towers providing data up to 80m. When multiple levels were present, all levels were used for assimilation and evaluation purposes. Using all tower levels available, approximately 235-250 reliable independent wind measurements were used at any given hour for either assimilation or evaluation purposes. All of the tall tower observations were provided as 10 minute averages, with the exception of the ERCOT network which came as 15 min averages.

Nacelle anemometers

In the NSA, WindLogics provided nacelle anemometer winds from 411 wind turbines at 23 different wind farms. These data were a small subset of the total number of wind turbine nacelle anemometers existent, and were selected by WindLogics to provide an adequate sample of the turbine winds across the entire set of wind farms providing data. WindLogics developed and applied wind speed and wind direction corrections to nacelle anemometer data. These corrections accounted for blade wash from the turbine on which the anemometer was mounted, but not wake effects from multiple upwind turbines.

Surface mesonet

In normal operations, the NOAA/ESRL RAP and HRRR models only assimilate surface mesonet observations from the NOAA/NWS ASOS network. Although many other public and private networks exist, the data quality often is sufficiently unreliable as to lead to degraded forecast accuracy if these surface mesonet data are assimilated. The NAM, however, does use these surface mesonet observations but only through judicious use of station reject lists.

For WFIP two networks of surface mesonet data were utilized by the NOAA/ESRL RAP model. The first were 6 stations deployed by DOE/PNNL in the SSA specifically for WFIP. The second network was from West Texas A&M University, which operates a surface mesonet of 56 stations that overlaps the SSA. Data from both networks was used for evaluation and for assimilation in the retrospective data denial experiments with the RAP, after additional QC was applied as discussed below in section 2.3.

2.2 Site Selection and Preparation, Leases, Data Transmission and Handling

Site selection and leases for all of the wind profiling radars was tasked to NOAA, with the exception of the DOE/ANL wind profiler deployed at Sioux City, Iowa, which was obtained by DOE/ANL, and the TTU wind profiler that was already running near Lubbock, Texas. Several scouting trips were required for each of the two study domains to find appropriate sites. NOAA/ESRL has permission to obtain expedited site leases, if the site is a government owned property (federal, state, or local county, city or municipality) and a no-cost lease can be agreed to. Typical sites include small airports, Forest Service or Bureau of Land Management property, water treatment plants, or road maintenance facilities. In most cases it took 6 to 9 months from the time the site was selected to obtain the required legal signatures for WFIP leases. NOAA also obtained electrical power and security fencing (if necessary) for each of these sites.

In the NSA, all of the sodars and the lidar were co-located with one of the wind profiling radars. WindLogics had to obtain their own separate lease for their two sodars with the same government entities that had agreed to the NOAA/ESRL leases. In the SSA, three sodars were co-located with NOAA sites, and AWS Truepower was responsible for obtaining leases for the remaining three sodars.

Site selection in the NSA was based on the concept of evenly sampling the study area domain with a WPR separation of approximately 200 km. In the SSA, site selection for both of the WPR's, the 3 remaining sodars, and the surface met stations was guided by an AWS Truepower correlation-based model study.

NOAA WPR data was transmitted from each site in real-time to NOAA/ESRL, normally using cell-phone communication. After arrival at NOAA/ESRL, it was then run through several automated QC algorithms (see Section 2.3 below) and then placed onto the NOAA Meteorological Assimilation Data Ingest System (MADIS) data repository and assimilated into the NOAA/ESRL RAP and HRRR models. Other data sets were obtained by the instrument owners (DOE labs, private sector partners), also typically using cell-phone communication, and then transmitted to NOAA/ESRL using the internet.

Several steps were taken to ensure the safety of proprietary data proved to NOAA. First all of the data was stored on a dedicated WFIP server behind NOAA's standard firewall. Second, this server has local access controls including Transmission Control Protocol (TCP) wrappers and a local host-based firewall. All traffic to/from this server is restricted by source/destination static Internet Protocol (IP) addresses and port numbers corresponding to the data providers. File transfer protocols for data ingest was limited to Secure Copy (SCP) or File Transfer Protocol-Secure Sockets Layer (FTPS), to ensure the confidentiality of data in transit. In cases where data was being pushed from data providers to NOAA, each data provider also had a username and password that gave them the ability to log in to their home directory and write files.

2.3 Data Quality Control and Instrument Performance

Wind profiling radars

Quality control of atmospheric observations is crucial if the data are to be assimilated into numerical weather prediction models, as the degradation of forecast skill from the assimilation of only a few bad data points can outweigh the benefit of assimilation of many good data points. In particular, radar wind profilers are known to sometimes suffer from large measurement errors due to a variety of causes, including migrating birds, ground clutter, and radio frequency interference. For this reason, a major effort was undertaken to improve the quality of the radar wind profiler data prior to real-time data assimilation.

Contamination of radar wind profiler data from nocturnal migrating birds was identified and quantified by Wilczak et al. (1995). Although techniques have been developed that helped reduce the level of contamination (Merritt, 1995), these were unable to completely remove the interference during periods of very dense bird migration. For this reason, data from operational wind profiling radar networks, such as the NOAA National Profiler Network apply additional simple quality control procedures to eliminate bird contaminated data based on time of day, wind direction, season, Signal-to-Noise Ratio (SNR), and spectral width thresholds. This procedure effectively eliminates all data that have characteristics of bird contamination, but can at times mistakenly flag and eliminate real atmospheric signal.

A more recent technique (Lehman, 2012) utilizes a Gabor frame expansion to identify periods with bird contamination. The discrete Gabor frame expansion is a method for decomposing wind profiler data simultaneously in time and frequency, which allows for a separation of the stationary and non-stationary signal components. A statistical filtering method can then be constructed to identify and remove the non-stationary, intermittent signal components from the data.

As part of WFIP, Bianco et al. (2013) investigated the ability of the Gabor frame technique to identify and then remove bird contamination from both 915 MHz and 449 MHz wind profiler data. A different tuning of the Gabor scheme was found to be required for the two frequencies of profilers, as well as additional levels of thresholding of the moment level data. In addition, a state-of-the-art Multi-Peak Picking (MPP) algorithm (Griesser and Richner, 1998) was implemented in parallel with the Gabor processing, followed by a moment level pattern recognition scheme (Weber et al., 1993). Almost all bird contamination in both the 915 and 449 MHz profilers was eliminated when the data were processed in this way, as shown in Fig. 2.9 using data collected and experimented with prior to the start of WFIP.

Because of weak signal strength during some of the winter months, data degradation due to ground clutter could occur in the 449 MHz profiler data, and from both ground clutter and radio frequency interference (RFI) from cell-phone communication with the 915 MHz profilers. A special WFIP real-time algorithm was developed and implemented midway through the field campaign for ground clutter interference (which bias the winds to lower wind speeds). This algorithm identified periods of strong ground clutter through comparison with the 10 m anemometer winds, measured by a prop-vane at each

of the wind profiler sites, with the lowest several levels of WPR data. In addition, a RFI check was implemented at two sites (DST and LDS) that had strong RFI. This check was based on the vertical continuity of the profiler wind speeds. The clutter and RFI QC checks are described in more detail in Appendix 1.

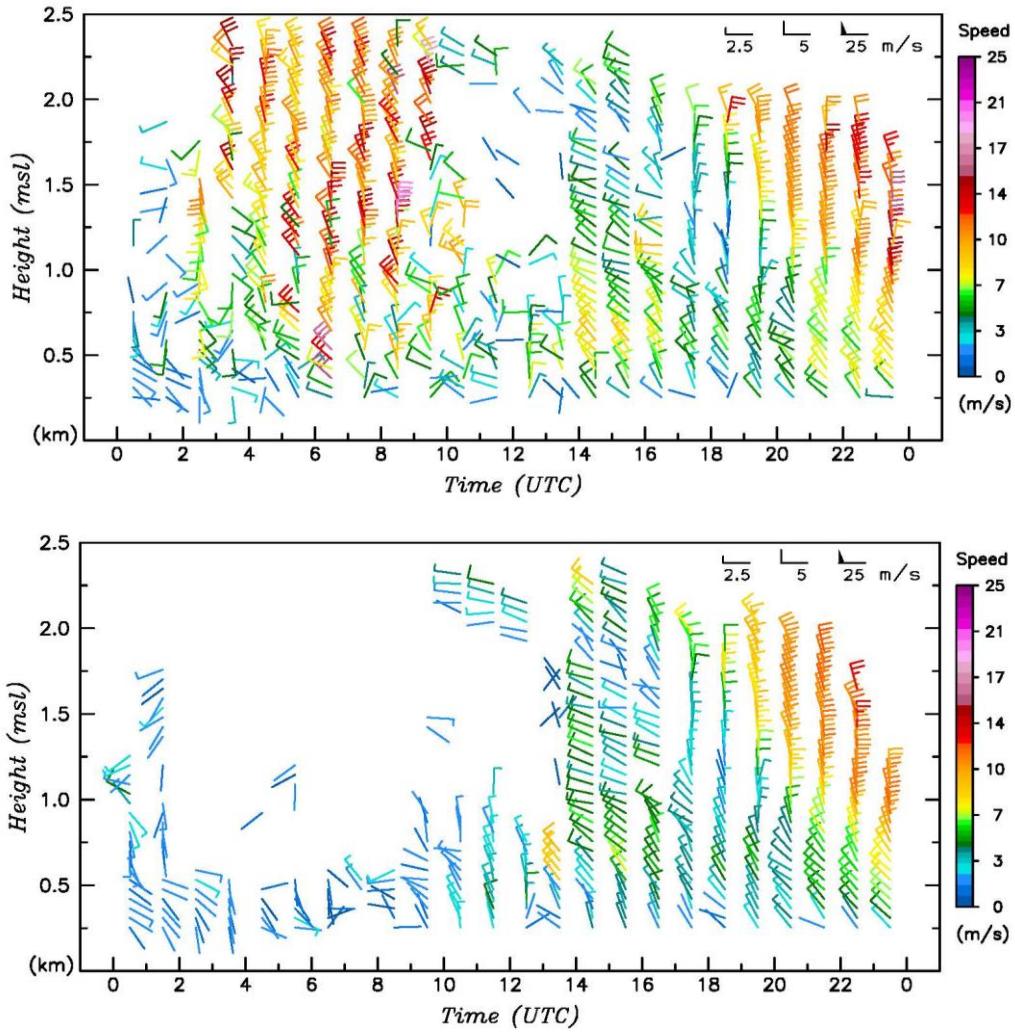
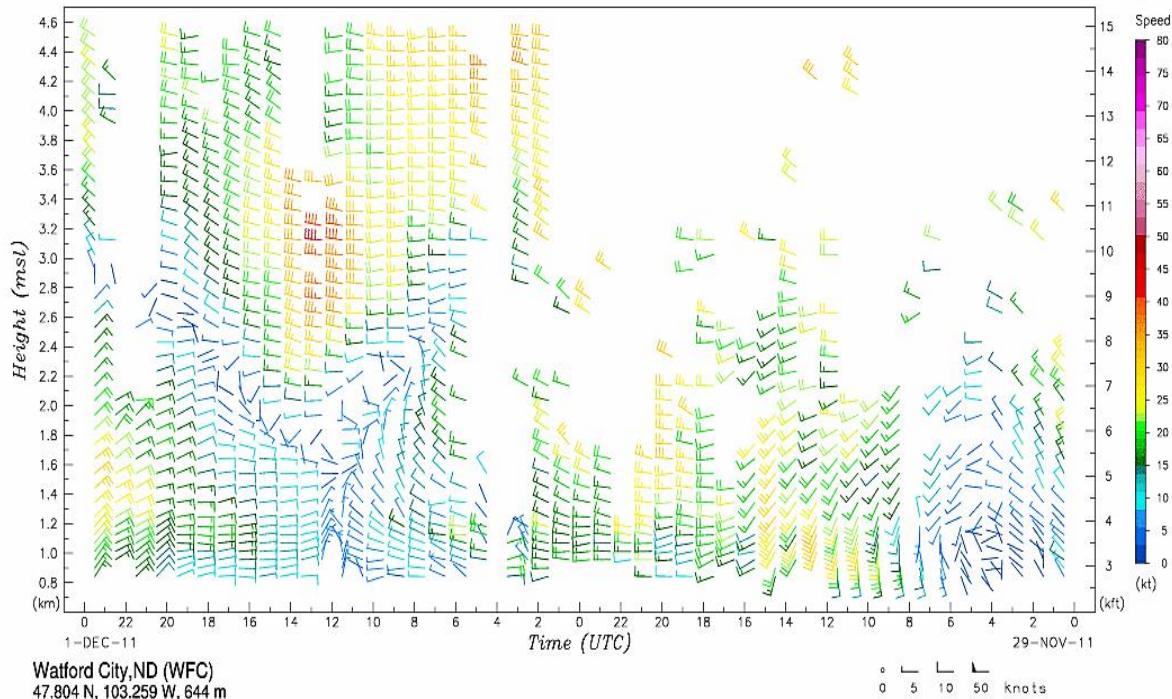


Fig. 2.9. Top panel) 915-MHz wind profiler time-height cross sections of hourly winds computed by a standard consensus procedure, using test data for the 9th of October 2010 at Chico, California. Periods with contamination from southward nocturnal migrating birds are apparent during hours 02-11 UTC. Bottom panel) The same data after processing by the combined Gabor, thresholding, and Weber-Wuertz pattern recognition scheme.

In addition to Gabor processing, WFIP funding also allowed for the implementation of real-time pattern-recognition signal processing (Weber and Wuertz, 1993). This processing first searches for patterns in the sub-hourly moment level radial wind components (typically measured by the radar every few minutes), eliminating outliers that do not fit the local pattern. A second level of pattern recognition is

then also implemented on the u, v, and w component winds after the hourly averaged winds have been computed. Patterns are searched within a time and height window that slides forward in time as new observations are acquired. The effects of the pattern recognition processing are shown in Fig. 2.10 for hourly averaged winds from the WFIP Watford City, ND, 915 MHz wind profiler, where the top panel shows winds produced by a standard consensus algorithm, while the lower panel shows winds after the Weber-Wuertz pattern recognition processing. The pattern recognition processing provides more good winds and erroneous values are removed, all while maintaining sharp gradients that occur in the atmosphere.



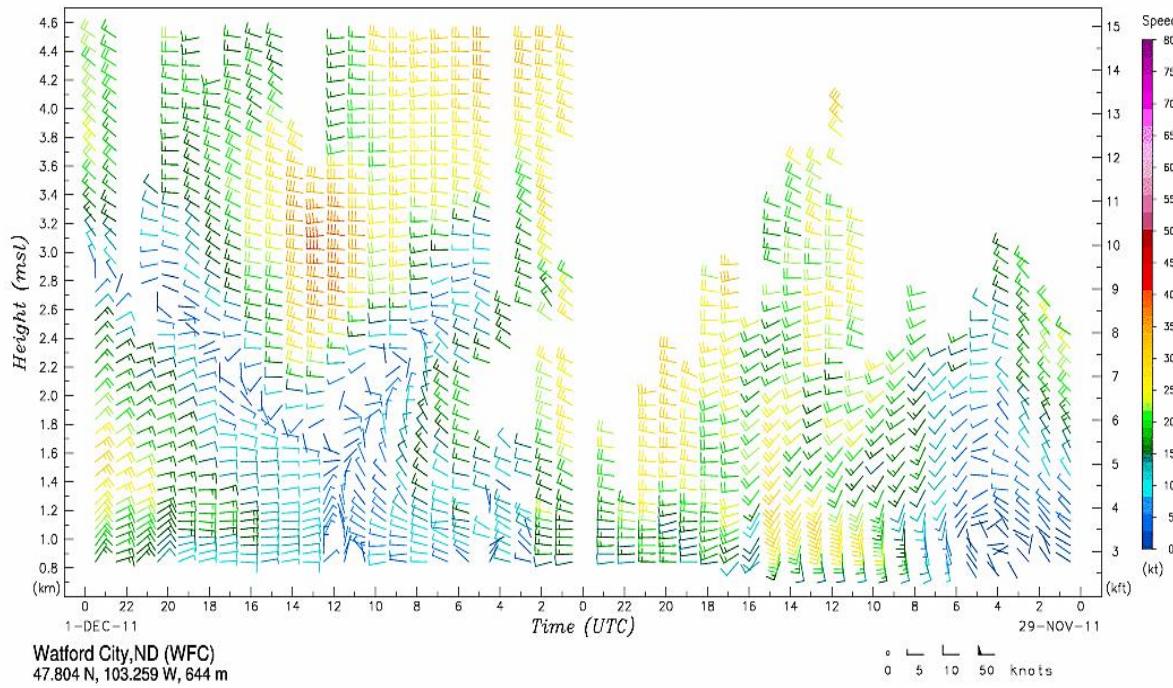


Figure 2.10. Hourly averaged winds from the WFIP Watford City, ND 915 MHz wind profiler. Top panel: winds produced by a standard consensus algorithm. Bottom panel: winds produced using the Weber-Wuertz pattern recognition processing algorithm.

RASS

The WPR RASS systems were sometimes also affected by cell-phone induced RFI, especially in the winter months when the atmospheric signal is weaker. An automated real-time algorithm was developed that identified RFI contaminated temperatures, based on the fact that these contaminated temperatures at all gates had almost identical temperatures. The algorithm is described in more detail in Appendix 1.

Sodars

The quality of the sodar data was found to be dependent on the age of the instrument. Newer commercial systems generally had few data quality problems, except for occasional bad winds during periods surrounding rain events. Some of the older sodar systems (up to 25 years old) also had persistent bad range gates, probably due to reflected acoustic signals from nearby structures, and questionable winds in the upper range gates where the signal became weak. No special QC was applied to any of the sodar data in real-time before assimilation.

Tall tower, nacelles, surface mesonet winds

On most of the tall towers, each measurement level had two anemometers located on booms positioned 180 degrees apart. This is done so that for all wind directions there will be at least one anemometer that is situated out of the tower's wake. Wind speeds measured within the wake can be significantly reduced from the free-stream wind speed. The common practice in the wind energy

industry is to avoid waked sensors by rejecting the lower wind speed sensor when two or more observations are available at the same height, and the same was done for the WFIP analysis.

A second problem that can occur with cup anemometers is the phenomena known as “cup over-speeding”. The physical mechanism underlying this phenomena is simply that cup anemometers respond more quickly to an accelerating wind than to a decelerating wind. Studies of cup anemometer over-speeding (Kristensen, 1998, 1999) indicate that the size of the error can be large (> 10%) for poorly designed anemometers under highly turbulent atmospheric conditions, but with careful design can be reduced to within about 1% if the cup anemometer is coupled with a fast response wind vane. We assume that the various manufacturer’s cup anemometers used on the tall towers were carefully chosen to have small over-speeding errors, and no corrections were applied.

A significant data quality issue with the cup anemometer observations occurred during snow and icing conditions, when the cup anemometer speeds would first gradually slow down as snow or ice began to accumulate, eventually stop, and then slowly return to normal operation as the snow and ice melted. An automated algorithm was developed that searched for icing characteristics within a 1-hour sliding window (6 points for the 10 min averaged data, 4 points for the 15 min ERCOT data). If the hourly mean wind speed was less than 1.0 ms^{-1} , the standard deviation of the wind speed was less than 0.2 ms^{-1} , and the temperature was less than 5 C, all of the observations within the hour window were eliminated. The window was then advanced by 10 or 15 minutes, and the process repeated. A similar check was done for wind direction measurements, eliminating all points within an hour that had a standard deviation less than 0.01 degrees if the temperature was also less than 5 C. The icing algorithm was only applied to the tower wind speeds for the two cold season data denial episodes in November and January. Similar procedures were also applied to the nacelle observations. In addition, speeds from nacelle anemometers were flagged as bad whenever the difference between an individual anemometer’s speed and the mean speed of all the nacelle anemometers within a wind plant was greater than 2 standard deviations of the speed of all the nacelle anemometers within the plant.

Another significant problem found in a substantial fraction of the tall tower sites was the occurrence of large offsets in the wind directions. An example of such an offset is shown in Fig. 2.11 (top panel), which displays the real-time hourly averaged wind directions for a 10 day period in June 2012 from the 29 and 58 m levels at a tall tower location. The black curve is the observed wind direction, and the red is from a RAP data denial control model simulation that does not assimilate any of the WFIP observations. The directions at the 29m and 58m levels in both the observations and the model agree so closely, that at most hours the two levels are indistinguishable. However, a large offset is present between the model and observations. The bottom panel of Fig. 2.11 shows the same data, but with the two observation levels rotated by -79.7 and -78.7 degrees. Once these constant offsets have been applied, the observed and model directions come into close agreement for the entire 10 day period.

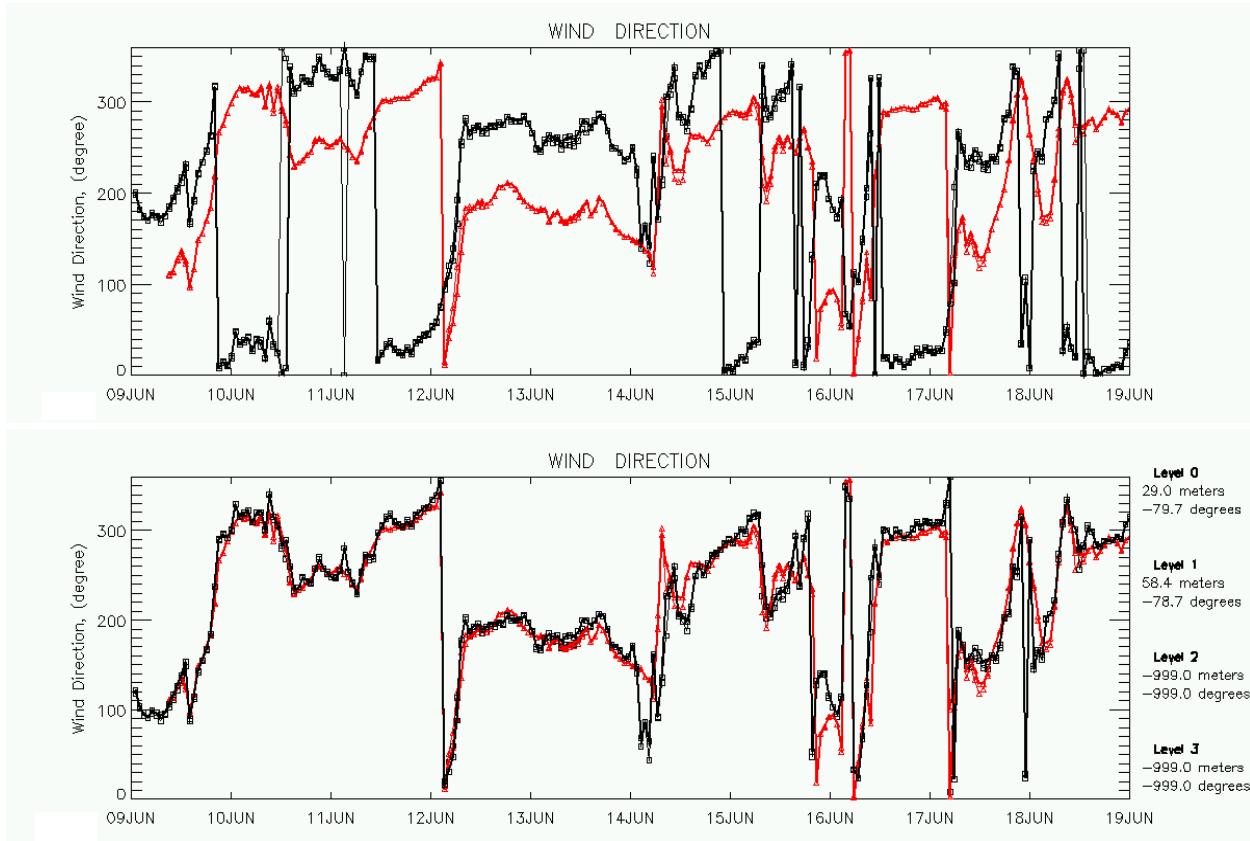


Figure 2.11. 10-day time series of wind directions from the RAP model (red curves) and observations (black curves) for two levels at 29 and 59 m. Top panel) the black curves are from the original raw observations; bottom panel) the two levels of observations have been rotated by a constant offset of -78.7 and -79.7 degrees.

Although this example shows an extreme case when a direction offset correction is both obvious and necessary, a large spectrum of offsets was found between tall tower directions and the forecasts. If the larger offsets are deemed to be instrumental error that should be corrected, the question then becomes what is the threshold for determining that the offset is instrumental and not a real forecast error?

Figure 2.12 is an idealized schematic diagram of the wind direction histograms for two towers, which illustrates two characteristics that can be used to determine when a tower's directions should be corrected. The blue histogram has a relatively narrow, sharply peaked distribution, and a large bias offset, similar to what would occur for the direction errors shown in Fig. 2.11. In comparison, the red histogram is broader, and has a smaller mean direction offset. If the direction error reflects a model deficiency, it is likely due to local, sub-grid scale effects, such as topographic variations. These forecast errors are then likely to change with wind direction, leading to a broader histogram, whereas an instrument offset will give the same error for all directions, leading to a narrower histogram. Therefore the blue histogram is more likely to be due to an instrument wind vane alignment error, and that is large

enough that it should be corrected, while the red histogram is more likely due to a model deficiency, or an instrument direction error that is sufficiently small that it does not require correction.

A non-dimensional measure of the width of the histogram is given by the number of observations O_{2N} that fall within a window of width $2N$ centered on the peak of the histogram (indicated in Fig. 2.12), divided by the total number of observations O_T in the histogram. Dividing by the width of the window $2N$ removes the dependency on the choice of the window width, and then multiplying by the wind direction mean bias MB gives the dimensionless factor DF that incorporates both the histogram width and the magnitude of the direction error and that can be used to quantify the necessity to correct the error:

$$DF = \left(\frac{O_{2N}}{O_T} \right) * \left(\frac{MB}{2N} \right) \quad (\text{eqn 2.1})$$

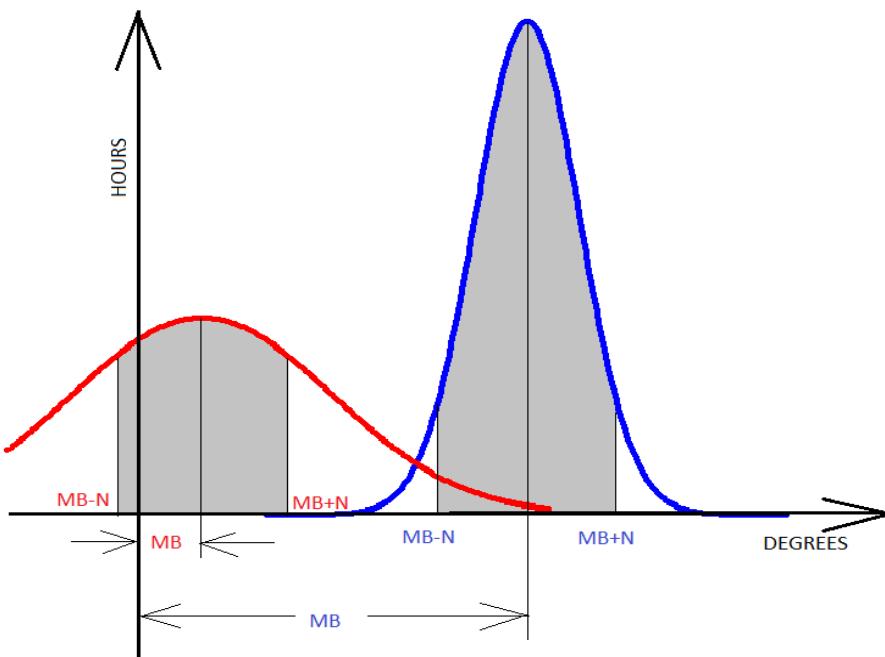


Figure 2.12. Idealized wind direction histograms for a tower with a large mean bias (MB) error and narrow distribution (blue curve), and a smaller mean direction bias with a broader distribution (red curve).

Figure 2.13 shows the values of DF for each of the tower levels in both study areas. Two values of the window width are used, 11 and 21 degrees (+/- 5 and 10 degrees around the center of the histogram). The two curves are very similar, indicating that DF to first order is independent of the choice of the window width. The choice of a threshold value of DF will then distinguish between those wind direction errors that are likely instrument errors and should be corrected (DF greater than the threshold) and wind direction errors that could be real and are not corrected. A value of $DF = 20$ is shown in the

figure, but we have chosen $DF = 15$, which gives approximately half of the DF values greater than the threshold and half smaller than the threshold.

The histogram of the mean direction bias of all of the towers/levels is shown in top panel of Fig. 2.14, for wind speeds greater than 3 ms^{-1} , indicating a shift of the centroid of the histogram towards positive values. The middle panel of Fig. 2.14 shows only those towers for which $DF < 15$, indicative of a model deficiency or an instrumentation error sufficiently small not to correct. The bottom panel on Fig. 2.14 shows the histogram of mean bias errors for those tower/levels with $DF > 15$. The declination angle (the difference between magnetic and true north) varies with location, and for the northern study area varies between 6 and 12 degrees, while for the southern study area varies between 8 and 10 degrees. The peak in the distribution of the mean direction errors in the lower panel clusters around values that would have occurred if the declination correction was not applied to these towers, while secondary peaks occur near values that would result if the declination angle correction was applied with the wrong sign, or if it was applied twice.

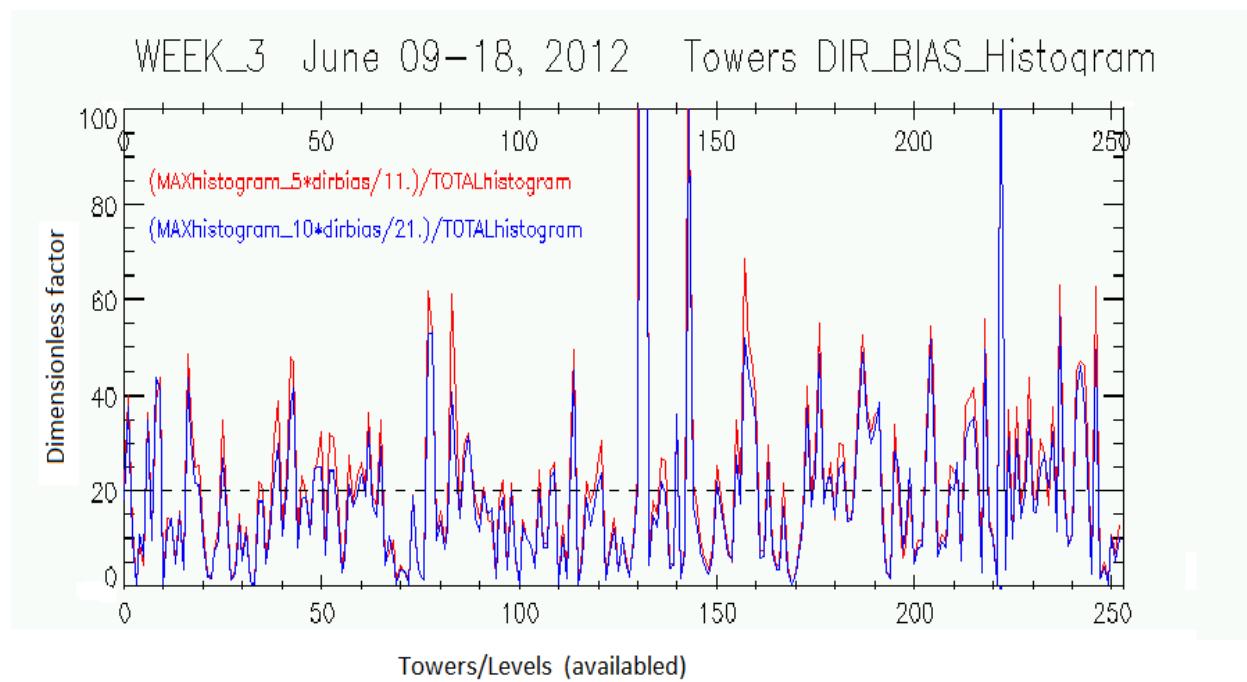


Figure 2.13. The dimensionless direction error factor DF for each of the towers/levels, using both a window width of 11 degrees and 21 degrees, for 10 days of observations. The horizontal dashed line represents a threshold for defining which towers will have their directions corrected and which will not.

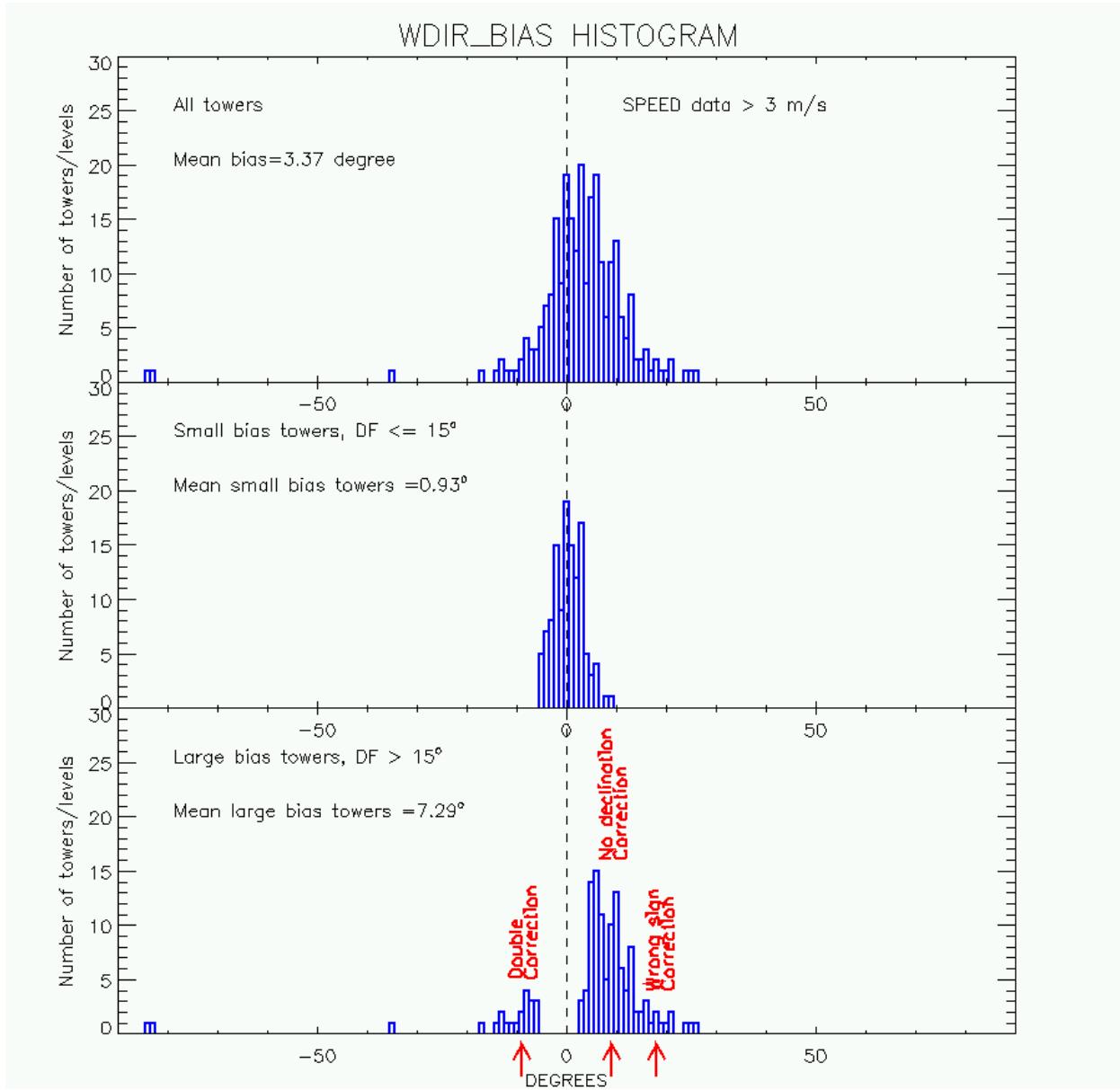


Figure 2.14. Histogram of the mean direction bias of all of the towers/levels, for wind speeds greater than 3 ms^{-1} , for the same 10 days as Fig. 2.13. The top panel is for all of the ~250 towers/levels, the middle panel is for only those that have $DF < 15^\circ$, and the lower panel is for those that have $DF > 15^\circ$.

The wind direction and icing QC algorithms as well as the nacelle outlier algorithm were developed towards the end of the field campaign and were not applied for any of the real-time model simulations. However, for the data denial experiments the icing correction was applied to the tall tower and mesonet stations, the direction correction was applied to the tall tower and mesonet data except for the DOE/PNNL sites (which did not need it), and the nacelle outlier algorithm was applied.

Because of the occasional observation errors present in the sodar data, and occasional remaining errors in the WPR data, these data were further manually edited before assimilation in the retrospective data denial simulations, largely relying on cross-comparisons of the co-located sodar, wind profiler, surface met station observations, and the model forecasts. This editing was done because it was felt that the remaining corrections for all three of these instrument systems could have been applied in real-time with additional development effort to create more sophisticated automated QC algorithms that combined the various observations.

2.3 Instrument Inter-comparisons

The accuracy of the three different remote sensing systems (WPR's, sodars, and lidar) is investigated through inter-comparisons of the data from co-located systems for several different periods. The only site where all three instrument types were deployed was at De Smet, SD. Also, none of the remote sensors were co-located with any of the tall-towers or turbine nacelle anemometers, so only the remote sensors can be inter-compared. Fig. 2.15 shows time-series of real-time hourly averaged winds in the layer in which all three systems provide observations, between 90 and 200m. The two lowest high-resolution mode WPR range gates at De Smet are 138 and 196m, with a range gate spacing of 58m. The five sodar and lidar levels from 100 to 200m were then averaged to correspond to the volume of atmosphere sampled by the WPR. The bottom panel of Fig. 2.15 shows the number of WPR, sodar, and lidar levels that were available at each hour. The maximum number of levels for the sodar and lidar is 5, and is two for the WPR. As can be seen, wind speeds from the three instrument systems generally follow one another quite well, with the exception of the WPR data being fast by 10-15% on the first day of the time-series, and occasional hours when one system or another is an outlier.

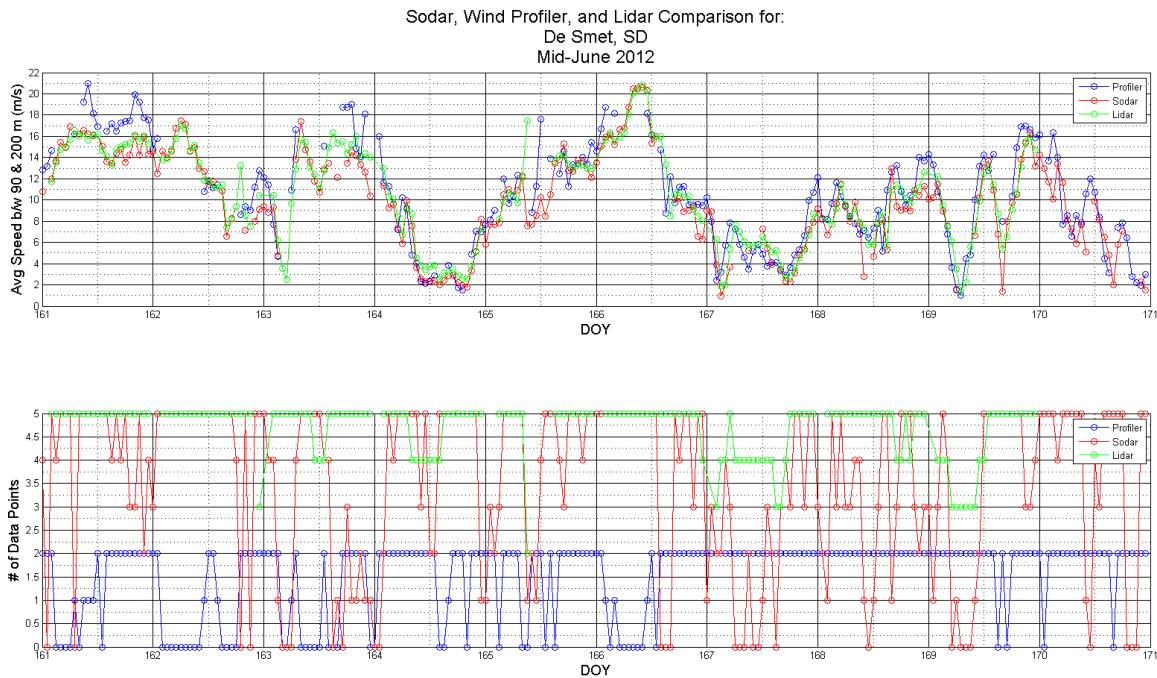


Figure 2.15 Inter-comparisons of WPR, sodar, and lidar data from the June 9-18, 2012 data denial time period, averaged between 100 and 200m AGL. The top panel shows the time series of wind speeds from these three instrument systems over the 10 day period. The bottom panel shows the number of observation levels that go into each average. The two WPR measurement levels were at 138 and 196m AGL.

Figure 2.16 presents scatter plot inter-comparisons of the real-time WPR and sodar data for the Oct. 13-20, 2011 data denial experiment period, for six sites that had co-located WPR's and sodars. The data are again averaged in the layer of overlapping observations for both instrument systems, nominally 90-200m. Good agreement between the two instrument types is found, although there is some variability from site to site. In particular, a speed offset and relatively higher scatter is found at Ainsworth NE, which had an older sodar system. Several large outliers were also present at Buffalo, SD.

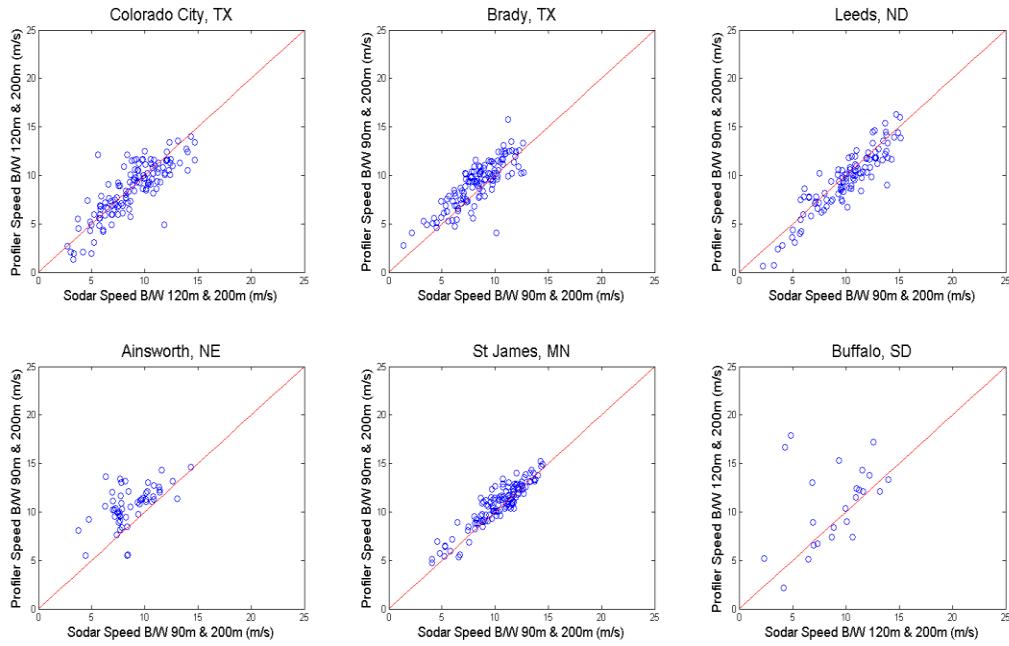


Figure 2.16 Scatter plots of the real-time, hourly averaged WPR and sodar data for six sites that had co-located systems, for the period Oct. 13-20, 2011.

Figure 2.17 repeats the analysis of Fig. 2.16, but uses data after it has passed through the additional quality control procedures used for the retrospective data denial assimilation simulations. The larger outliers have been removed, with overall better agreement between the WPR's and sodars. Additional comparisons between the WPR's and sodars after the additional QC has been applied are shown in Appendix 2 for each of the 6 DD episodes.

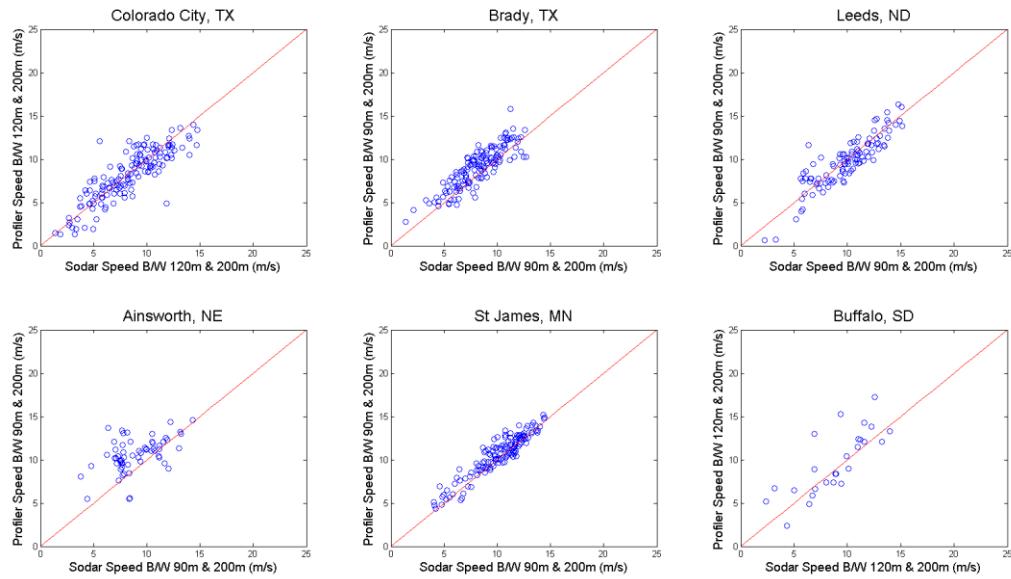


Figure 2.17 Scatter plots of the hourly averaged WPR and sodar data for six sites that had co-located systems after additional QC was applied for the data denial experiment, for the period Oct. 13-20, 2011.

The co-located sodars and WPR's deployed during WFIP allows for a detailed inter-comparison of instrument bias at the range gates where the two instrument systems have overlapping data. Quantifying these biases is important to determine the confidence we can have in the accuracy of each sensor type when operated in a real-world operational setting. In no cases were sodars and industry provided tall-towers co-located, and direct evaluation of potential biases in the tall tower data is not possible. Instead, in Section 6.3 sodar, WPR, and tall tower biases with the model will be calculated, and then these biases will be compared to see if they are in approximate agreement.

3. NOAA Models

Because of the focus on short-term forecasts, the principal NOAA models used during WFIP were the hourly updated 13 km resolution Rapid Update Cycle (RUC), the 13 km resolution Rapid Refresh (RAP), and the 3 km resolution High-Resolution Rapid Refresh (HRRR) (Fig. 3.1). In addition, because the NWS is developing an hourly updated version of the North American Mesoscale (NAM) model, some evaluation of the NAM model was also made, although using forecasts initialized only 4 times per day. The HRRR model in particular has a large potential for application to wind energy forecasting, as its 3 km grid better resolves terrain features affecting turbine-height winds, and also explicitly resolves atmospheric convection such as thunderstorms, which produce outflows responsible for wind ramp events.

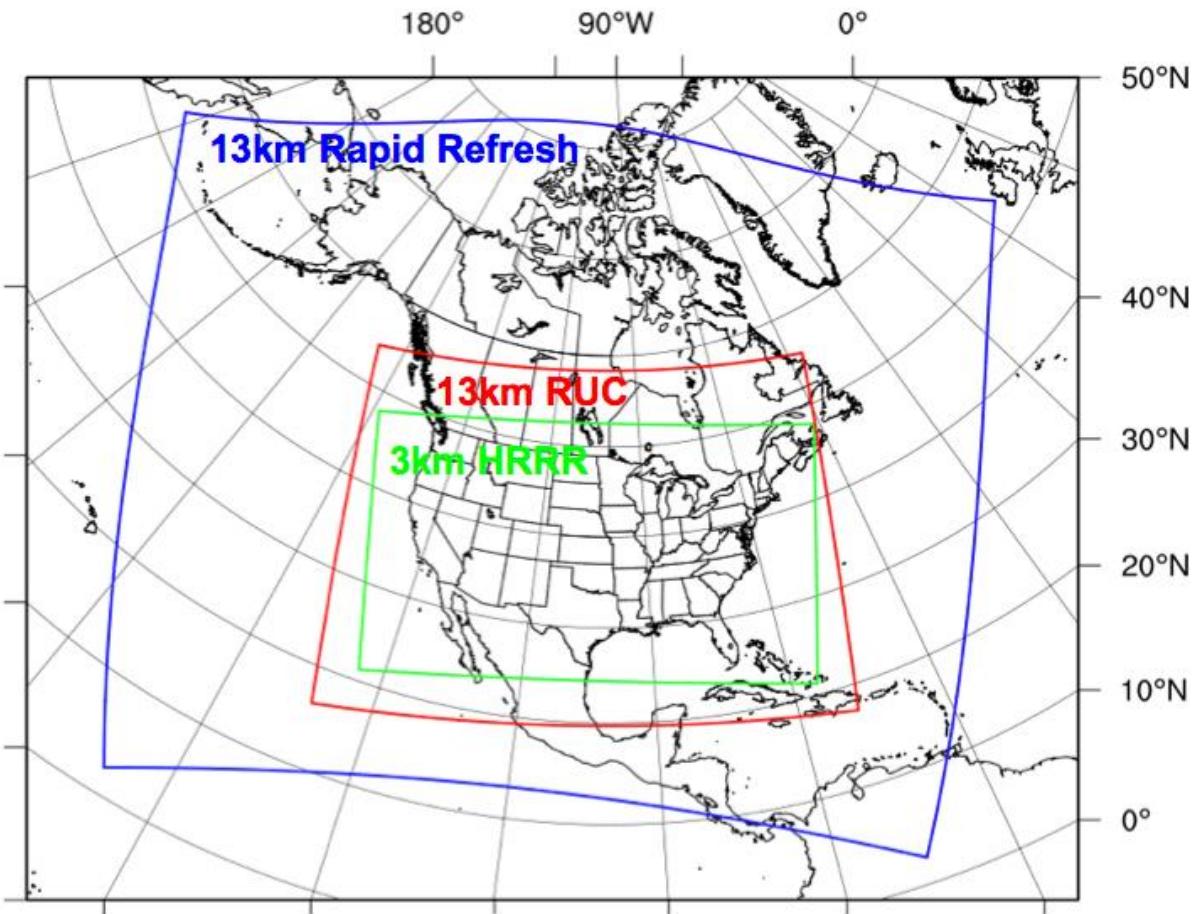


Figure 3.1. Model domains for the 13 km Rapid Refresh (blue), 13 km RUC (red) and the 3 km HRRR (green).

3.1 Rapid Update Cycle (RUC)

The RUC forecast system was run operationally at the National Center for Environmental Prediction (NCEP) from 1994-2012. The RUC features a 13 km domain covering the contiguous U.S. (Fig. 3.1) and is distinctive in two primary aspects: its hourly assimilation cycle and its use of a hybrid isentropic–sigma vertical coordinate. The use of a quasi-isentropic coordinate for the analysis increment allows the influence of observations to be adaptively shaped by the potential temperature structure around the observation, while the hourly update cycle allows for a very current analysis and short-range forecast. Although the RUC was discontinued operationally during WFIP, a semi-operational version was maintained and run at NOAA's Earth System Research Laboratory. The NWS operational RUC was used as a control forecast system with no assimilation of WFIP-special wind observations.

The forecast model component of the RUC uses an updated version of the explicit mixed-phase bulk cloud microphysics originally described as the “level 4” scheme of Reisner et al. (1998) with modifications following Thompson et al. (2004) to parameterize the effects of moist processes. Sub-grid

scale convection is parameterized by the Grell-Devenyi scheme (Grell and Devenyi 2002). The land-surface physics are parameterized by the RUC land surface model (LSM; Smirnova et al. 1997, 2000). The RUC LSM contains a multilevel soil model, treatment of vegetation, and a two-layer snow model, all operating on the same horizontal grid as the atmospheric model. The level 3.0 boundary layer scheme of Burk and Thompson (1989) parameterizes the turbulent mixing. The exchange coefficients regulating the fluxes of heat, moisture, and momentum between the land and atmosphere are prescribed by Monin-Obukhov similarity theory, specifically the three-layer scheme described in Pan et al. (1994). The radiative transfer in the short- and longwave spectrums are parameterized by the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997). Table 3.1 summarizes the RUC model configuration.

3.2 Rapid Refresh (RAP)

The RAP serves as the National Center for Environmental Prediction's regional short-range rapidly updating forecast system, which provides hourly updated forecasts out to 18 hours. The RAP replaced the operational RUC model at NCEP in May 2012, midway through the WFIP field campaign. The primary differences between the RAP and RUC include: (1) the model component of the operational RAP is based upon the Advanced Research Weather version of the Weather Research and Forecasting (WRF-ARW) model (Skamarock 2008), and (2) the data assimilation component uses the 3D-variations Grid Statistical Interpolation scheme (GSI; Wu et al. 2002).

The version of the RAP implemented for WFIP features a 13 km C-grid domain covering North America (Fig. 3.1). The ESRL/RAP model code used for the WFIP project was a more advanced version of the operational RAP and came from ESRL's real-time parallel-test developmental code at the time WFIP experiments began. As in operations, boundary conditions for the RAP were obtained from the previous cycle's forecast from the Global Forecast System (GFS) model.

The RAP uses a modified version of the WRF-ARW with explicit mixed-phase bulk cloud microphysics originally described by Thompson et al. (2008) to parameterize the effects of moist processes. Deep sub-grid scale convection is parameterized by the Grell 3D scheme and shallow-convective processes are parameterized by the Grell shallow-cumulus scheme. Land-surface physics are parameterized by the RUC LSM (Smirnova et al. 1997, 2000). The Mellor-Yamada-Janjic (level 2.5) boundary layer scheme (Janjic 2001) parameterizes the turbulent mixing. The exchange coefficients regulating the fluxes of heat, moisture, and momentum between the land and atmosphere are prescribed by Janjic (1994). The radiative transfer in the shortwave spectrum is parameterized by the Goddard scheme (Chow and Suarez 1994) and longwave spectrum is parameterized by the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997). Table 3.2 summarizes the RAP model configuration.

The real-time ESRL RAP forecasts were run out to 15 hours with output at 60 minute intervals, while output was stored at 15 minute intervals for the data denial simulations.

3.3 High Resolution Rapid Refresh (HRRR)

The HRRR features 3 km grid spacing with a domain covering the contiguous U.S. (Fig 3.1). The HRRR is not yet run operationally at NCEP but is planned for implementation in 2014. However, the HRRR is run in a quasi-operational 24/7 developmental mode at NOAA/ESRL and already has a wide user base, including NOAA Weather Forecast Offices and private sector organizations. The primary purpose of the HRRR is to improve the operational capability of forecasting high-impact convective storms, which play an important role in the ramping of low-level winds. The version of the HRRR in development during WFIP did not perform additional data assimilation on the 3-km grid. The initial and boundary conditions were obtained by direct interpolation from the RAP. The HRRR was run hourly, out to 15 hours, within ESRL's high-performance computing facility.

The forecast model component of the HRRR, like the RAP, uses a modified version of the WRF-ARW. Since most convective processes can be adequately resolved at 3-km grid scales, no deep- or shallow-convection schemes are used. The rest of the physical processes are parameterized using the same schemes employed in the RAP (above). Table 3.3 summarizes the HRRR model configuration.

3.4 NAM and NAM CONUSNEST

The NAM serves as the National Weather Service's regional short to mid-range NWP system which provides forecasts out to 84 hours four times a day at 00 UTC, 06 UTC, 12 UTC, and 18 UTC. The current configuration of the operational NAM was implemented in October of 2011 and is based upon the Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjic, 2003; Janjic, 2005; Janjic and Black, 2007; Janjic and Gall, 2012).

The version of the NAM implemented for WFIP data denial studies featured two domains (Fig. 3.2), a parent 12 km domain and a one-way nested 4 km domain covering the contiguous United States (CONUS). The coverage of both of these domains is identical to that covered by the operational NAM. The NMMB model code used for the WFIP project was a more advanced version of the operational NMMB and came from NCEP/EMC's real-time, parallel-test developmental code at the time WFIP experiments began. The general configurations of both the 12 km and 4 km domains may be found in Tables 3.4 and 3.5, respectively. As in operations, boundary conditions for the 12 km parent domain were obtained from the previous cycle's forecast from the GFS model. All 12 km forecasts ran out to 84 hours and the 4 km forecasts ran out to 60 hours. Both domains produced hourly output to 36 hours.

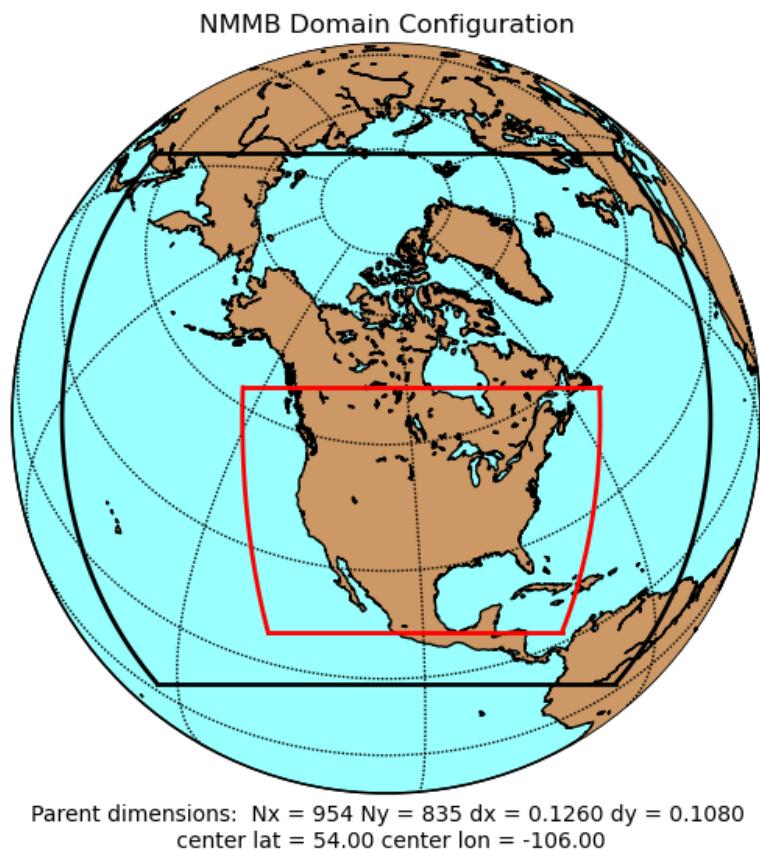


Figure 3.2. NAM Parent 12 km (black) and 4 km CONUSnest (red) computational domains used during WFIP.

13 km RUC Description	Configuration
Points in x, y, z directions	451, 337, 51
Microphysics parameterization	Thompson et al. (2004)
Boundary layer parameterization	Modified Burk-Thompson (1989)
Convective parameterization	Grell and Devenyi (2002)
Long/short wave radiation parameterization	Mlawer et al. (1997)
Land surface model	Smirnova et al. (1997, 2000)

Table 3.1. 13 km RUC domain configuration.

13 km RR/RAP Description	Configuration
Points in x, y, z directions	758, 567, 51
Microphysics parameterization	Thompson et al. (2008)
Boundary layer parameterization	Janjic (2001)
Convective parameterization	Grell 3D/Grell shallow-cumulus scheme
Long/short wave radiation parameterization	Chow and Suarez (1994)/Mlawer et al. (1997)
Land surface model	Smirnova et al. (1997, 2000b)

Table 3.2. 13 km Rapid Refresh domain configuration.

3 km HRRR Description	Configuration
Points in x, y, z directions	1800, 1060, 51
Microphysics parameterization	Thompson et al. (2008)
Boundary layer parameterization	Janjic (2001)
Convective parameterization	Turned off
Long/short wave radiation parameterization	Chow and Suarez (1994)/Mlawer et al. (1997)
Land surface model	Smirnova et al. (1997, 2000)

Table 3.3. 3 km HRRR domain configuration.

12 km NAM Parent Description	Configuration
Points in x, y, z directions	954, 835, 60
Microphysics parameterization	Ferrier et al. (2002, 2011)
Boundary layer parameterization	Janjic (2001)
Convective parameterization	Janjic (1994)
Long/short wave radiation parameterization	Iacono et al. (2008), Mlawer et al. (1997)
Land surface model	Ek et al. (2003)
Gravity wave drag parameterization	Alpert (2004)

Table 3.4. 12 km NAM domain configuration.

4 km CONUSnest Description	Configuration
Points in x, y, z directions	1371, 1100, 60
Microphysics parameterization	Ferrier et al. (2002, 2011)
Boundary layer parameterization	Janjic (2001)
Convective parameterization	Janjic (1994): Modified to be less active for higher resolution
Long/short wave radiation parameterization	Iacono et al. (2008), Mlawer et al. (1997)
Land surface model	Ek et al. (2003)
Gravity wave drag parameterization	None

Table 3.5. The NAM 4 km CONUSnest domain configuration.

3.5 RAP and HRRR improvements

During the course of WFIP numerous improvements were made to the RAP and HRRR models, and these improvements were then included in the data denial experiments that were run after the end of the field campaign. These improvements were the result of many different funded efforts focused on model development, and some occurred as a result of WFIP. Improvements that occurred during the course of WFIP (incorporated during 2012) are listed in Table 3.6. These include improvements made to the model physics, model numerics, data types assimilated, and data assimilation procedures, including the assimilation of wind profiling radar data.

3.6 HPC & Data Storage Requirements

The High Performance Computing (HPC) and data storage requirements for just the data denial simulation experiments can be substantial. If the full domain output was saved for the RAP model for the 55 data denial experiment days, over 280 Terabytes (TB) would have been required. Since this was not possible, model output was saved only over a truncated domain spanning the central U.S., reducing the data storage requirements to 35 TB.

The NAM WFIP data denial simulations occupy approximately 67 Terabytes of archived disk. This exceptionally high disk usage reflects the retention of several very large files which have been saved to restart NAM and NAM CONUSnest model forecasts if necessary. Furthermore, the addition of the NAM's 4 km CONUSNEST domain leads to a substantial increase in the amount of required disk space. All data denial simulations, both RAP and NAM, were run on Zeus, the NOAA research and development high performance computing system. Model forecast jobs used 1200 processors to run the nested configuration of the NAM 12 km parent and 4 km CONUSNEST domains and 196 processors for the RAP domain. The GSI used 240 processors for each data assimilation step that was run for the NAM's 12 km and 4 km domains.

	Model	Data Assimilation
RAP (13 km)	WRFv3.3.1+ Physics changes (convection, microphysics, land-surface, PBL) Numerics changes (w damping upper boundary conditions, 5 th -order vertical advection) MODIS land use, fractional 30→10 min shortwave radiation New reflectivity diagnostic	WPR assimilation heights Soil adjustment, Temp-departure radar-hydrometeor building Precipitable Water assim mods Cloud assim mods Tower/nacelle/sodar observations GLD360 lightning GSI merge with trunk
HRRR (3 km)	WRFv3.3.1+, Physics changes (microphysics, land-surface, PBL) Numerics changes (w damping upper boundary conditions, 5 th -order vertical advection) MODIS land use, fractional 30→05 min shortwave radiation New reflectivity diagnostic	

Table 3.6. Improvements made to the NOAA/ESRL RAP and HRRR research models in 2012.

4. Data Assimilation

Numerical weather prediction is an initial value problem and the atmosphere is a nonlinear, chaotic system which, when modeled, exhibits a strong sensitive dependence on initial conditions (Lorenz 1963). Therefore it is important that the best available initial conditions be used to initialize NWP models in order to yield the best possible forecasts. The process in which the best available initial conditions are obtained is through a procedure known as data assimilation, which combines a model

forecast with observations to provide an estimate of the current state of the atmosphere. This estimate, i.e. the analysis, is then used as the initial state from which forecasts are initialized.

The hourly data assimilation of the RUC forecast system uses an older three dimensional variational assimilation (3Dvar) technique. The analysis framework includes ingest and preprocessing of the observations and the calculation of innovations [discussed in Devenyi and Benjamin (2003)]. The background field is the previous 1-hr RUC forecast in its native coordinate. The RUC employs a diabatic digital filter initialization technique (Huang and Lynch 1993; Benjamin et al. 2004b) to develop physically-balanced circulations associated with the latent heating inferred from radar observations and applies a cloud analysis, using satellite data and surface ceiling observations, to initialize an the three-dimensional cloud field.

Both the RAP and NAM forecast systems use the Gridpoint Statistical Interpolation system (GSI; Wu et al. 2002) for data assimilation. The GSI analyzes the following variables: streamfunction, velocity potential, surface pressure, temperature, and normalized-relative humidity [a multivariate relation involving specific humidity, temperature, and pressure (Holm et al., 2002)].

The GSI is a complex variational data assimilation system which is capable of assimilating a diverse set of observations. Such observations include, but are not limited to, radiosondes, wind profilers, Doppler radar radial velocities, satellite radiances, surface observations, etc. With the advent of WFIP the capability to assimilate both wind turbine nacelle and tall tower observations has been developed for the RAP and NAM forecast systems within the GSI framework.

The implementation of the GSI system for the WFIP project was under the context of 3DVar, which minimizes a cost function that measures the distance to the background forecast and observations (Kalnay, 2003). The model analysis is then globally adjusted to all the observations available during the assimilation period (Talagrand, 1997). For GSI 3DVar, the following incremental cost function is minimized

$$J = \frac{1}{2} [\mathbf{x}^T \mathbf{B}^{-1} \mathbf{x} + (\mathbf{Hx} - \mathbf{y})^T \mathbf{R}^{-1} (\mathbf{Hx} - \mathbf{y})] \quad (\text{eqn. 4.1})$$

Where \mathbf{x} is a column vector of analysis increments, $\mathbf{x} = \mathbf{x}^a - \mathbf{x}^f$. Here superscripts 'a' and 'f' denote an analysis and forecast, respectively. The vector \mathbf{x} , and its constituents, has length n which corresponds to the total number of model gridpoints times the number of analysis variables (e.g. streamfunction, temperature, etc.). Matrix \mathbf{B} is the background error covariance matrix and is of dimension $n \times n$. \mathbf{H} is the (possibly nonlinear) observation operator which maps forecast variables to observations and has dimension $p \times n$, where p corresponds to the total number of observations to be assimilated. Finally, \mathbf{y} is a column vector of observation innovations and takes the form of $\mathbf{y} = \mathbf{y}_{\text{obs}} - \mathbf{Hx}^f$ and has length p , where subscript 'obs' denotes the actual observations (e.g. nacelle wind speeds). To find the analysis increment which minimizes the cost function the iterative preconditioned conjugate gradient algorithm of Derber and Rosati (1989) is used.

In the practical implementation of 3DVar the background error covariance matrix, \mathbf{B} , must be estimated *a priori*. The structure of this matrix is quite important, as it largely determines how the information from the assimilated observations is spread out into the analysis, i.e. the analysis increments (Daley, 1991; Kalnay, 2003). In the GSI \mathbf{B} may be estimated using the so-called NMC method (Parrish and Derber, 1992) or the ensemble-analysis method (Houtekamer et al. 1996). Both techniques follow the methodology of using the average of many sets of forecast differences, verifying at the same time, to estimate \mathbf{B} . Along with the estimation of \mathbf{B} , the GSI includes additional steps to model the cross-variable covariances using linear balance relationships via statistical regression, which allows for a coupling of mass and wind fields in the resulting analysis (Parrish et al., 1997; Wu et al., 2002). Finally, the spatially univariate correlations are modeled using an isotropic recursive filter, which has the response of a Gaussian function and spreads the analysis increments to nearby gridpoints (Purser et al., 2003).

The observation error covariance matrix, \mathbf{R} , contains instrumentation errors, representativeness errors, and errors associated with the observation operator, \mathbf{H} . In practice it is assumed that all observation errors are independent and uncorrelated (Lorenc, 1986; Kalnay, 2003), thus rendering \mathbf{R} a diagonal matrix of error variances. In GSI observation errors for conventional observations, e.g. nacelle winds and radiosondes, are specified through an external error table file which stores the observation error standard deviations as a function of vertical pressure level. Observations also have a certain amount of self-descriptive meta-data associated with them which can yield additional information about observation quality. This extra data can be used to quality control observations during the analysis process by adaptively inflating observation error and/or rejecting observations completely. For example, observations may be rejected through a gross error test, which checks the observation against the background forecast, if the magnitude of the difference is too large, the observation may be rejected.

The GSI settings and configuration used for both the RAP and NAM are described in the following subsections.

4.1 RAP/HRRR

The RAP data assimilation system used for WFIP is a developmental version of GSI, which includes updates from the operational RAPv1. RAP-GSI assimilates all standard observations as well as WFIP-special wind observations from towers and nacelles throughout the WFIP study region.

The GSI version the RAP implemented for WFIP, as well as in operations, used the so-called “partial cycling” procedure. Partial cycling for the RAP involves a twice-daily 6-hr spin-up cycle from an initial condition taken from a 3 hr GFS forecast, valid at 03 and 15 UTC. The 1-hr forecast from the 6th cycle is then injected into the regular hourly cycled system at 09 and 21 UTC. Note that for the RAP, only the atmosphere component uses this partial cycling technique, whereas the soil moisture and temperature

fields are continuously cycled. This is done so that the soil state is kept physically consistent with the RUC Land Surface Model (LSM) diffusion of heat and moisture.

Reducing the effects of imbalances introduced during the data assimilation step is addressed with a diabatic digital filter. Digital filtering produces an initial atmospheric state that is balanced within the context of the model's dynamics (Huang and Lynch, 1993) by filtering high-frequency noise from an unbalanced initial state. In hourly cycled systems, noise can accumulate with each successive cycle, so the use of digital filter initialization is important for the mitigation of noise (e.g. Benjamin, 2004b). In the RAP, a diabatic digital filter is also used for the assimilation of radar reflectivity, where a prescribed latent heating is added to the model's temperature tendency term, proportional to the radar retrieved strength, to help spin-up a physically consistent ageostrophic circulation associated with the precipitating storm-scale structures. The RAP uses a filter window length of 40 minutes, which is invoked at the beginning of every hourly cycle. The RAP also includes a cloud analysis procedure, using satellite data and surface ceiling observations, to initialize an accurate three-dimensional cloud field.

The current developmental version of the HRRR does perform data assimilation on the 3 km grid, but during WFIP there was no data assimilation performed and no diabatic digital filter used within the HRRR framework. The HRRR solely benefited from the data assimilation performed within the RAP by means of interpolated RAP analyses used to generate initial and boundary conditions.

4.2 NAM/NDAS and CONUSnest

The version of the NAM and NAM Data Assimilation System (NDAS) implemented for WFIP, as well as in operations, also uses the partial cycling procedure. Partial cycling for the NAM/NDAS involves using the atmospheric variables from a 6 hour forecast from the Global Data Assimilation System (GDAS) as the first guess for the atmospheric state at the beginning of the 12 hour long analysis-forecast-analysis window of the NDAS (Fig. 4.1). The land states, however, are still cycled from the previous, most recent NDAS cycle to maintain physical consistency with the model's Land Surface Model.

The WFIP project allowed for testing and introduction of several, new experimental features within the NAM/NDAS. The major additions include adding analysis-forecast steps during the NDAS for the 4 km CONUSnest, switching on the use of a diabatic digital filter initialization technique, and the introduction of the capability to assimilate special wind energy observations (nacelle and tall tower observations) into the NAM/NDAS system for the first time.

In the current operational configuration of the NDAS, the 4 km CONUSnest domain is not cycled and is initialized from a downscaled/interpolated field from the 12 km parent NAM domain. However it has been planned to add the 4 km CONUSNEST to the NDAS assimilation procedure (e.g. Fig. 4.1). In this configuration the 4 km CONUSNEST would go through an assimilation cycling procedure just as the 12 km parent domain does, thus allowing the initial conditions for the CONUSNEST forecast to be more consistent with its spatial resolution. NCEP/EMC recognized that the WFIP project was a good

opportunity to implement such a capability for testing and a substantial effort was invested in adding this feature. As a result of this work, this feature has also been included in a development version of an hourly-updated version of the NAM/NDAS.

The issue of initialization and reducing the effects of imbalances introduced during the data assimilation step has been a longstanding challenge in NWP (e.g. Daley, 1991). In the operational version of the NDAS extra divergence damping is applied to mitigate the accumulation of excessive noise (i.e. from TM12-TM03 in Fig. 4.1), where TMXX is the model initialization time minus XX hours. However, alternate initialization techniques also exist which accelerate the model adjustment process, such as digital filtering, which produces an initial atmospheric state that is balanced within the context of the model's dynamics (Huang and Lynch, 1993). Furthermore the use of digital filter initialization has also been considered a necessity for the mitigation of noise in the implementation of hourly, rapidly updating forecast models (e.g. Benjamin, 2004b), something which is being actively pursued for the NAM. In the WFIP version of the NAM/NDAS system a diabatic digital filter (Lynch et al., 1997) was applied immediately after each analysis, for both domains, using a filter window length of 40 minutes.

Finally, the development of the ability to assimilate nacelle and tall tower observations from the wind energy community is another benefit from the WFIP project. Given the availability of these new wind energy data sets, the WFIP project should help accelerate the ingest of these observation types into the operational RAP and NAM/NDAS.

4.3 Additional Observational Data Processing

NCEP operational observation processing was used for encoding the profiler, SODAR, and RASS data into the file format used by the GSI assimilation system, known as prepBUFR (prepared Binary Universal Form for the Representation of meteorological data). During the encoding process, NCEP's profiler complex quality control algorithm was applied to both the profiler and sodar observations, which is based upon observation differences from GDAS forecasts (e.g. Gandin, 1988).

Once these observations were encoded into prepBUFR, the new prepBUFR files containing all conventional observations in addition to the special WFIP profiler, SODAR, and RASS observations were transmitted to ESRL for the inclusion of nacelle and tall tower observations. These prepBUFR files were then used for assimilation within the NAM and RAP forecast systems.

Finally, given the high density of nacelle data and the fact that each nacelle anemometer is mounted directly behind rotating turbine blades, an averaging method was necessary to create a more robust estimate prior to assimilation. The method chosen was a three step approach: (1) a mean of all nacelle observations was taken for every $30 \times 30 \text{ km}^2$ region, (2) all nacelle observations which deviated from the mean by more than two standard deviations were excluded, (3) the remaining nacelle observations were again averaged to get a single estimate within the region. This method generally excluded about 10% of the nacelle observations and reduced the amount of single nacelle observations from about 300 to about 18 averaged observations.

4.4 GSI 3DVar parameter settings

The WFIP version of the GSI was implemented for both the RAP and NAM in a manner which closely mimicked that used in operations. In particular, the GSI data assimilation system eliminates observations based upon the following gross error test:

$$|Obs - Background| > (obserror * gross error) \quad (\text{eqn. 4.2})$$

where “Obs” refers to a particular observation and “Background” refers to the model forecast equivalent of the observation, interpolated to the observation location (i.e. \mathbf{Hx}^f).

The background error statistics, \mathbf{B} , for the RAP are the same as those used in the operational RAP, which originated from the GDAS with modified recursive filter correlation lengths. Background error statistics for the NAM are the same as those used in the operational NAM for cycles at TM09-TM00 (Fig. 4.1). These background error statistics were derived using 60 three hour forecast pairs based upon the method of Houtekamer et al. (1996). At TM12, the beginning of the NDAS window (Fig. 4.1), the background error statistics from the GDAS system are used since the first guess forecast at this time corresponds to a six hour forecast from the GDAS.

The observation errors for all conventional observations assimilated into the RAP and NAM in WFIP were set to be identical to the errors used in the operational versions of those models, with the exception that the WFIP profiler and sodar observation errors were reduced, since extra quality control was undertaken to ensure high data quality. The values of observation error and gross error for the real-time forecasts are set as:

Real time runs:

	profilers(u,v)	Sodars(u,v)	Towers(u,v)	Nacelles (spd)	Mesonet(u,v/T,q)
Obs error	3.7-10.0	3.7-10.0	3.5	3.5	1.5/1.0
Gross error	5.0	5.0	1.5	1.5	5.0/7.0

Table 4.1. Real-time forecast GSI values of observation error and gross error for the assimilated WFIP instrumentation types.

where the profiler and sodar observation errors are equal to 3.7 ms^{-1} up to 700 mb, then increase by 0.2 ms^{-1} every 50 mb above that, up to a max value of 10.0 ms^{-1} . The values of observation and gross errors for the data denial simulations are set as:

Data denial (DD) simulations:

	profilers(u,v)	Sodars(u,v)	Towers(u,v)	Nacelles (spd)	Mesonet(u,v/T,q)
Obs error	2.0-5.0	2.0-5.0	1.6	1.6	1.5/1.0
Gross error	7.0	7.0	7.0	7.0	5.0/7.0

Table 4.2. Data denial simulation GSI values of observation error and gross error for the assimilated WFIP instrumentation types.

For the real-time forecasts the gross error test will reject profiler and sodar observations below 700mb when they differ from the background field by more than $5.0 \times 3.7\text{ms}^{-1} = 18.5 \text{ ms}^{-1}$, and tower/nacelle observations by more than $1.5 \times 3.5\text{ms}^{-1} = 5.25\text{ms}^{-1}$, while for the data denial simulations the gross error test will reject profiler and sodar observations when they differ from the background field by more than $7.0 \times 2.0\text{ms}^{-1} = 14.0 \text{ ms}^{-1}$, and tower/nacelle observations by more than $7.0 \times 1.6\text{ms}^{-1} = 11.2\text{ms}^{-1}$. The settings for the profiler's and sodars for both the real-time and DD simulations and for the tall tower/nacelles for the DD simulations are set sufficiently lax that virtually all observations are always accepted. For the non-QC'd real-time tall tower/nacelle observations, the tighter parameter settings eliminated some observations.

Finally, the time window for the observations was also modified. The operational NAM and RAP use a very short time window of ± 6 minutes of the analysis time to select the observations to assimilate. This effectively left out a large number of WFIP field experiment observations from the analysis. Therefore the time window for these observations in the data denial NAM and RAP runs was expanded to ± 21 minutes of the analysis time to ensure successful ingest of the WFIP field experimental observations.

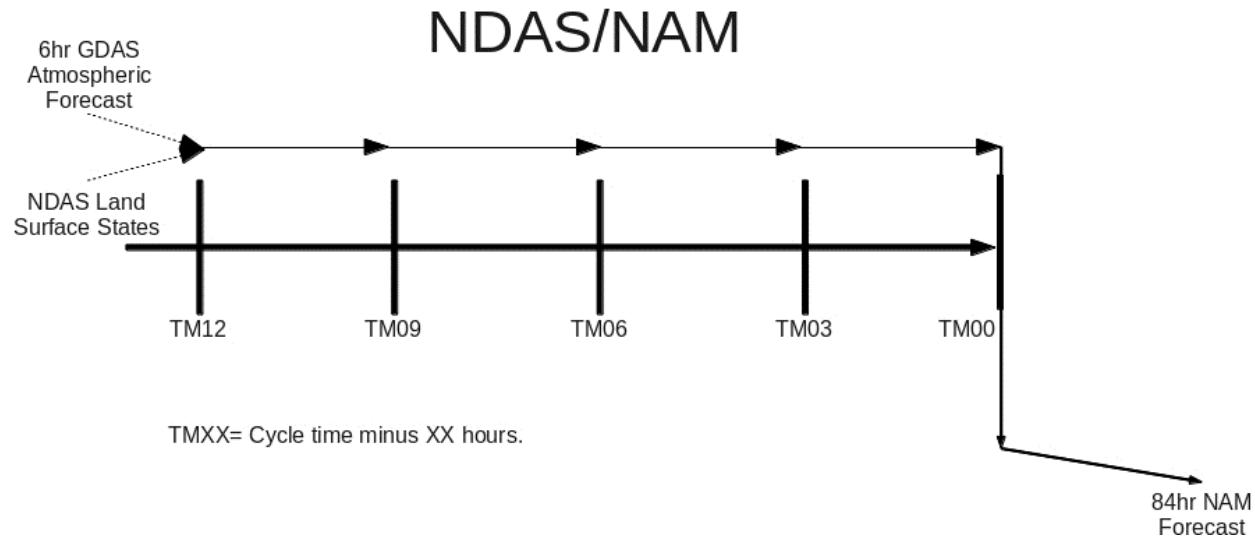


Figure 4.1. NAM/NDAS data assimilation cycling diagram. Each forecast cycle begins with a 12 hour analysis-forecast window during which analyses are conducted at three hour intervals (TM12, TM09, etc.). TM00 refers to the forecast initialization time (e.g. 00, 06, 12, or 18 UTC). At TM12 the first guess for the atmosphere is a 6 hour forecast from the GDAS. The land states are, however, still cycled from the previous NAM/NDAS cycle.

5. Evaluation of Real-Time Forecasts

5.1. Real-time model evaluation web site

To assist in maintaining the WFIP instrumentation in continuous working order throughout the year-long field campaign and in identifying potential model problems, the observations and model forecasts were displayed continuously on a real-time publically accessible web site, updated on a sub-hourly basis. The web sites for the NSA and SSA can be found at:

<http://wfip.esrl.noaa.gov/psd/programs/wfip/North/>

<http://wfip.esrl.noaa.gov/psd/programs/wfip/South/>

An example screen from the NSA web page is shown in Fig. 5.1. Buttons on the web page allow for selection of seven different models, four observation types (profilers, sodars, lidar, surface met), each of the WPR, public sodar, and lidar sites, and each of the model initialization times. In addition, non-public versions of the web site were developed for both of the private sector partners that displayed the proprietary sodar, tall tower and nacelle observations, as well as comparisons of the various models with those proprietary observations. The web sites show both vertical profile data and surface data time-series for 24 hour periods that are updated each hour, with the corresponding model forecasts displayed out to the length of the forecast made, typically 15h for the RR and HRRR models. Also separate buttons for the RR and HRRR models (in orange) allow for visualization of which observations at the initialization hour were accepted by the model data assimilation system. The ability to peruse previous day's data is maintained through the date selection tool on the web site calendar. In the example shown in Fig. 5.1, the high-resolution WPR wind vector (barbs) and speed (filled color) data at Buffalo ND (top panel) are compared to a 15 h forecast from the ESRL RR model (lower panel) initialized at 00 UTC, Sept. 1, 2012. In this particular example the model is seen to have large speed discrepancies of up to 15 ms^{-1} at forecast hours 6-9. The real-time web site was essential during WFIP for monitoring instrument status, allowing for engineers and technicians to rapidly respond to instrument problems thereby minimizing data outages. Also, the ability to compare observations with model output in many cases allowed for the identification and correction of subtle problems with instruments that would not have been detected by examining the observations alone.

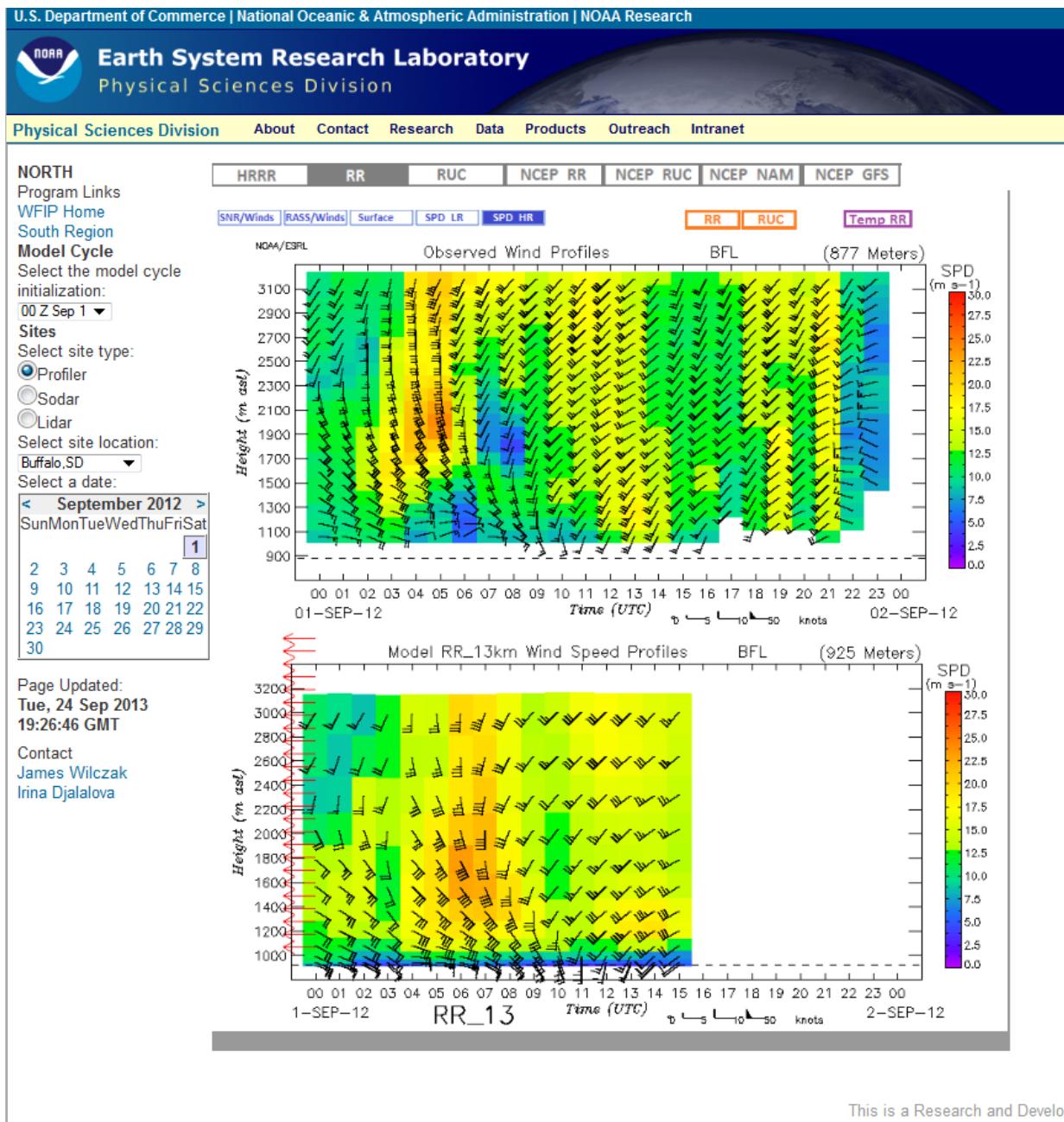


Figure 5.1 A screen-shot of the main page of the model/observation evaluation web page

5.2. Conversion of wind speed to power

In order to properly evaluate the skill of an NWP model at forecasting winds for wind energy, it is essential to convert from wind speed to the equivalent power that a wind turbine would produce. This is necessary because wind speed errors produce corresponding turbine power production errors only for a range of moderate wind speeds. Errors at low speeds do not matter as the speeds are too low for the

turbine to produce any power, and errors at very high speeds do not matter because the turbine will be producing at full capacity in any case. In addition, the wind speed is nonlinearly related to the power generated which can make interpretation of wind speed forecast errors difficult to translate directly into forecast errors in wind power.

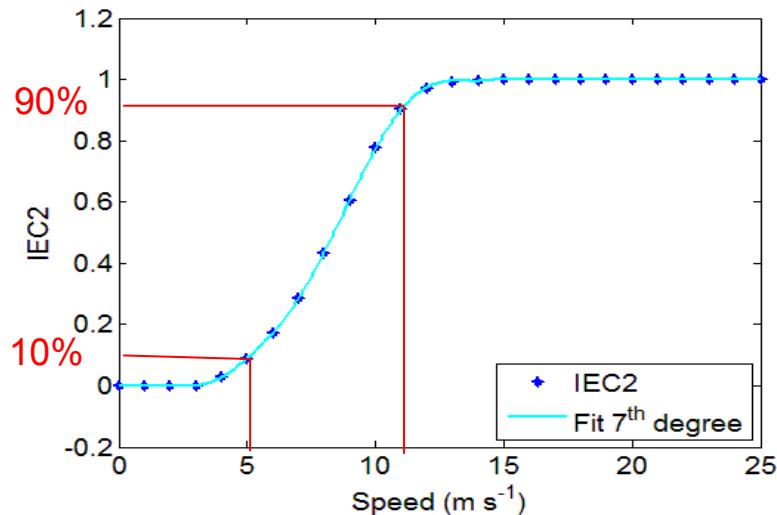


Figure 5.2. An IEC class 2 wind turbine power curve, which shows the expected wind power produced as a function of wind speed for this class of turbine.

The conversion of wind speed to wind power follows a wind turbine's power curve, an example of which is shown in Fig. 5.2 for a standard IEC Class 2 wind turbine, which is the most common type of wind turbine used in the U.S. Midwest. As can be seen, this type of turbine produces only 10% of its maximum possible power at a speed of 5 m s⁻¹, increasing to 90% of full power near 11.5 m s⁻¹. The power curve shown in this figure has been used throughout the NOAA WFIP analysis to quantify model forecast errors and forecast error improvements.

5.3. Bulk error statistics: RAP and RUC models

From the start of WFIP through April 30, 2012, the NOAA/NWS/NCEP operational hourly-updated model was the Rapid Update Cycle (RUC) model. This operational model did not assimilate any of the special WFIP observations. Basic MAE bulk-statistics comparisons of the NCEP/RUC model with the ESRL/RAP model are shown in Fig. 5.3 for the vector wind evaluated using the 39 real-time tall towers in the NSA. Since the NCEP/RUC model did not assimilate in the new observations while the research ESRL/RAP model did, the improvement in forecast skill of the ESRL/RAP over the NCEP/RUC combines fundamental model improvements of the RAP over the RUC, as well as the impacts of assimilation of the WFIP data. Because the real-time observations were not quality controlled to the same level as in the DD

simulations, the contribution to the improvement from assimilation of the new observations in these real-time evaluations may be reduced accordingly. The fundamental model improvements of the RAP over the RUC are the result of many years of research and development that preceded WFIP, and also reflect a few changes that were made as a result of WFIP. The percent improvement of the ESRL/RAP over the NCEP RUC is quite significant, as large as 13% at for forecast hour 1, decreasing to 6-7% for forecasts hours 7-15.

During this same time period of Oct – April 2012, NCEP was running a test version of the RAP model, which replaced the operational RUC model on May 1, 2012. Also shown in Fig.5.3 is the percent improvement of the ESRL/RAP over the NCEP/RAP model. The NCEP_RAP assimilated a subset of the WFIP observations (one wind profiling radar and 5 sodars in the NSA; 3 sodars in the SSA; none of the tall towers, nacelle anemometers, or surface mesonet). Since these observations will have added some skill to the NCEP_RAP, the improvement of the ESRL_RAP relative to the NCEP_RAP model will provide a conservative estimate of what the improvement would have been had none of the WFIP observations been assimilated into the NCEP_RAP. This improvement peaks at 3-4% at forecast hour 1 and slowly decreases to near 1% out to hour 15.

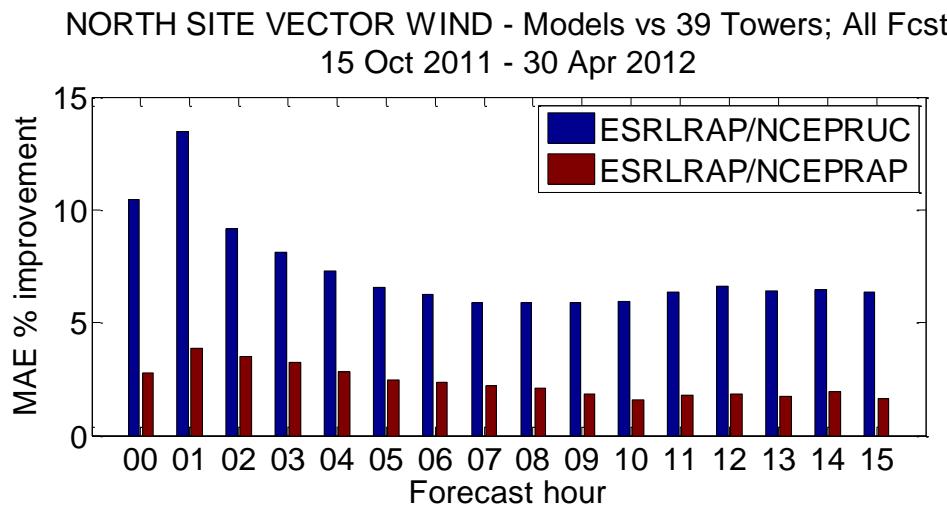


Figure 5.3. MAE percent improvement of the ESRL/RAP model over the NCEP/RUC model for the vector wind as a function of forecast length, calculated using observations from 39 real-time tall tower sites in the Northern Study Area (blue bars) during the first 6.5 months of the WFIP field campaign. Red bars indicate the same except for the ESRL/RAP over the NCEP/RAP.

Figure 5.4 also shows improvements in bulk statistics for the NSA, except in this case for wind power, calculated by converting wind speeds from both the model and tall tower observations to power using the power curve shown in Fig. 5.2. The top panel shows percent improvement for the coefficient of determination (correlation coefficient squared), and the lower panel the percent improvement in MAE.

Again, large improvements are found for the improvement in the ESRL/RAP over the NCEP/RUC model, with more modest improvements in the ESRL/RAP versus NCEP/RAP comparison.

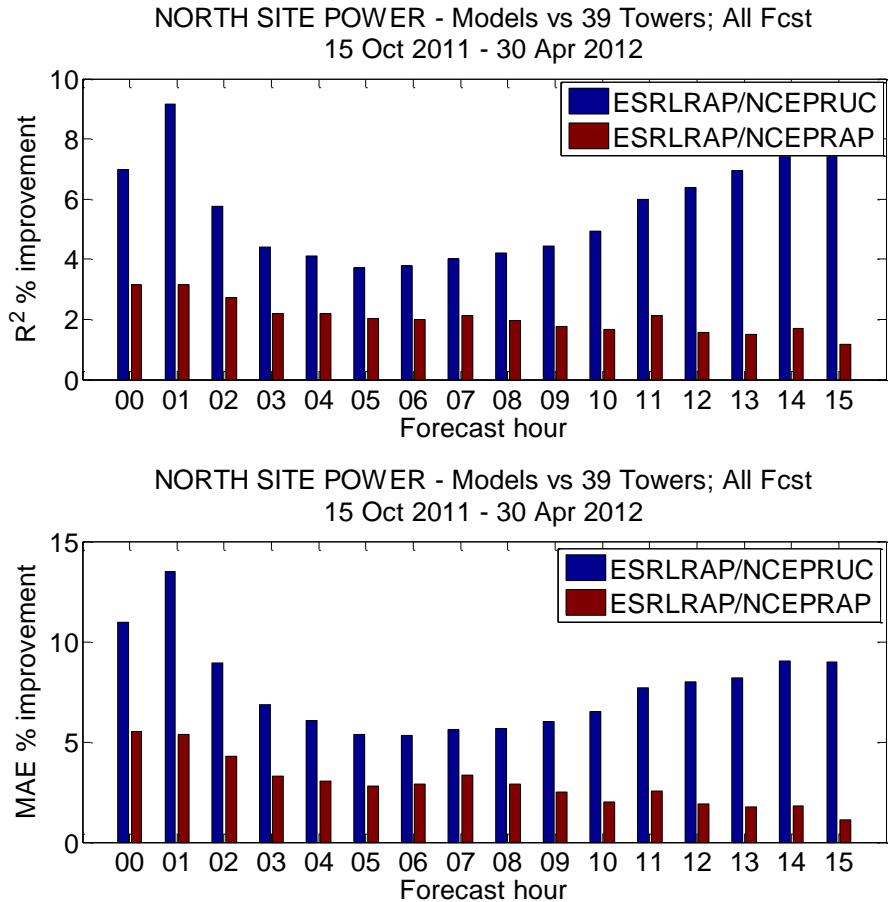


Fig. 5.4. The same as for Fig. 5.3, except for coefficient of determination R^2 and MAE percent improvement of wind power.

Figure 5.5 shows vector wind percent improvement statistics identical to Fig. 5.3 except for the SSA, using observations from 15 real-time ERCOT towers. The percent improvement in the ESRL/RAP versus the NCEP/RUC comparison again starts out large for short forecast lengths, then decreases to 2-3% by forecast lengths of 15 hours. The ESRL/RAP versus NCEP/RAP comparison shows near constant improvement of 2-4% at all forecast hours.

The improvements in power forecasts for the SSA are shown in Fig. 5.6. The improvement for the coefficient of determination R^2 for the ESRL/RAP to NCEP/RUC comparison is larger in the SSA than the NSA (Fig. 5.4), while the MAE and also the ESRL/RAP to NCEP/RAP improvements are similar in both domains.

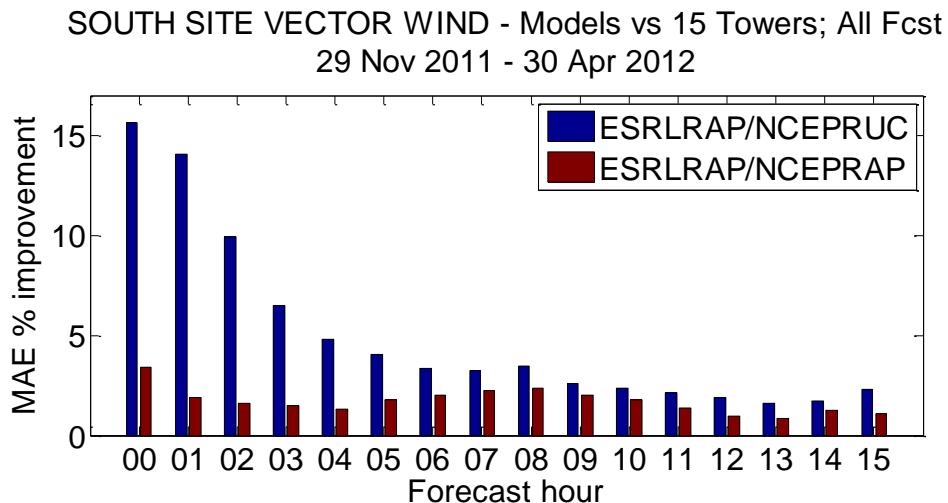


Figure 5.5. MAE percent improvement of the ESRL/RAP model over the NCEP/RUC model for the vector wind as a function of forecast length, calculated using observations from 15 real-time tall tower sites in the Southern Study Area (blue bars) during the first 6.5 months of the WFIP field campaign. Red bars indicate the same except for the ESRL/RAP over the NCEP/RAP.

The overall conclusions from this analysis are that: 1) a significant improvement in the NWS operational hourly-updated forecasts available at the start of WFIP was technically possible from a combination of research forecast models and additional observations; this improvement ranged from 15-4% for 1-6 hour hub-height wind and power forecasts of MAE and R^2 ; and 2) the switch of the NWS operational forecast model from the RUC to the RAP that occurred half-way through WFIP represented a significant improvement in operational forecast accuracy for the wind energy community.

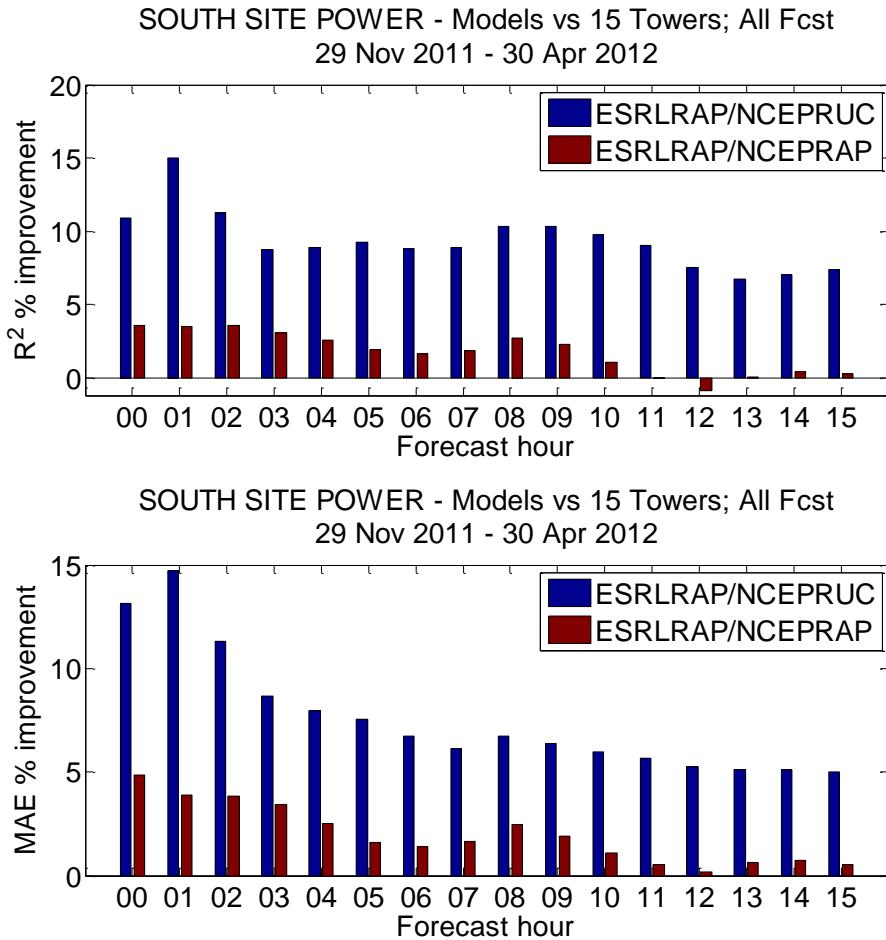


Fig. 5.6. The same as for Fig. 5.5, except for coefficient of determination R^2 and MAE percent improvement of wind power.

5.4. Bulk Error Statistics: ERSL RAP and HRRR

Next we show comparisons of the real-time ESRL RAP and HRRR models. These models are very similar, with the most significant differences being the higher 3 km resolution of the HRRR compared to the 13 km resolution of the RAP, and the fact that the HRRR uses only an explicit convection parameterization scheme. The ESRL/RAP assimilated the special real-time WFIP observations, and since the HRRR was initialized off of the ESRL/RAP, it was impacted by the same new observations. Standard RMSE bulk statistics are computed for the NSA and SSA using the real-time tall tower data that were available, and tend to show in general lower skill by up to 4-6% for the HRRR than for the RAP for most forecast hours (Fig. 5.7). Similar reductions in HRRR skill of 2-6% were found for MAE, RMSE, and R^2 for the scalar wind speed (not shown).

A reduction in bulk statistics forecast skill for a higher resolution model is often found in weather forecasting analysis, and can be explained by the fact that although the higher resolution HRRR model can more realistically simulate thunderstorms and other small scale convective atmospheric weather systems, small misplacements of these features in time or space will result in worse point evaluations of statistical skill than when a smoothly varying forecast from a coarse resolution model is used (e.g. Rife et al., 2004). Therefore one must exercise caution when comparing high resolution forecasts to comparatively lower resolution forecasts when using traditional metrics (e.g. RMSE). An approach to address this issue has very recently been introduced via a neighborhood approach and adopting probabilistic approaches to forecast verification at observing sites (Mittermaier, 2013). Such a forecast verification approach may be an interesting technique worth investigating in a future study. Finally, we note that although the bulk statistics do not show improvement from the HRRR relative to the RAP, the ramp statistics analysis contained in the WindLogics WFIP final report shows that the HRRR provides additional value over the RAP in predicting the frequency of ramp events.

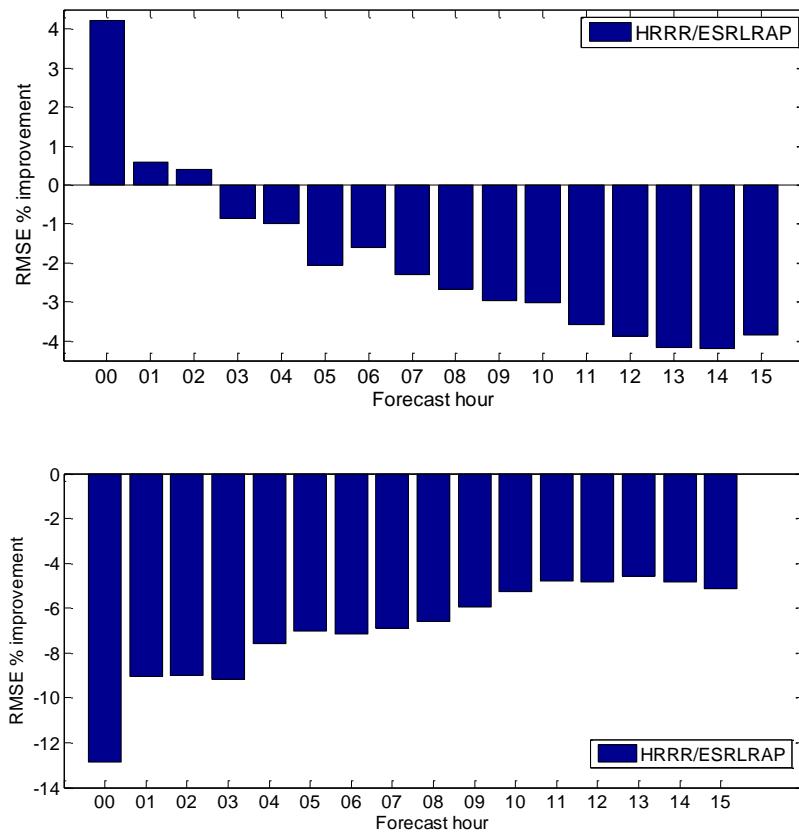


Figure 5.7. RMSE improvement of the vector wind for the HRRR model over the ESRL/RAP model, for the NSA from 15 Oct. 2011 – 30 Apr. 2012 (top panel), and the SSA from 29 Nov. 2011 – 30 Apr. 2012 (bottom panel), as a function of forecast length.

6. Data Denial Simulations

One of the primary goals of WFIP was to determine the impact of the special WFIP observations on model forecast skill of turbine hub-height winds. Isolating the impact of the new observations required carefully controlled data denial simulations, where the identical numerical weather prediction model was run twice: first a *control run* that assimilated only the routinely available observations, and second, an *experimental run* that assimilated both the routine and the special WFIP observations. Differences in forecast skill between these two simulations determine the impact that the special WFIP observations alone had on improving model forecast skill. Both ESRL and NCEP ran data denial simulations, ESRL using the RAP model and NCEP using the NAM and NAM CONUSnest. The ESRL HRRR model did not have its own data assimilation system, but was initialized using the ESRL RAP assimilation system at each hour. Therefore only ESRL RAP, NAM, and NAM CONUSnest simulations are utilized in the WFIP data impact analysis, and not the HRRR.

6.1. Observations assimilated

The special data that was assimilated into the NOAA models for these experimental simulations included in the northern study area vector winds and RASS temperatures from 9 WPR sites, vector winds from 5 sodars and 132 tall towers, and scalar wind speeds from 441 turbine nacelle anemometers (Table 6.1). In the southern study area the assimilated new observations included vector wind profiles from 3 WPR's two of which also provided RASS temperature profiles, vector winds from 7 sodars and 51 tall towers, and for the RAP model, mesonet near-surface vector winds, temperature and humidity, and pressure from 62 sites. In the NSA the nacelle scalar wind speeds were assimilated using the same technique used to assimilated satellite scatterometer scalar wind speeds over the ocean. All WFIP observations were quality controlled as described in Section 2.3 before the data was assimilated using assimilation parameter settings as described in Section 4.4.

	WPR vector winds	WPR-RASS temperatures	Sodar vector winds	Tall tower vector winds	Nacelle speeds	Surface mesonet Vector winds, T q, p
Northern Study Area	9	9	5	132	441	0
Southern Study Area	3	2	7	51	0	62

Table 6.1 Data types and quantities assimilated in the data denial simulation experiments for both the Northern and Southern Study Areas.

6.2. Data denial simulation dates

Because of limitations in computing resources, data denial simulations were run for only a limited subset of days from the WFIP field campaign. The intent in selecting these days was to get a distribution through all four seasons of the year, and also to select days that were of meteorological interest to the private sector partners in both the Northern and Southern Study Areas. Six separate data denial episodes were chosen, ranging in length from 7 to 12 days, for a total of 55 days (Table 6.2). These episodes were selected based on the presence of ramp events in the observations and forecasts, the occurrence of challenges to grid operators from wind power variations, and the availability of the special WFIP observations (i.e., if possible avoiding periods when instruments were off-line due to icing or other conditions).

Episode 1	30 Nov – 6 Dec 2011	7 days
Episode 2	07 Jan – 15 Jan 2012	9 days
Episode 3	14 Apr – 25 Apr 2012	12 days
Episode 4	09 Jun – 17 Jun 2012	9 days
Episode 5	16 Sep – 25 Sep 2011	10 days
Episode 6	13 Oct - 20 Oct 2011	8 days

Table 6.2 Dates for six data denial studies.

6.3. Model bias estimation

The different types and numbers of instruments deployed during WFIP allows for a detailed determination of model bias. This estimate will help inform the direction of future improvements to the model, highlight potential instrumental problems, and also will be useful in determining the types of bias-correction methods to be considered for calculating improvements in model skill from assimilating the observations. Data for the bias analysis will be restricted to the 55 days used for the data denial simulation, as these days had observations with the highest level of data QC applied. The bias analysis for the most part is shown only for the data denial control runs from the ESRL RAP model, which did not assimilate any of the special WFIP observations, largely because the bias does not change dramatically between the control and experimental simulations. However, to illustrate this point, biases for the control and experimental simulations are shown for the tall tower observations.

Wind profiling radars

The wind profiler biases as a function of height for each of the 6 separate DD episodes are shown in Fig.6.1, for all 12 sites, and then separately for the NSA and SSA. The bias in the NSA is almost always positive (the model speed greater than observed), and follows a distinctive pattern with the largest bias of about 1.5 ms^{-1} occurring in the lowest levels, then decreasing in the layer between 500-1500m to about $+0.5 \text{ ms}^{-1}$. In contrast the bias in the SSA is close to zero in all of the DD episodes except December. For the NSA the largest biases occur in the cold season months (Jan., Dec., and Oct.). This suggests that at least part of the bias may be due to residual clutter and RFI, which tend to be worse during the colder winter months when the atmospheric reflectivity is weaker.

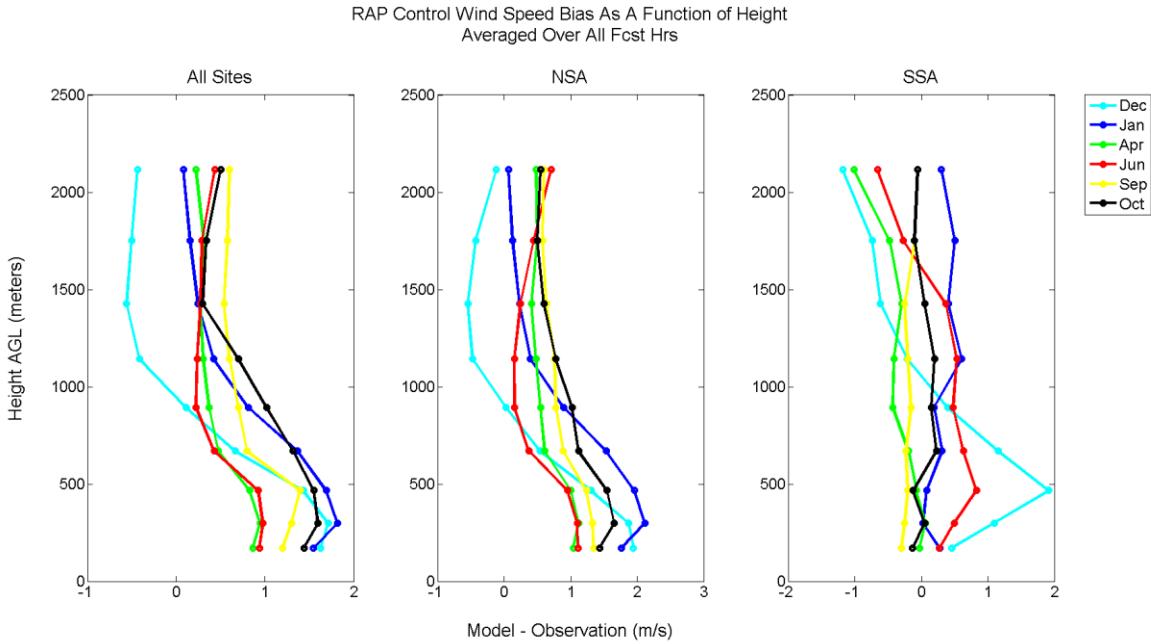


Figure 6.1. RAP control simulation wind profiler radar biases (model-observation) as a function of height, for each of the 6 DD episodes, averaged for all 15 forecast hours. The left panel is for all 12 WPR's, the middle panel is for the 9 NSA profilers, and the right panel is for the 3 SSA WPR's.

The WPR bias dependence on forecast hour is shown in Fig. 6.2 for both the NSA and SSA. The biases here are layer averages from 0-500m, and all 6 DD episodes are averaged together. The bias in the SSA is slightly negative for hours 00 – 01, turning positive with a value near $+0.5 \text{ ms}^{-1}$ for hours 04-15. In contrast, the bias for the NSA starts positive and becomes increasingly positive with each forecast hour.

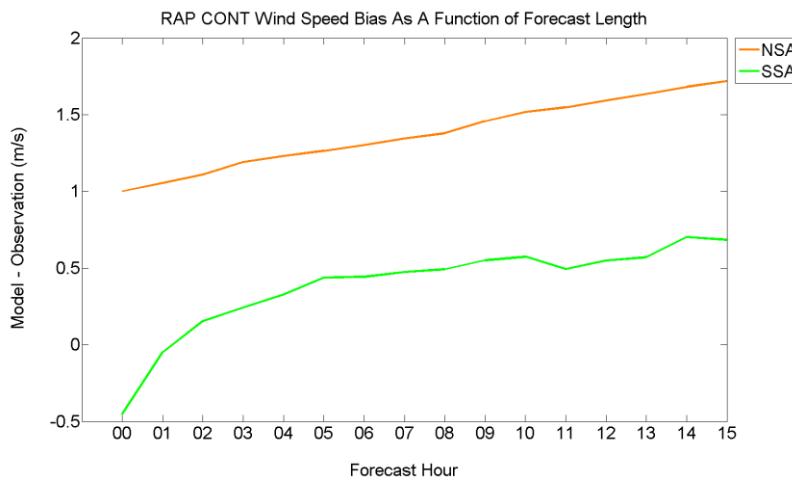


Figure 6.2. RAP control bias calculated using the WPR observations, as a function of length of forecast, averaged for all 6 DD episodes and over the layer 0-500m AGL, for the NSA (orange) and SSA (green).

To investigate any possible diurnal variation in the bias, we plot the bias as a function of verification hour (0-23 UTC) in Fig. 6.3, averaging the data into the cold season (Dec., Jan., and Oct.) and warm season (Apr., Jun., and Sept.) episodes, again averaged over the lowest 500m AGL. For the cold season the biases at all forecast lengths tend to be fairly uniform across the time of day, while for the warm season the biases in both the NSA and SSA are reduced during the daytime hours between 16-04 UTC (11-23 CST). This suggests that the presence of a deep, convective boundary layer reduces the magnitude of the wind speed bias.

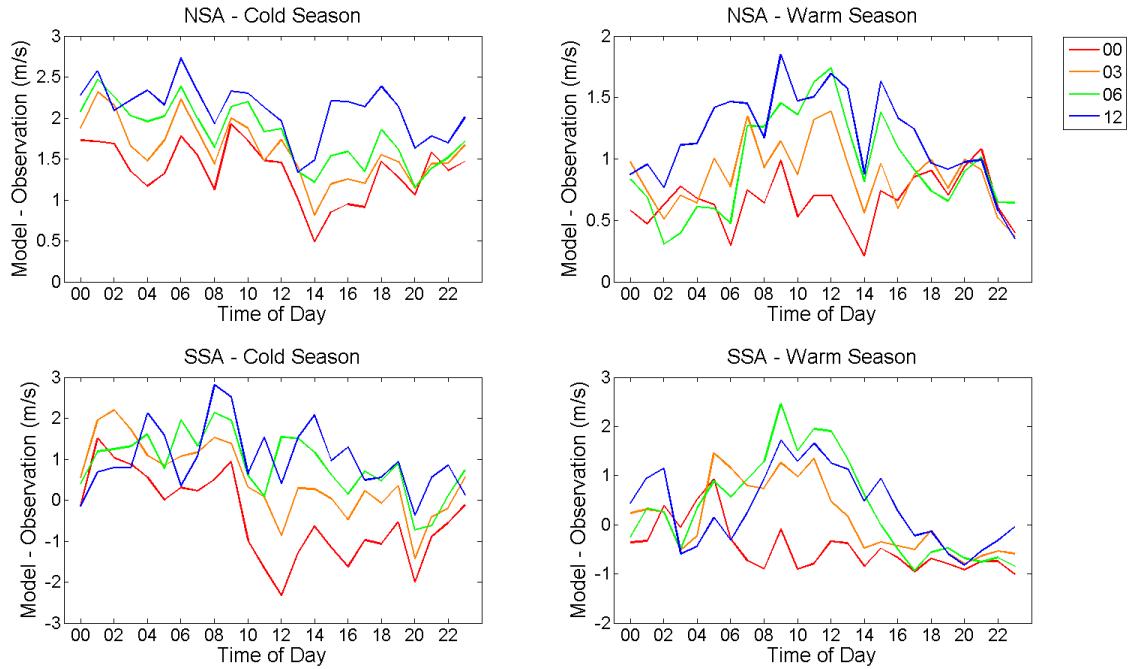


Figure 6.3. RAP control bias using the wind profiler speed observations, as a function of forecast verification time (UTC), for all 6 DD experiments, averaged over the layer 0-500m AGL. Individual curves show forecasts lengths of 0, 3, 6, and 12 hours.

Lastly, we consider the WPR bias as a function of wind speed. The bias is computed for 3 ms^{-1} wind speed intervals ($0-3 \text{ ms}^{-1}$, $3-6 \text{ ms}^{-1}$, etc. out to $18-21 \text{ ms}^{-1}$), again averaged over 0-500m, and averaged over all 6 DD episodes. The bias is a very strong function of wind speed and follows the same pattern in the NSA and SSA, changing by $3-4 \text{ ms}^{-1}$ over the range of wind speeds.

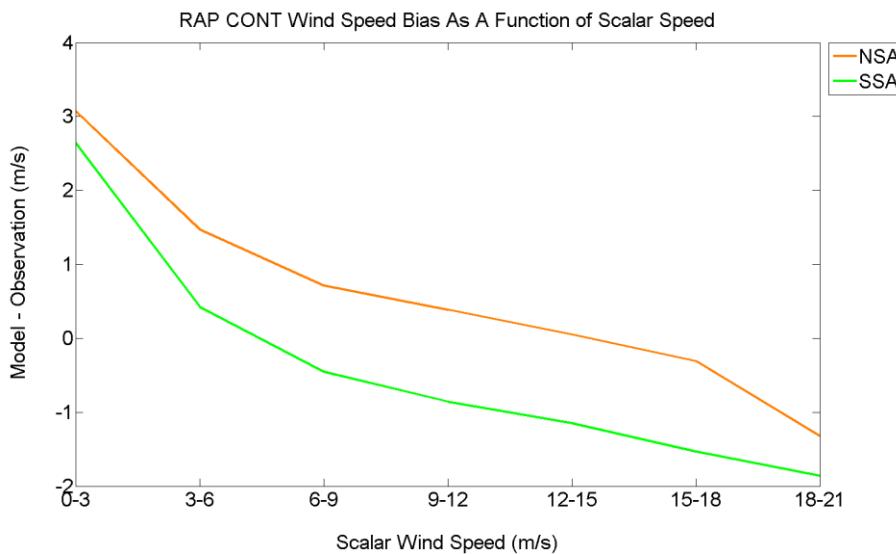


Figure 6.4 RAP control bias at forecast hour 00 calculated using the wind profiling radar observations, as a function of observed wind speed, averaged over the lowest 500m AGL and over all 6 DD episodes, for the NSA (orange) and SSA (green).

Sodars

The bias analysis is now repeated using the 12 sodars deployed during WFIP. Figure 6.5 displays the sodar biases as a function of height for each of the 6 separate DD episodes, for all 12 sites, and then separately for the NSA and SSA. The sodar bias for the NSA averages approximately -0.35 ms^{-1} for the NSA, and $+0.5 \text{ ms}^{-1}$ for the SSA. The NSA bias is nearly constant with height, whereas the SSA bias is small at the lowest level, increases at 80m, then decreases again at 160m. No clear seasonal pattern is present in either the NSA or SSA.

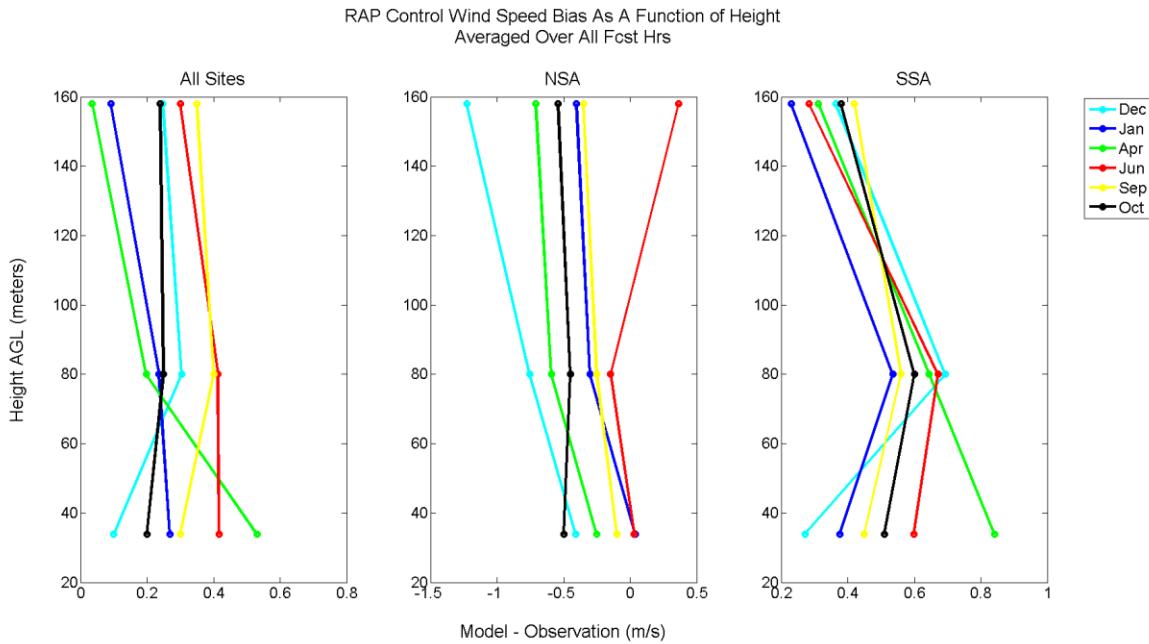


Figure 6.5. RAP control simulation biases as a function of height, using the sodars for verification, for each of the 6 DD episodes, averaged for all 15 forecast hours. The left panel is for all 12 sodars, the middle panel is for the 5 NSA sodars, and the right panel is for the 7 SSA sodars.

Next, the sodar bias dependence on forecast length is evaluated, using the average 0-200m sodar bias from the 3 lowest model levels, and averaging all 6 DD episodes together (Fig. 6.6). In both study areas the bias starts off at hour 00 at its most negative value, and increases with forecast length. The NSA bias is small and negative for most forecast hours, while the NSA bias is significantly positive, in agreement with Fig. 6.5.

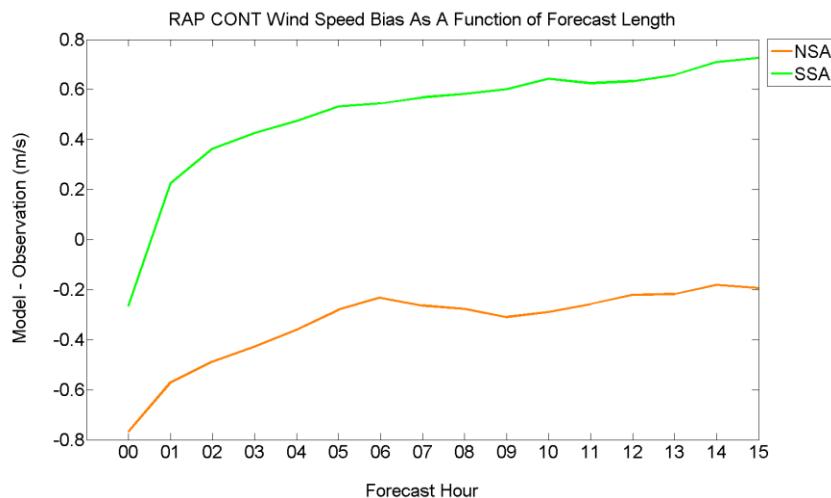


Figure 6.6. RAP control bias calculated using the sodar observations, as a function of length of forecast, averaged for all 6 DD episodes and over the layer 0-200m AGL, for the NSA (orange) and SSA (green).

The RAP-sodar bias as a function of forecast verification hour is shown in Fig. 6.7. The data are averaged into the cold season (Dec. Jan. and Oct.) and warm season (Apr., Jun., and Sept.) episodes, and again are averaged over the lowest 200m AGL.

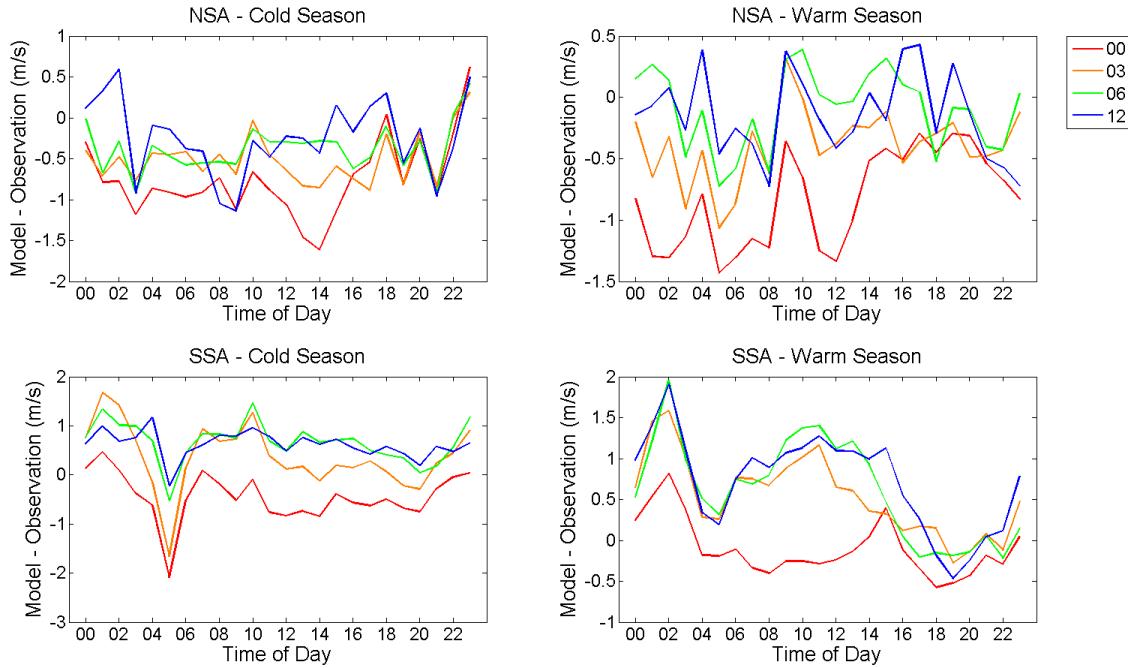


Figure 6.7. RAP control bias using the sodar speed observations, as a function of forecast verification time (UTC), for all 6 DD experiments, averaged over the layer 0-200m AGL. Individual curves show forecasts lengths of 0, 3, 6, and 12 hours.

The RAP control bias using the sodar speed observations as a function of the observed speed is shown in Fig. 6.8 for the NSA and SSA. The bias is remarkably similar in both the NSA and SSA for most speed bins. Similar to the WPR bias, the sodar bias is a strong function of wind speed, being approximately $+1 \text{ ms}^{-1}$ for small observed speeds, decreasing nearly linearly to near -2 ms^{-1} for the $15\text{-}18 \text{ ms}^{-1}$ bin. The largest wind speed bin has a worse bias in the SSA, but there are few observed values in this speed range.

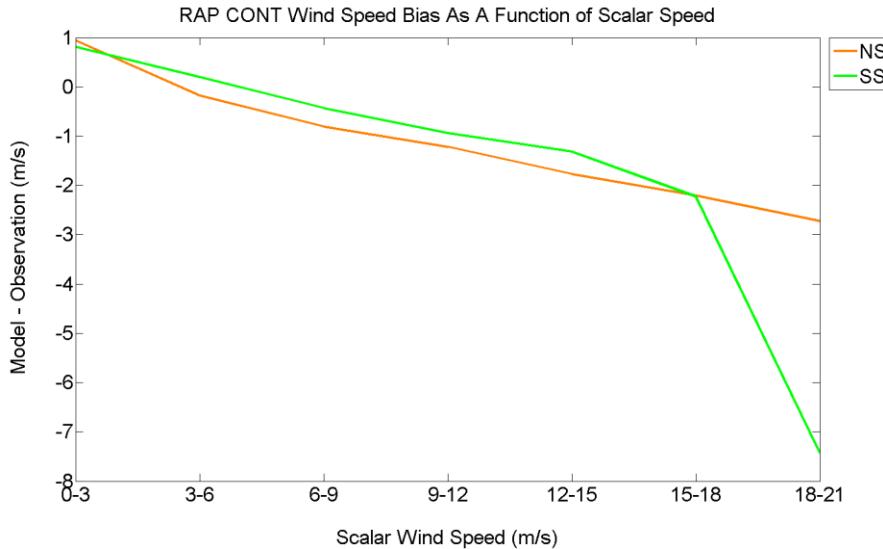


Figure 6.8 RAP control simulation wind speed bias at forecast hour 00 determined using the sodar observations, as a function of the observed wind speed, averaged over all 6 DD episodes, for the NSA (orange) and SSA (green).

Tall towers

A similar bias analysis was also carried out for the tall tower observations used during WFIP. Figure 6.9 displays the bias as a function of forecast length, for the NSA (orange) and SSA (green), and for the RAP control (solid lines) and experimental (dotted lines). The RAP bias when using the tall tower observations is negative for both the NSA and SSA, but worse in the NSA. The bias is largest at the initialization time, rapidly reaches a plateau from 01-12 hours, and then slightly increases in the last several forecast hours.

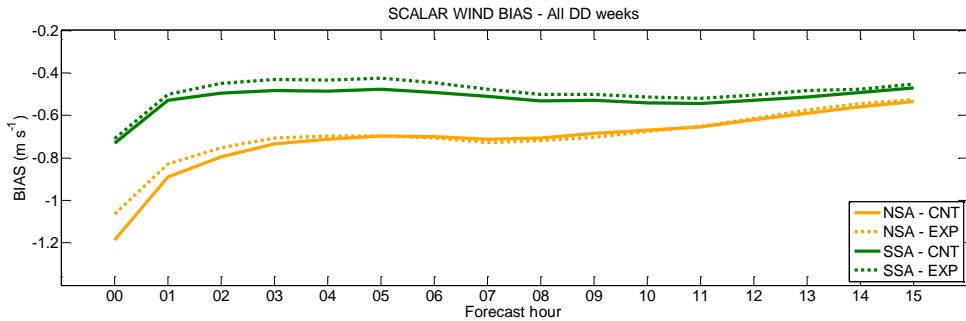


Figure 6.9. RAP bias using the tall tower observations, as a function of length of forecast, averaged for all 6 DD episodes, for the NSA (orange) and SSA (green), and for the control (solid lines) and experimental (dotted lines) simulations.

The bias as a function of season is shown in Fig. 6.10 for both the NSA and SSA. Here some seasonal trend is found in the NSA, with June and September having the smallest biases, while December, October, and especially January have the largest biases. The SSA biases also vary considerably from

episode to episode, with January and December having the opposite extremes. The mostly random episode-to-episode variation in the bias suggests that the bias has a significant day-to-day flow-dependent component, and that one-week averaging periods for each episode are too short for this variability to reach an equilibrium value.

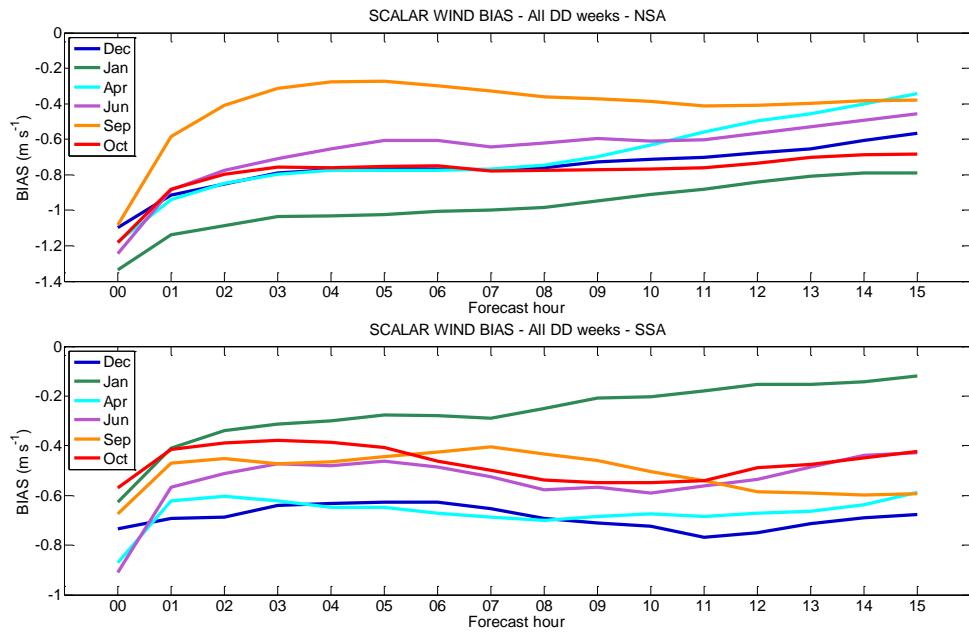


Figure 6.10. As in Fig. 6.9, except showing the biases for individual DD episodes. The top panel is for the NSA, the bottom panel for the SSA.

The RAP-tall tower speed control bias as a function of validation hour is shown in Fig.6.11. Here the biases are simultaneously shown as a function of the validation hour and the forecast length. The cold and warm season patterns are quite similar in the NSA except for an offset of less negative biases in the warm season. In both the NSA and SSA the bias oscillates with a 12 hour period, with two clear maxima and two minima. For moderate forecast lengths (02-09 hours) the largest (most negative) biases occur during the nighttime (03-10 UTC; 22-05 CST) and afternoon (18-22 UTC; 13-17 UTC) hours. In the SSA, during both the cold and warm seasons the bias is also large during the afternoon hours (17-22 UTC; 12-17 UTC), and in the warm season the second period of large bias in the nighttime hours (05-08 UTC; 00-03 CST) is also present. Also, the variation of the bias versus forecast hour verification time and length of forecast are broadly consistent with those found for the sodars in Fig.6.7, including more negative biases present at the model initialization time, and the biases being more negative in the NSA than the SSA. The fact that the biases are most negative at the model initialization time (especially in the NSA) suggests that this may not be a result of inaccurate model physical parameterizations, but may be related to the data assimilation procedure.

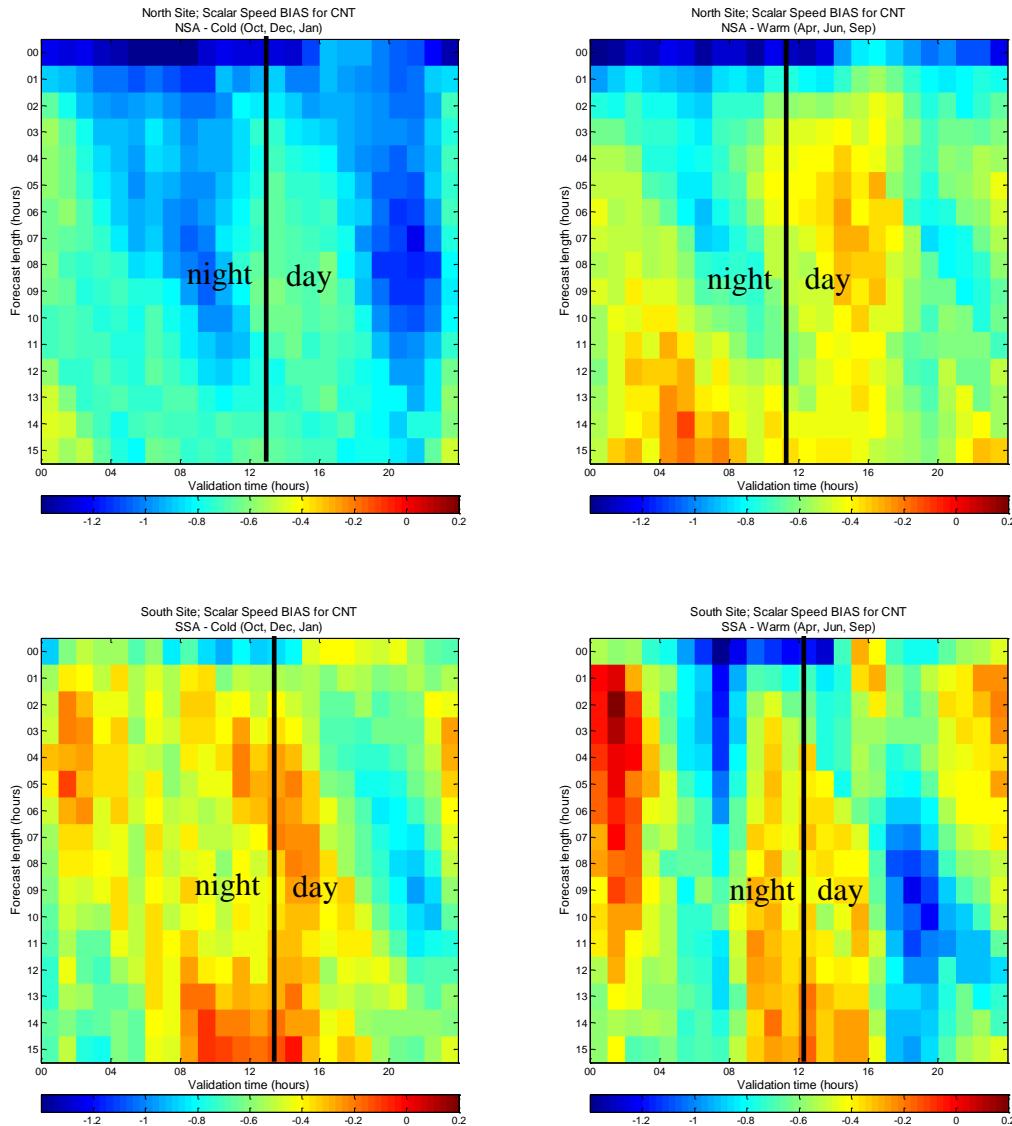


Figure 6.11. RAP control speed bias calculated using the tall tower observations, as a function of forecast verification time (UTC, x-axis), and forecast length (y-axis), averaged for all 6 DD experiments. The top panels are for the NSA, bottom for the SSA, left panels for the cold season, right panels for the warm season.

Finally, the RAP control bias using the tall tower observations is shown as a function of observed wind speed (Fig. 6.12), for forecast hours 00, 03 and 06. Similar to the wind profiler and sodar derived biases, the bias starts out positive for small wind speeds, and decreases nearly linearly to values of -3 ms^{-1} at speeds of approximately $15\text{-}18 \text{ ms}^{-1}$, after which it decreases even faster.

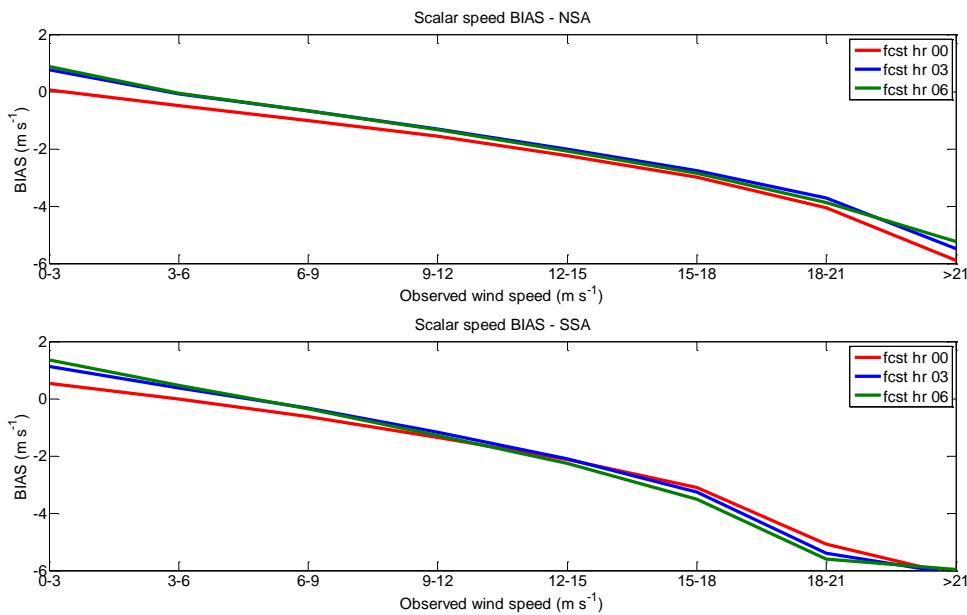


Figure 6.12 RAP control bias using the tall tower observations as a function of observed wind speed, averaged over all 6 DD episodes, for the NSA (top) and SSA (bottom), for three different forecast lengths.

Bias evaluation synopsis

The speed biases in the RAP control simulations have been found to be quite large, with many similarities and some differences across type of observation platform (WPR's, sodars, or tall towers) and between the NSA and SSA. Similarities in biases across the instrument platforms indicate the presence of a real bias in the model (unless all three instruments suffer from the same bias errors). Conversely, differences between the biases when using the different instrument types indicate problems with one or more of the instrument types.

The most prominent similarity found in the biases for all of the instrument platforms is a speed-dependent bias, where the model bias becomes increasingly negative as the speed increases. The speed dependent bias appears to be similar in both the NSA and SSA, and very approximately would be expressed as: model bias = $1.0 \text{ ms}^{-1} - 0.2 * (\text{observed speed})$.

The wind profiling radars in the NSA have an apparent low speed bias in the lowest 500m of approximately 1.3 ms^{-1} . This bias can be partitioned into an approximately 1.0 ms^{-1} bias for speeds greater than 3 ms^{-1} , and 1.5 ms^{-1} for speeds less than 3 ms^{-1} . A low speed bias can be caused by clutter or RFI, and suggests that despite efforts to QC these effects, some residual errors are still present in the profiler observations.

A comparison of the sodar and tall tower observations suggests that either the sodars have a low speed bias of approximately -1 ms^{-1} at low wind speeds and a high speed bias of $+1 \text{ ms}^{-1}$ at high speeds, or the

towers have a high speed bias of $+1.0 \text{ ms}^{-1}$ for low speeds and a low speed bias -1.0 ms^{-1} for high speeds, or a combination of both effects is present (with smaller magnitudes for each).

Comparison of these bias estimates is of course limited by the fact that the various observations are not all co-located. In particular, none of the tall tower observations are co-located with either sodars or WPR's. A precise determination of the model and instrument biases would require co-located sensors at many sites, as the model bias at any single grid point may not be representative of the model bias as a whole.

6.4. Wind profiler evaluation

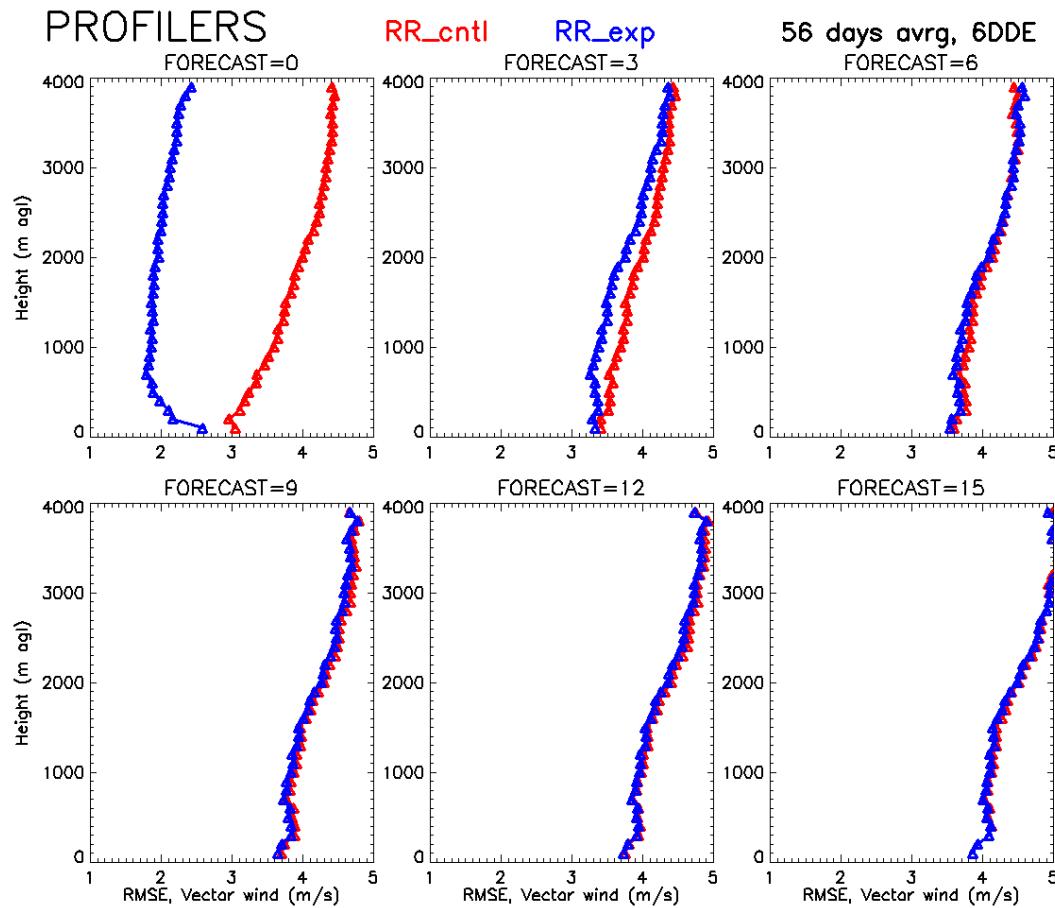
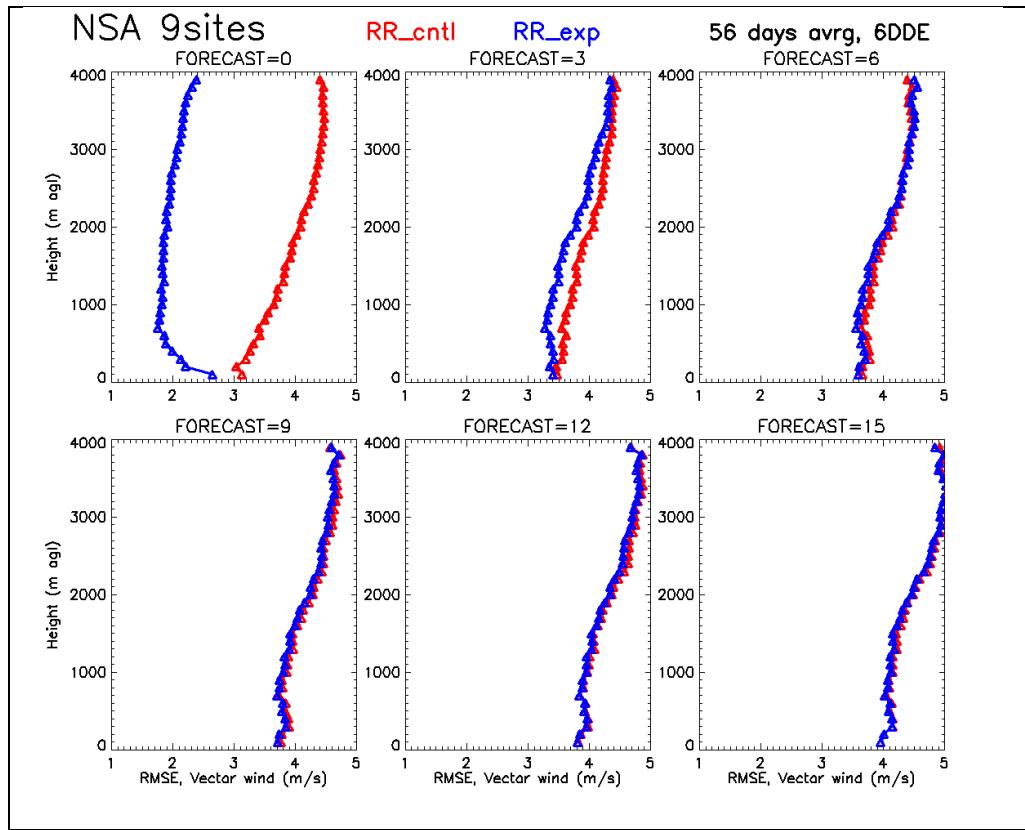


Fig. 6.13. Vertical profiles of vector wind RMSE averaged over all 12 WPR sites and all 55 DD RAP simulation days, at the model initialization time and three hour forecast length increments. Red is for the control simulations and blue is the experimental simulations that assimilate the special observations.

The RMSE of the vector winds (Fig. 6.13) for the control simulations almost always increases with height, perhaps because the relative paucity of upper atmosphere observations leaves the model initial fields

with larger errors than near the surface, where more numerous routine observations exist. The difference between the vector wind RMSE for the control and experimental simulations is largest at forecast hour 0 (the initialization time) and becomes smaller with the length of the forecast. At forecast hour 0 the RMSE for the experimental simulations is smaller than the control by as much as 2.0 ms^{-1} at 2000 m AGL. This reduction in the experimental run RMSE relative to the control is less pronounced at lower heights, especially in the lowest 500m, which results from the model initialization also trying to fit other WFIP observations in this layer. At forecast hour 03, a reduction in RMSE in the experimental simulation exists up to 4 km AGL.



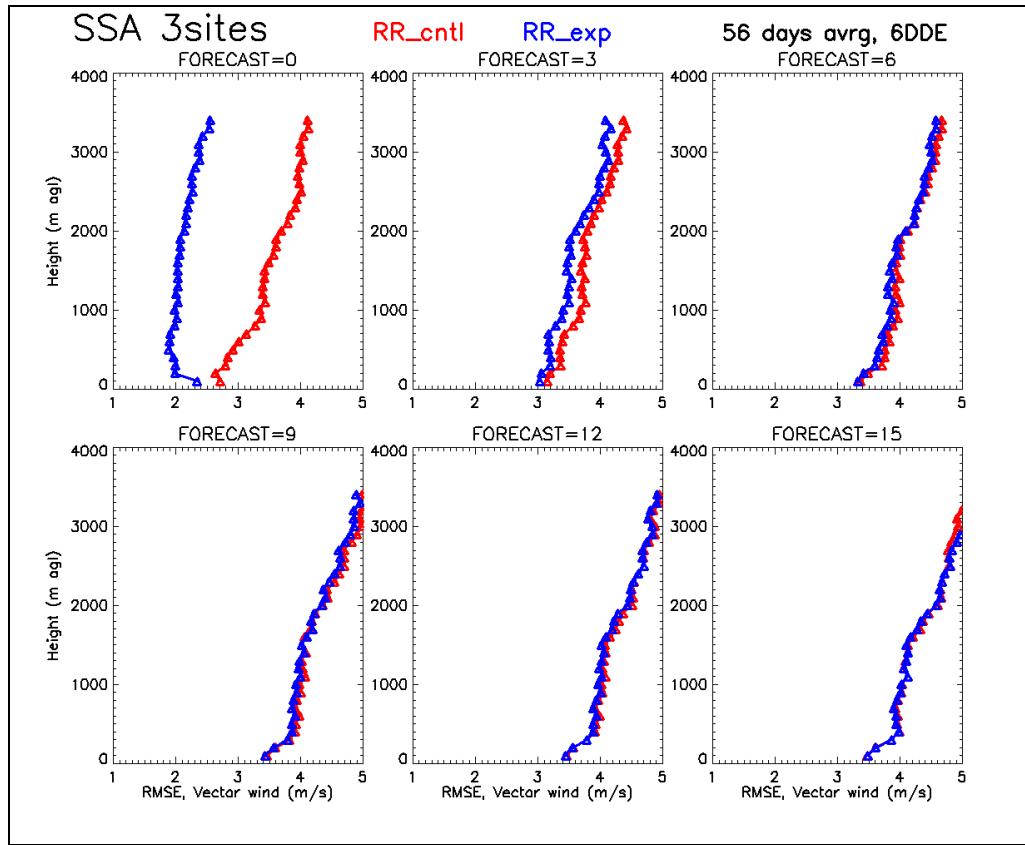


Fig. 6.14. As in Fig. 6.13 for the 9 NSA sites (top set of panels) and 3 SSA sites (bottom set of panels).

Vertical profiles of RMSE for the NSA and SSA are shown separately in Fig. 6.14. In general the behavior of the two domains is qualitatively similar, although the RMSE is somewhat larger in the NSA than the SSA.

RMSE vertically averaged over the lowest 2 km AGL and over the 55 DD episode days is shown in Fig. 6.15 again for the control (red) and experimental (blue) simulations, for both the vector wind (top 4 panels) and scalar wind speed (bottom 4 panels). The MAE difference between the control and experimental simulations is shown by the black curves, with 95% confidence intervals indicated. Error bars represent the 95% confidence intervals defined as $(\pm 1.96 \sigma / \sqrt{n'})$, where n' is the effective number of samples determined from the one-sample time-lagged autocorrelation r_1 , with

$$n' = n \frac{(1 - r_1)}{(1 + r_1)}$$

The improvement is largest at the initialization time, becoming insignificant beyond forecast hours 8-9, and is larger for the vector wind than for the scalar wind speed, indicating an improvement in wind direction exists as well as speed.

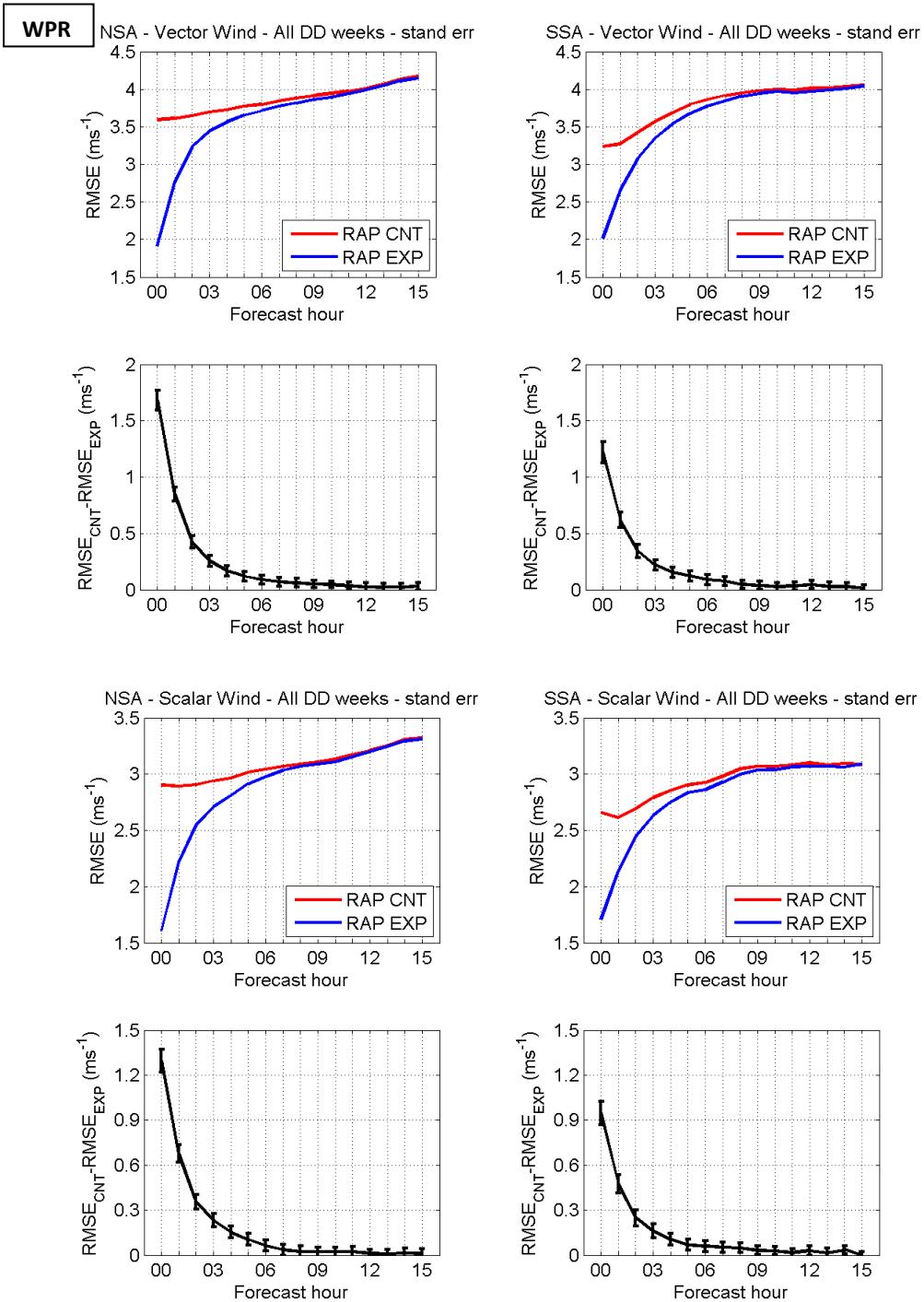


Fig 6.15. Wind profiling radar layer averaged (0-2000m AGL) RMSE, averaged for the 6 control DD episode simulations (red) and the experimental DD simulations (blue). Top 4 panels are vector wind, and bottom 4 panels are scalar wind speed. Left panels are average of 9 NSA profilers; right panels average of 3 SSA profilers. Panels with black curves show difference between control and experimental simulation RMSE's in the corresponding panel above, and error bars indicate 95% confidence intervals.

6.5. Sodar evaluation.

The improvement in RMSE between the experimental and control DD simulations evaluated using the sodar observations in the lowest 200m is shown in Fig. 6.16. The sodar data was interpolated to the exact heights of the 3 model levels below 200m (approximately 30m, 80m, and 180m), RMSE's were calculated at each of these levels, and then averaged. The magnitudes of the sodar RMSE control run values at hour 00 and hour 06 from Fig. 6.16 compare favorably with those from the wind profilers lowest range gate (Fig. 6.13). In the SSA the sodar and profiler RMSE are nearly identical (both approximately 2.7 ms^{-1} at hour 00 and 3.3 ms^{-1} at hour 06), while in the NSA the wind profiler values (3.2 and 3.6 ms^{-1} at hours 00 and 06) are slightly larger than those for the sodars (2.7 and 3.3 ms^{-1}). We interpret this to mean that the first range gates of the wind profilers in the NSA had somewhat lower accuracy than the sodars, while in the south the two had very similar accuracies. We also note that the control simulation has lower accuracy in the SSA than the NSA, especially for forecast lengths greater than 3 hours.

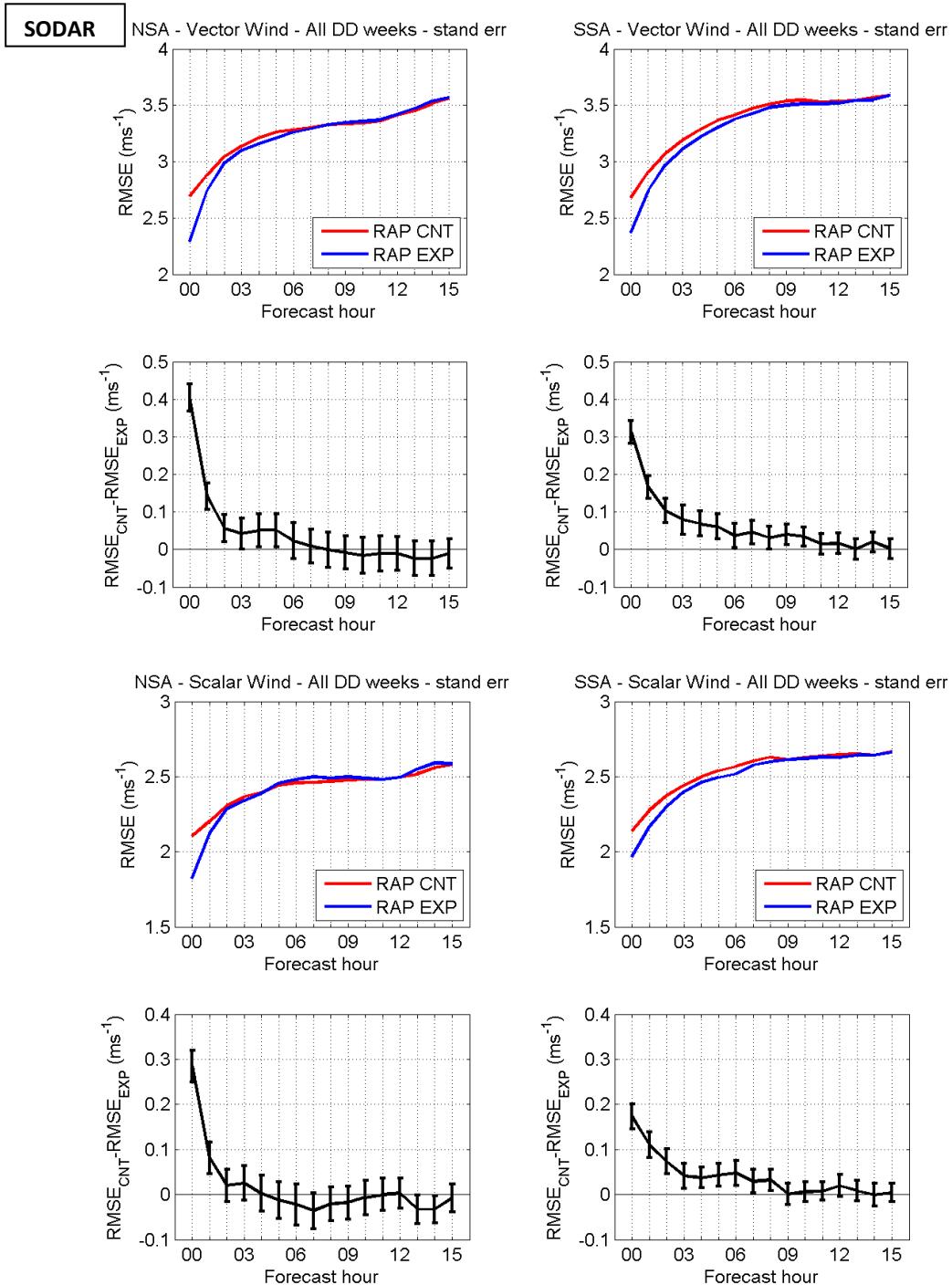


Figure 6.16. The same as Fig. 6.15 except using sodar observations, with the layer average between 0-200m.

6.6. Tall tower evaluation

Next the tall tower data sets are used to evaluate the impact of assimilation of the new WFIP observations on the ESRL RAP forecasts. Private forecasting companies almost always apply some type of bias correction to their wind power forecasts, some of which can be quite complicated, such as machine learning algorithms. We do not want to duplicate the complex bias correction schemes that the private forecasting sector has developed, but to first order want to understand if the improvement from the assimilation of the WFIP observations is dependent on the choice of the bias correction scheme. To this end we first investigate the dependence of the improvement on several simple types of bias correction schemes. We then investigate how the forecast improvement depends on forecast length, season, forecast verification time, and wind speed, for both the NSA and SSA. We also investigate how the forecast improvement varies in time hour-by-hour in one of the DD simulations, and evaluate the data's impact on rare large errors. Finally we investigate the sensitivity of the results to the geographic position of towers, checking if tower sites far from the main body of the WPR's, sodars, and other tall towers have less skill.

6.6.1. Bias correction sensitivity

Three different bias-correction methods are evaluated, where in all cases the corrected forecast is the raw forecast minus a calculated bias. The first bias is simply the average wind speed calculated independently for each of the 15 forecast hours over an entire DD simulation at each tower, minus the observed wind speeds at the same times, referred to as the mean bias correction. The second method takes the additional step of calculating separate biases for each hour of the diurnal cycle, again separately for each of the 15 forecast hours. The third method separates the biases according to wind speed, using binned intervals of 3ms^{-1} , again separately for each of the 15 forecast hours. After the forecasts are corrected for the speed bias, the speeds are then converted into power forecasts as described in Section 5.2, using a standard turbine power curve.

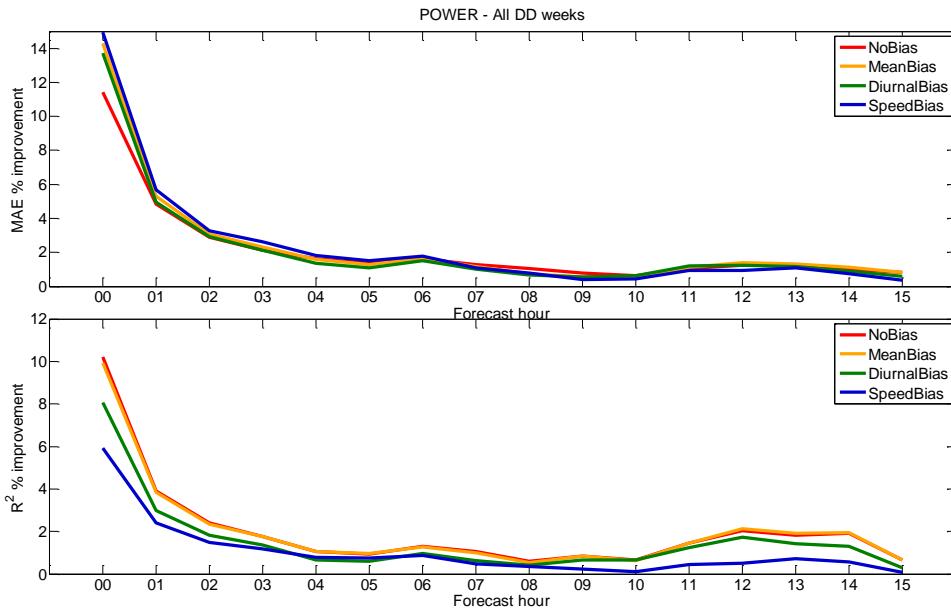


Figure 6.17. MAE and R^2 percent improvement of the experimental RAP simulation assimilating the special WFIP observations over a control that does assimilate the WFIP observations, for no bias correction, and the mean, diurnal, and speed dependent bias corrections. The calculations are done using the 55 DD episode days, northern and southern study areas combined.

Although the choice of bias correction method can have a significant impact on forecast skill, the various methods have a much smaller impact on the relative improvement of the experimental simulations over the controls, as the same technique is applied to both. This is seen in Fig. 6.17, which shows percent improvement for MAE and R^2 for the average of all 6 DD episodes and the NSA and SSA combined. For MAE, at the initialization time (forecast hour 00), using no bias correction leads to a slightly smaller improvement, but at all other forecast hours the MAE differences are negligible. A greater dependence on bias correction method is found for R^2 , with somewhat smaller improvements found for the diurnal cycle and speed bias corrections. We have chosen to use the simple mean bias correction for the remainder of the analysis, knowing that it does not significantly alter the MAE improvement statistics, but may have some effect on the R^2 statistics.

6.6.2. NSA/SSA & Forecast length

The impacts of assimilating the new WFIP observations are broken out separately for the NSA and SSA in figs. 6.18-20 for the vector wind and power. Figure 6.18 shows MAE for the vector wind (top two panels) and power (5th and 6th panels) for the control (red curves) and experimental (blue curves) simulations, as a function of forecast length, averaged over all six data denial episodes, using all tall tower sites for verification. For power, the MAE is expressed as a percent of the maximum wind power capable of being generated (the rated power). The SSA has higher values of MAE (for both vector wind and power) than the NSA, perhaps due to more prevalent low-level jets, the presence of complex terrain

(many of the plants are on mesa tops), and possibly more frequent convection. As seen previously when using the radar wind profilers and sodars for verification (Figs. 6.15 and 6.16), the MAE reduction for the experimental simulations is largest at the initialization time and this reduction becomes smaller at longer forecast hours. The MAE difference between the control and experimental simulations is shown by the black curves, with 95% confidence intervals indicated. The MAE difference at the model initialization time (hour 00) is similar between the NSA and SSA for both the vector wind and power, but stays positive at a statistically significant level for longer time in the NSA. For power, the positive improvement is statistically significant through forecast hour 07 for the NSA, and though forecast hour 03 for the SSA.

Figure 6.19 expresses the increase in vector wind forecast skill as an MAE percent improvement. The MAE improvement at the initialization time is large in both areas (16% in the NSA and 14% in the SSA), reflecting the degree to which the GSI data assimilation scheme is able to better fit the tower observations. The improvement then decreases fairly rapidly in the next few forecast hours, reaching an approximately 3% improvement at forecast hour 06 in the NSA. The percent improvement is larger in the NSA than the SSA at all forecast hours. The greater magnitude and longer duration of the positive improvement in the NSA is likely due to the fact that there were more observations assimilated in the NSA, and they spanned a larger geographic footprint than in the SSA.

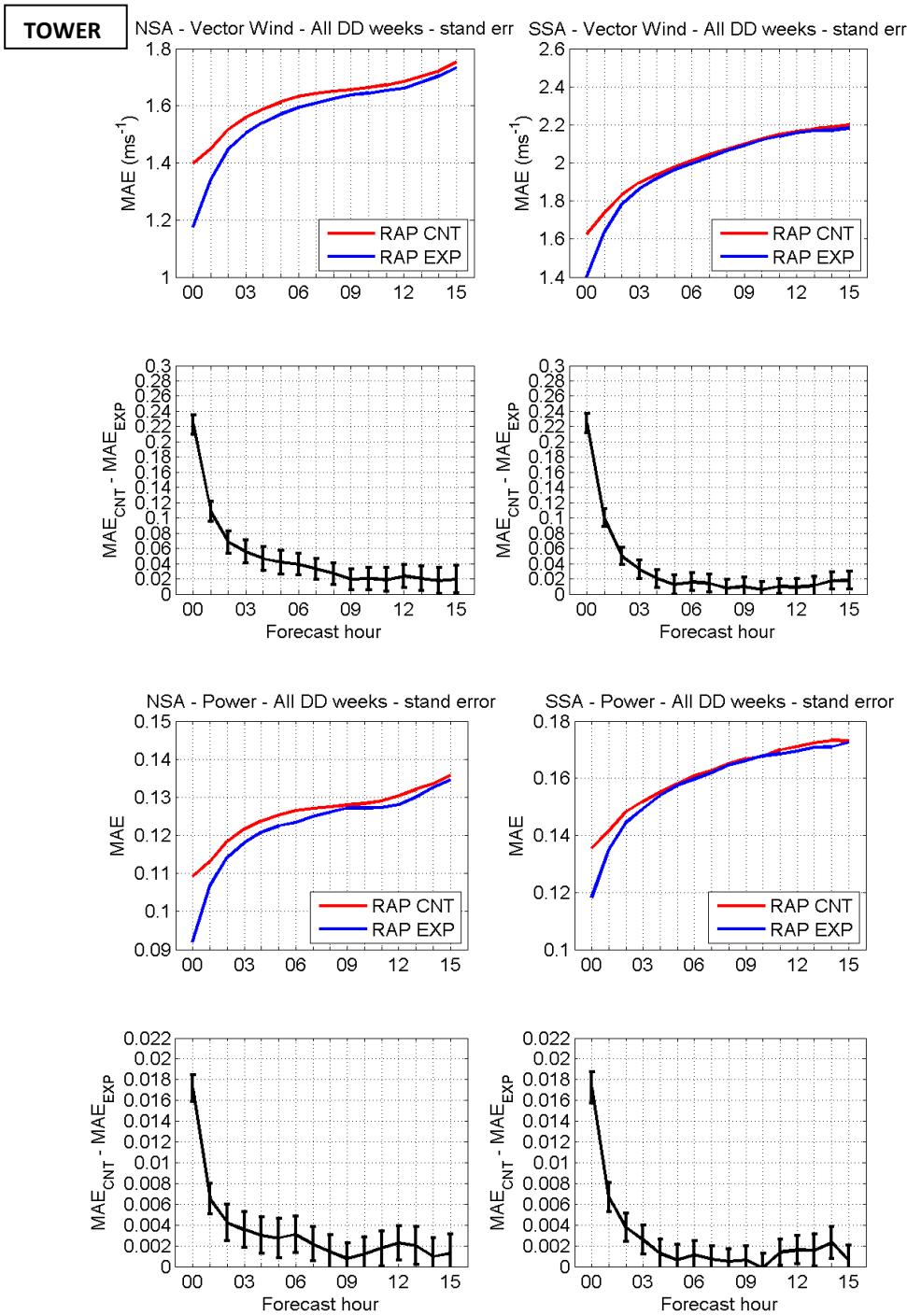


Figure 6.18 RAP tall tower-derived RMSE, averaged for the 6 control DD episode simulations (red) and the experimental DD simulations (blue). Top 4 panels are for vector wind, and bottom 4 panels are for power. Left panels are for the 9 NSA and right panels are for the SSA. Panels with black curves show difference between control and experimental simulation RMSE's in the corresponding panel above, and error bars indicate 95% confidence intervals.

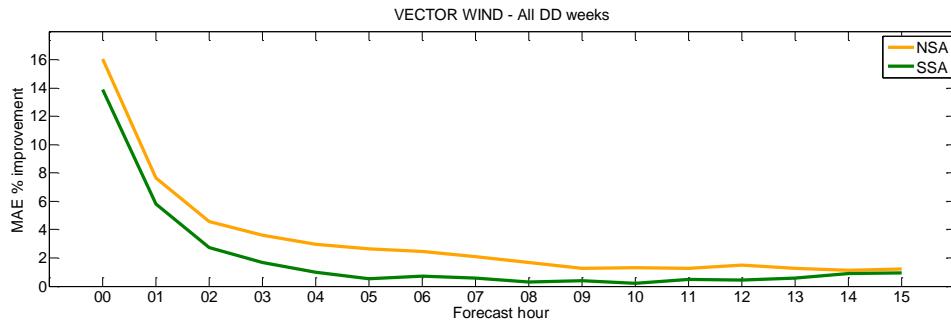


Figure 6.19. MAE percent improvement for the vector wind, for the NSA (orange curve) and SSA (green curve) for all 55 DD episode days.

Figure 6.20 shows the percent improvement for power, again for the NSA and SSA, and for both MAE and R^2 . The MAE improvement for the power looks qualitatively similar to that for the vector wind, with larger improvements in the NSA than SSA. For R^2 the improvement is more similar in the two areas, with a larger improvement falling in either the NSA or the SSA depending on the forecast hour.

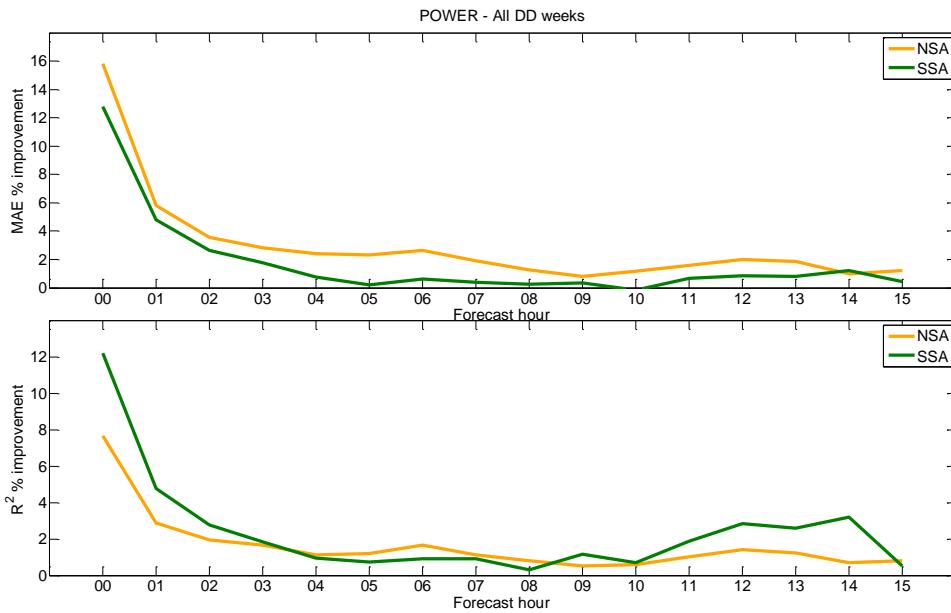


Figure 6.20. The same as Fig. 6.19 except for power.

6.6.3. Seasonal variation

Since the various DD episodes were chosen to sample all seasons of the year, breaking out the percent improvement for each DD episode provides a seasonal analysis. The percent improvement of the vector wind MAE is shown in Fig. 6.21, with cooler colors for the winter months and progressively warmer colors for succeeding months ending with magenta for October.

In the NSA the improvement is largest for the two autumn months (September and October) with significant improvement lasting out to forecast hour 15. December, January, April, and June have considerably lower MAE improvement for forecast hours beyond hour 04. In the SSA October again starts out as one of the best months of forecast improvement, but that early improvement is lost for forecast hours beyond hour 05. Overall it is difficult to identify any particular season of the year that has clearly superior forecast improvement in both the NSA and SSA for the entire range of forecast hours.

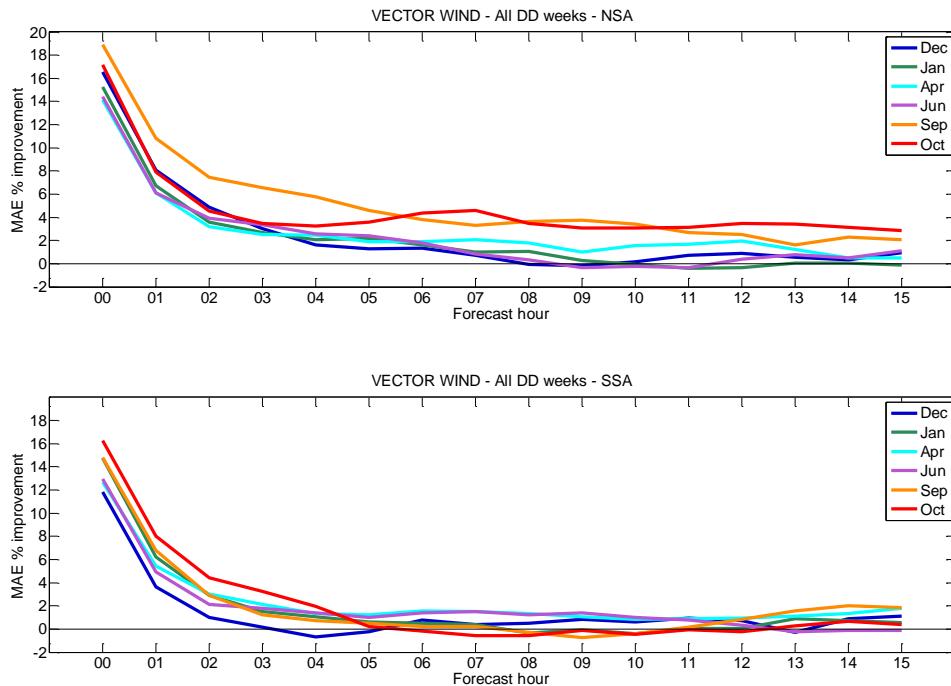


Figure 6.21 MAE percent improvement in the vector wind broken out by DD episode, for all 55 DD episode days, for the NSA (top panel) and SSA (bottom panel).

The seasonal variation in MAE and R^2 for power is shown in Fig. 6.22 for the NSA and SSA. For the NSA September and October again show relatively greater MAE improvements, while January and notably June show the worst improvement. The SSA is almost the mirror image of the NSA, with October the worst month, while June is the best, especially for later forecast hours. Evidently the DD episodes are of short enough duration that the individual DD episode statistics are heavily influenced by particular meteorological events that occur. Further analysis is required to better understand the variation from one DD episode to another and to characterize the meteorological events that influence them.

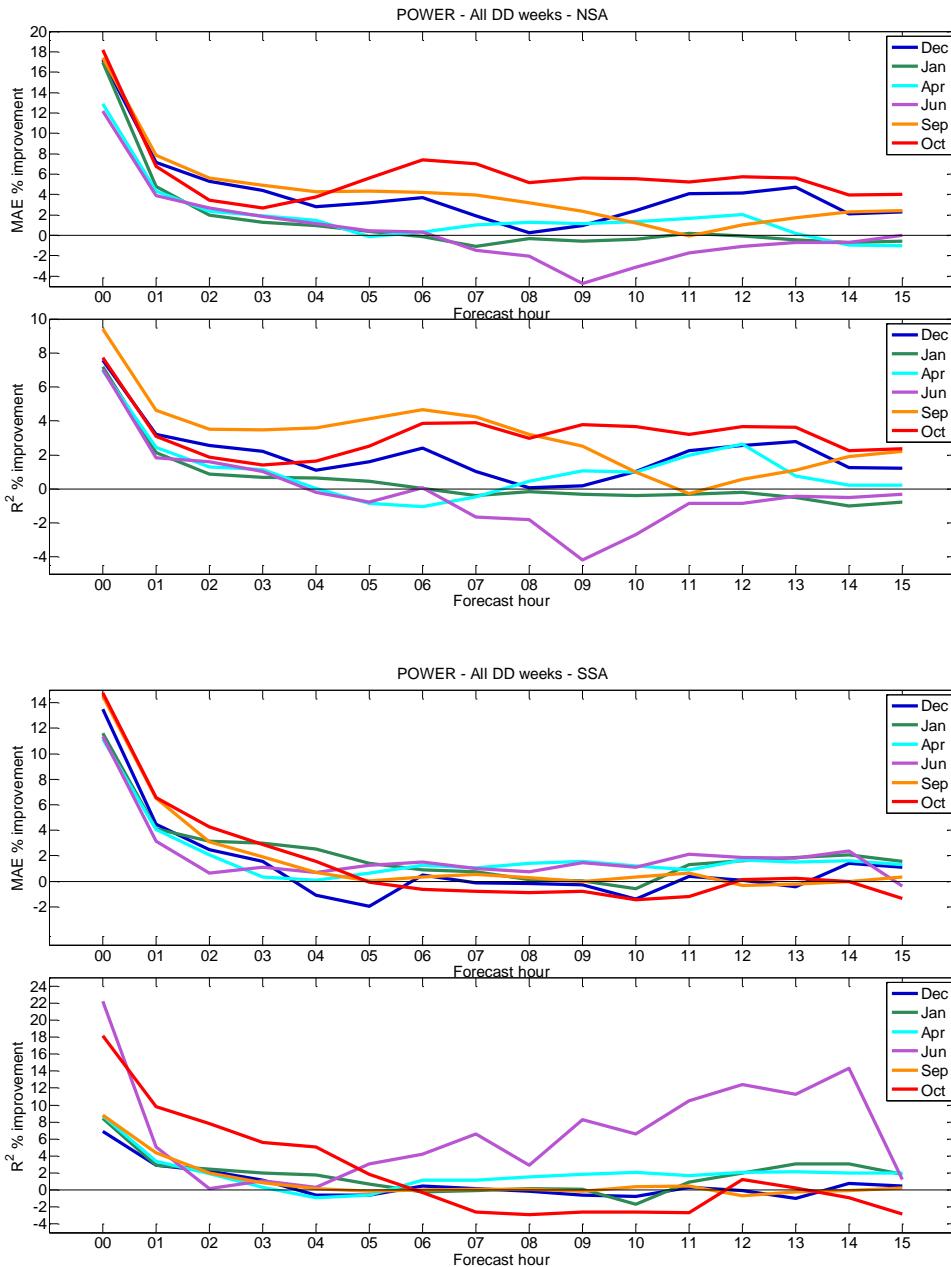


Figure 6.22. The same as Fig. 6.21 except for power, and for both MAE and R^2 .

6.6.4. Validation hour sensitivity

Next the diurnal variation of forecast MAE and MAE percent improvement is evaluated by visualizing these both as a function of forecast validation time (0-23 UTC) as well as forecast length (0-15 hours). Since the diurnal cycle near the Earth's surface is greatly different in summer and winter, and forecast errors may also be dependent on the surface vegetation and associated roughness lengths, we further

separate the data into cold season (October, December, January) and warm season (April, June and September) periods.

Figure 6.23 shows the power MAE for the NSA, and Fig. 6.24 the power MAE for the SSA. The y-axis is the forecast length starting at the initialization time at the top and forecast hour 15 on the bottom, and the x-axis is the time of the day that the forecast is valid for, running from 00 UTC on the left to 23 UTC on the right. Both study areas fall mostly in the U.S. Central Time Zone, so UTC – 06h = local standard time. Arrows showing the average times of sunrise (SR) and sunset (SS) for the center of the two study areas are indicated at the bottom of the figure. The MAE is larger in the SSA than the NSA (as was also found in Fig. 6.16 when using the sodar observations), and larger in the warm season than the cold season. Due to a variety of meteorological phenomena, including that the northern area has more synoptic scale systems while the southern area has more thunderstorm scale convection, more frequent and stronger LLJ's, and somewhat more complex topography, ERCOT has a harder time than MISO in forecasting and integrating wind. Also, in all 4 panels of Figs. 6.23 and 6.24 the MAE is smallest for forecasts whose verification times are during the daytime hours, especially for those forecast that were also initialized during the daytime hours, clearly showing the difficulty of models to forecast the stable boundary layer correctly

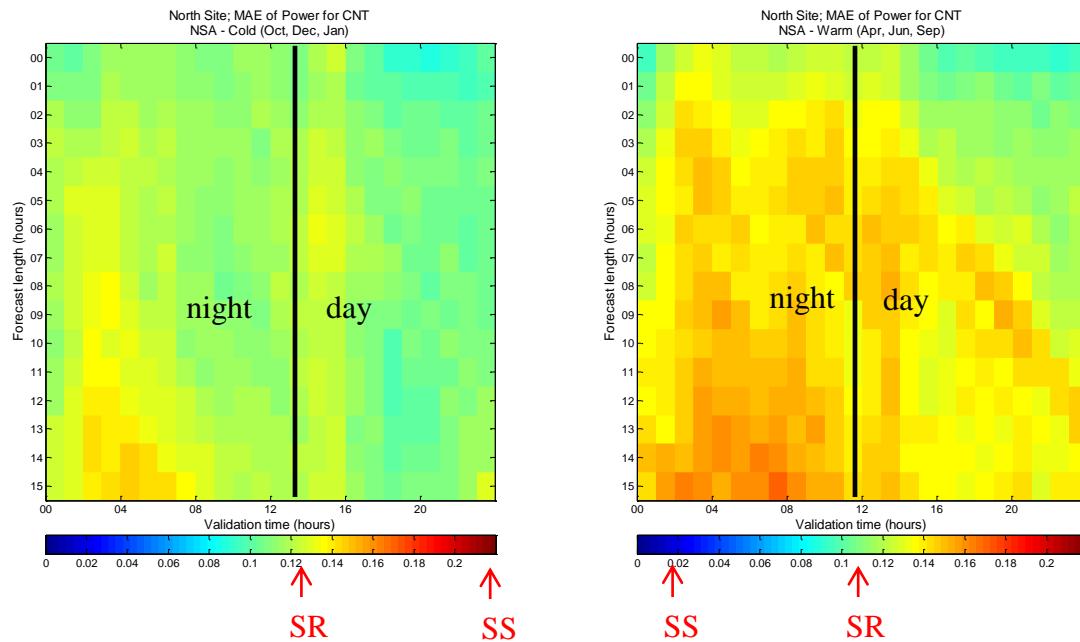


Figure 6.23. MAE of power in the NSA as a function of forecast length (y-axis) and forecast validation time (x-axis). Warm colors are larger forecast errors, cold colors smaller errors. The left panel is the cold season average of the October, December and January DD episodes, the right panel is the warm season average of the April, June and September DD episodes. The average times of sunrise (SR) and sunset (SS) are indicated by arrows, and the vertical black line is at sunrise.

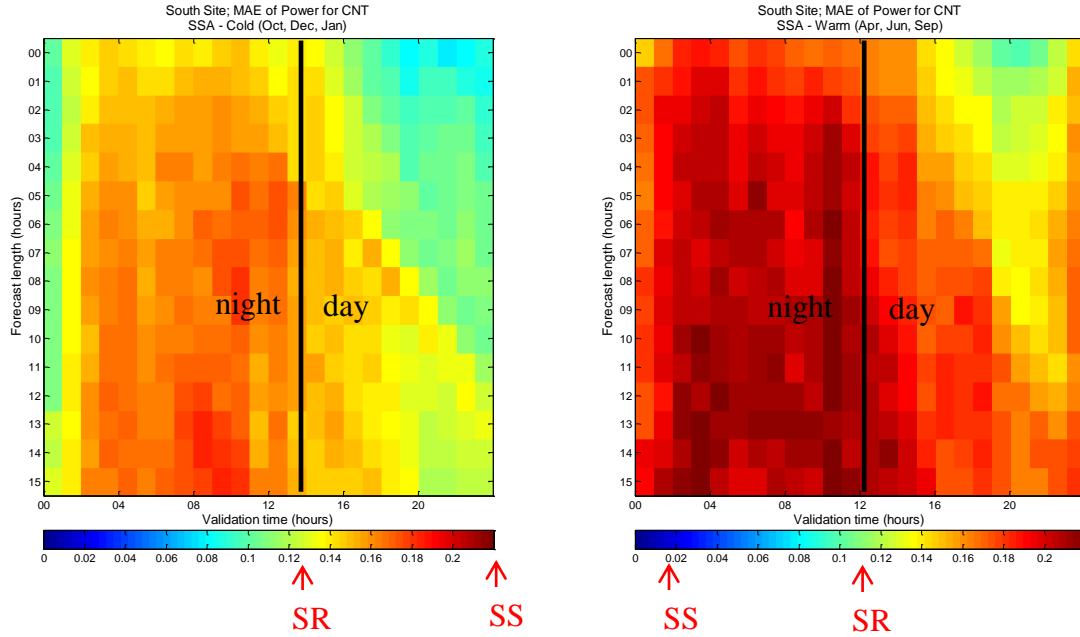


Figure 6.24. Similar to Fig. 6.23 except for the SSA power MAE.

Figure 6.25 shows the improvement in power MAE for the NSA in the same format. The improvement in power MAE in both seasons is obviously larger at short forecast lengths (the top of the figure) than for longer forecast lengths, but is also greater during the daytime hours than the nighttime hours. For the warm season the larger improvements actually start 3-4 hours before sunrise, and remain until about 4 hours before sunset. Forecasts of 5-8 hours length that are verified during the morning to mid-afternoon hours have some of the largest improvements. The cold season improvements are fairly similar to the warm season, except that the period with the largest improvement is more centered on the daytime hours, and the worst improvement is just before sunrise.

The MAE improvement for the SSA is shown in Fig. 6.26. For the cold season the largest improvements again occur in the daytime hours, although the difference between day and night is not as pronounced as in the NSA. For the warm season, the SSA shows periods of improvement both day and night, with the worst improvements just before sunrise, mid-afternoon, and near sunset.

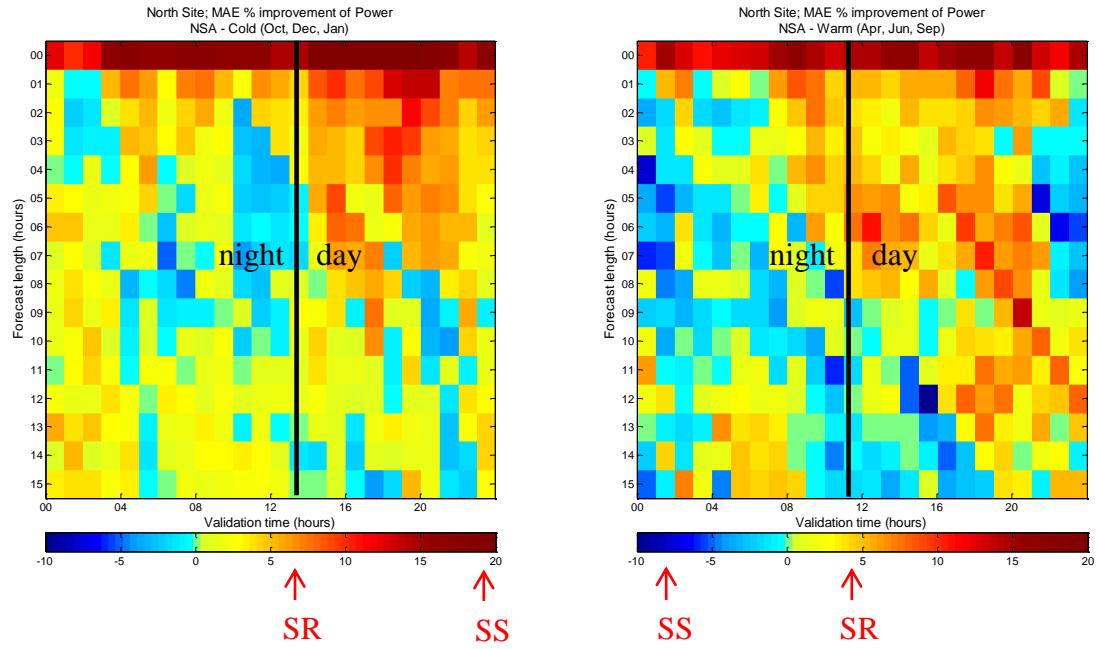


Figure 6.25. Power MAE percent improvement of forecast power in the NSA as a function of forecast length (y-axis) and forecast validation time (x-axis). Warm colors are a positive forecast improvement, cold colors negative. The left panel is the cold season average of the October, December and January DD episodes, the right panel is the warm season average of the April, June and September DD episodes. The average times of sunrise (SR) and sunset (SS) are indicated by arrows.

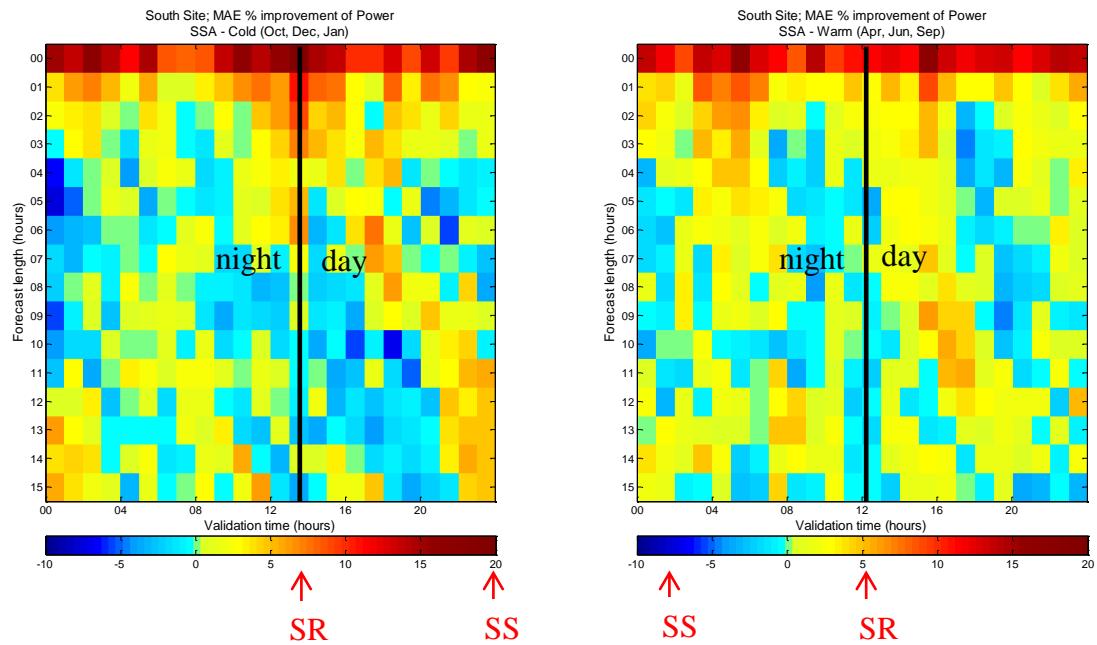


Figure 6.26. Similar to Fig. 6.25 except for the SSA.

6.6.5. Observed power dependence

We next investigate the dependence of the forecast improvement on the value of the observed power (derived from the tall tower wind data); that is, are times of low power generation improved as much as times of high power generation? Figure 6.27 shows the percent MAE improvement for the NSA and SSA as a function of the observed power, for forecast hours 00, 03, 06, and 12. The improvement at initialization time decreases as the power increases in the NSA, but this trend is less obvious in the SSA. Although for individual forecast hours the variation of improvement with power is rather noisy, on average across the three forecast hours shown, the percent improvement tends to increase slightly for larger power values with the possible exception of the largest observed power bin.

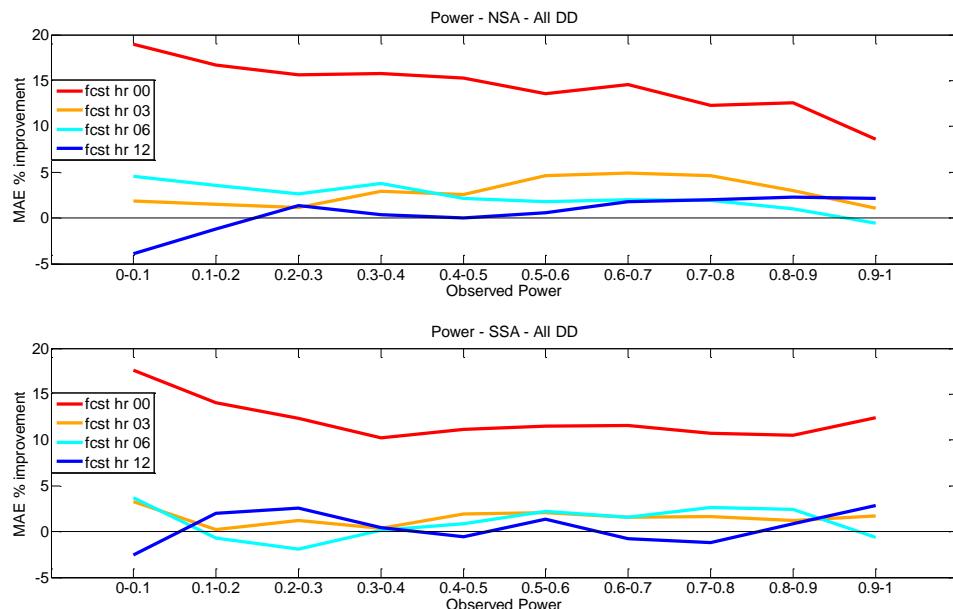


Figure 6.27. Percent improvement in MAE for four forecast lengths as a function of observed power, averaged for all 6 DD episodes, for the NSA (top panel) and for the SSA (bottom panel).

6.6.6. Large forecast errors

The dependence of the forecast error improvement on the size of the forecast error is investigated next. Figure 6.28 is a scatter plot of the power error in the control simulations (x-axis) versus the power error in the experimental simulations (y-axis) at the model initialization time, and is used to describe the method used. If there were no improvement the data would fall on the 1-1 line shown in magenta. On average both positive (model power is greater than observed) and negative (model power is less than observed) model errors are reduced, as demonstrated by the best fit line shown in teal. To determine the improvement for errors larger than 80% of the capacity of a hypothetical wind plant at the tall tower location, red dashed lines are drawn at 0.8 for both the control and experimental errors. The MAE improvement for all errors greater than 0.8 is then determined by averaging the absolute differences between control and experimental errors for all data points that lie above and to the right of the red dashed lines. Similarly, the MAE improvement for all negative errors worse than 80% is found by

averaging the absolute differences between control and experimental errors for all data points that lie below and to the left of the blue dashed lines. The dashed lines are then moved to ± 0.6 , ± 0.4 , ± 0.2 , and 0, and the error improvement is calculated. The error improvement at $+0.4$ therefore shows the percentage improvement for all errors greater than 0.4, and includes the errors greater than 0.6 and 0.8. The MAE percentage error improvements are shown in the bar chart of the lower panel. The percent improvement in the fit to the data at the initialization time is quite uniform for all power error thresholds, with the exception of the largest negative errors (less than -0.8) that has a smaller improvement.

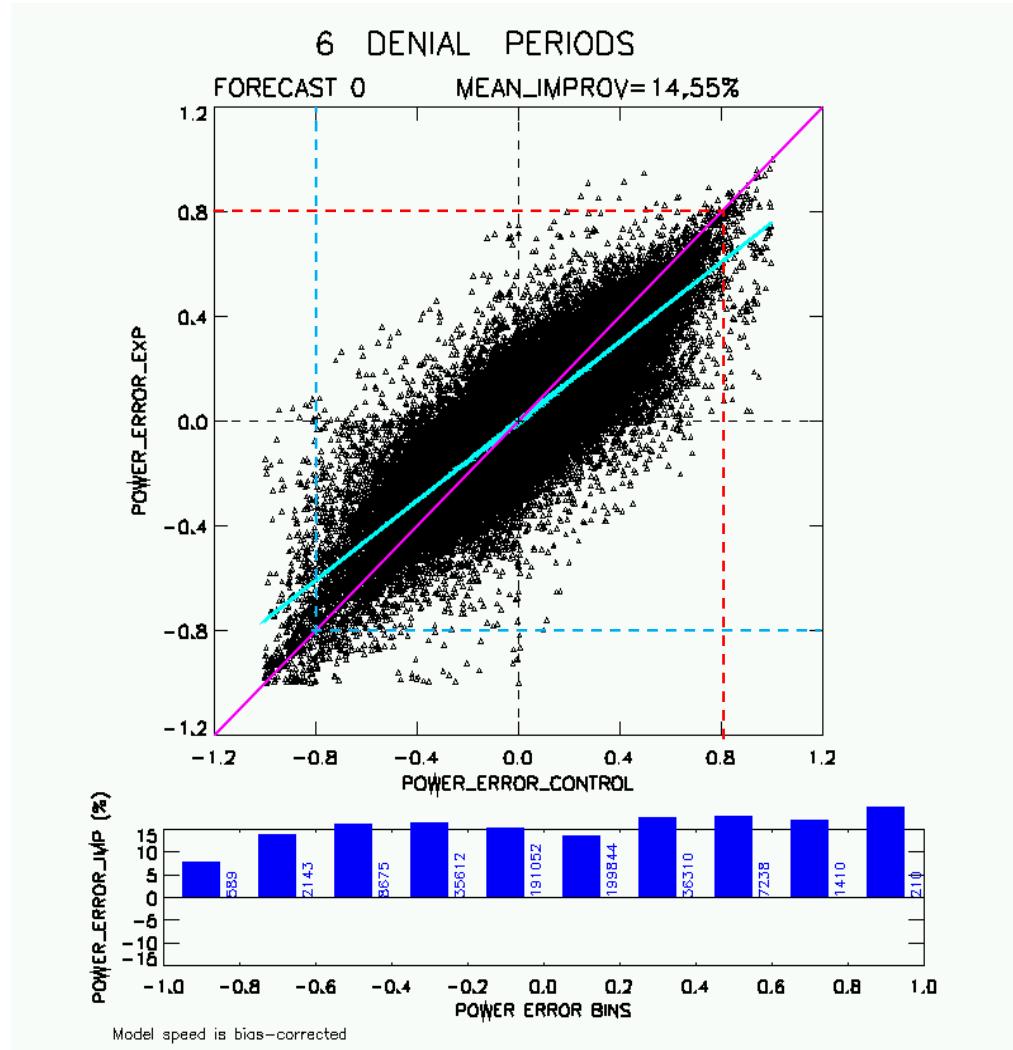


Figure 6.28. Top panel: scatter plot of the control and experimental simulation power errors at each tower site using 15 min data, for all 6 DD episodes, NSA and SSA combined, at forecast hour 00. The 1:1 line is shown in magenta, and the teal line is the best fit to the data. Dashed red and blue lines define the large error thresholds, in this example at 80% power capacity. Lower panel: MAE percent improvement for all errors greater than the threshold values. Blue numbers indicate the number of points in each bin.

Because the MAE percent improvement can be rather noisy for a single forecast hour due to the small number of events in the largest error categories, average errors are shown over forecast hours 1-6 and hours 7-12 in Fig. 6.29. Considering both forecast periods, for positive forecast errors no obvious dependence on forecast error is found. For negative forecast errors, the improvement is greater for smaller forecast errors, and is negative for the most negative errors. The reasons for this behavior are unknown.

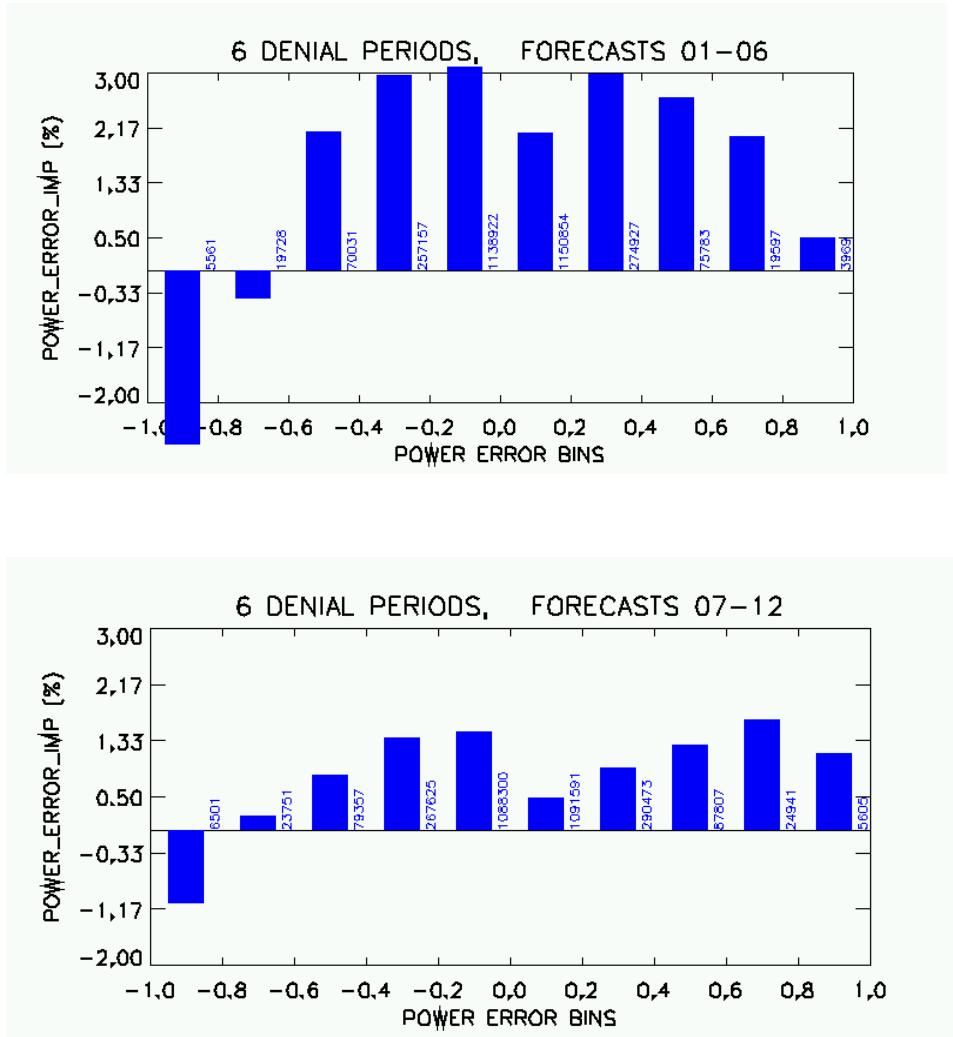


Figure 6.29. Power MAE percent improvement for control simulation errors larger than the threshold bin error size, for all 6 DD episodes, NSA and SSA combined. Top panel: the average for forecast hours 1-6. Bottom panel: the average for forecast hours 7-12.

6.6.7. Effects of spatial averaging

The statistics shown up to this point have all been averages of errors calculated at individual point locations. These statistics are appropriate if one is interested in the skill in forecasting for an individual wind plant that fits within a single model grid cell. For some applications one would instead be interested in comparing spatially averaged power forecasts with spatially averaged model forecasts. For example if a number of dispersed wind plants were feeding power into a transmission line and the overall power flowing through that transmission line is the quantity of interest. Spatially averaged forecast skill can differ from the average skill of individual point locations if the point locations have compensating errors, where an over-forecast at one point balances an under-forecast at another point.

To evaluate the effects of spatial averaging, we used forecasts and observations from the NSA, since that domain had tower data spread over a larger geographic area than the SSA. First an 8x8 grid of grid boxes was overlain on the NSA domain, with each grid box approximately 100 km (north-south) by 150 km (east-west). Within each of these grid boxes all of the speed observations and forecasts for all of the towers within the box were averaged at each hour. The MAE was then computed for this aggregated set of 64 observations and forecast locations. The process was then repeated using a 4x4 grid, a 2x2 grid, and finally averaging the observations and forecasts for all of the tower sites together (a 1x1 grid), and then calculating the MAE. Since each tower site has equal weighting in both the 1x1 grid and when each tower site is evaluated individually, we also weighted the various grid boxes by the number of towers within them so that each tower again has the same weight as any other tower, even if some are averaged together with more neighbors than others. This would more closely represent the actual aggregate power and power forecast improvement from an uneven geographical distribution of wind plants in the domain.

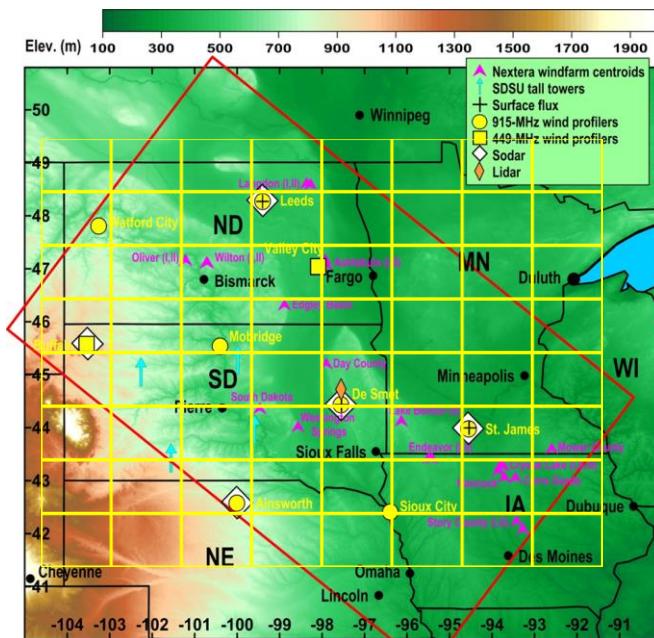


Figure 6.30. Spatial averaging boxes using an 8x8 grid for the NSA. The 4x4 grid is formed by combining 4 neighboring cells in the 8x8 grid, and the 2x2 grid by combining 16 neighboring cells.

The forecast power MAE's for the various degrees of spatial averaging are shown in Fig. 6.31, with the solid curves for the control simulations and the dashed curves for the experimental simulations assimilating the new WFIP observations. The reduction in MAE provided by spatial averaging is very large: almost a factor of three reduction in MAE from treating each tower individually to when all towers are aggregated together. Clearly, to the extent that grid operators are unencumbered by transmission constraints and can aggregate all wind power production together, the larger the spatial domain the better.

The difference between the dashed lines and solid lines shows the improvement that assimilation of the new WFIP observations provides at the various degrees of spatial averaging. Interestingly, although the MAE itself decreases continuously with more spatial averaging, the absolute improvement remains fairly constant for all size averages until it finally decreases in the 1x1 box when all towers are combined into a single aggregate. This indicates that for moderately large aggregation areas (the 2x2 boxes are 400 km x 600km) that forecasts can still be improved with assimilation of new observations.

This last point is highlighted in Fig. 6.32, which shows the MAE speed percent improvement for the various degrees of aggregation. The MAE percent improvement at hours 00, 01, and 02 increases with the degree of spatial averaging, and is largest for the 2x2 grid first for the first 7 forecast hours.

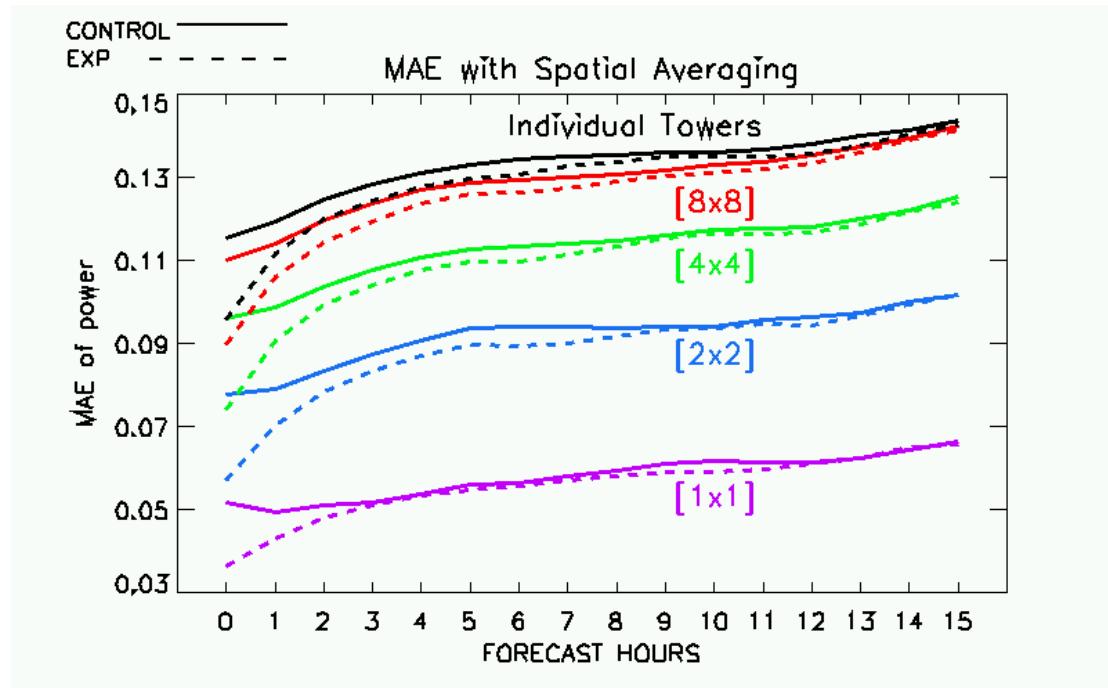


Figure 6.31. Power MAE with different degrees of spatial averaging for all 6 DD episodes for the NSA. The solid lines are for the control simulations, dashed for the experimental. The black lines are with no spatial averaging, and the purple lines for the maximum spatial averaging with all tower location observations and forecasts aggregated into single time series.

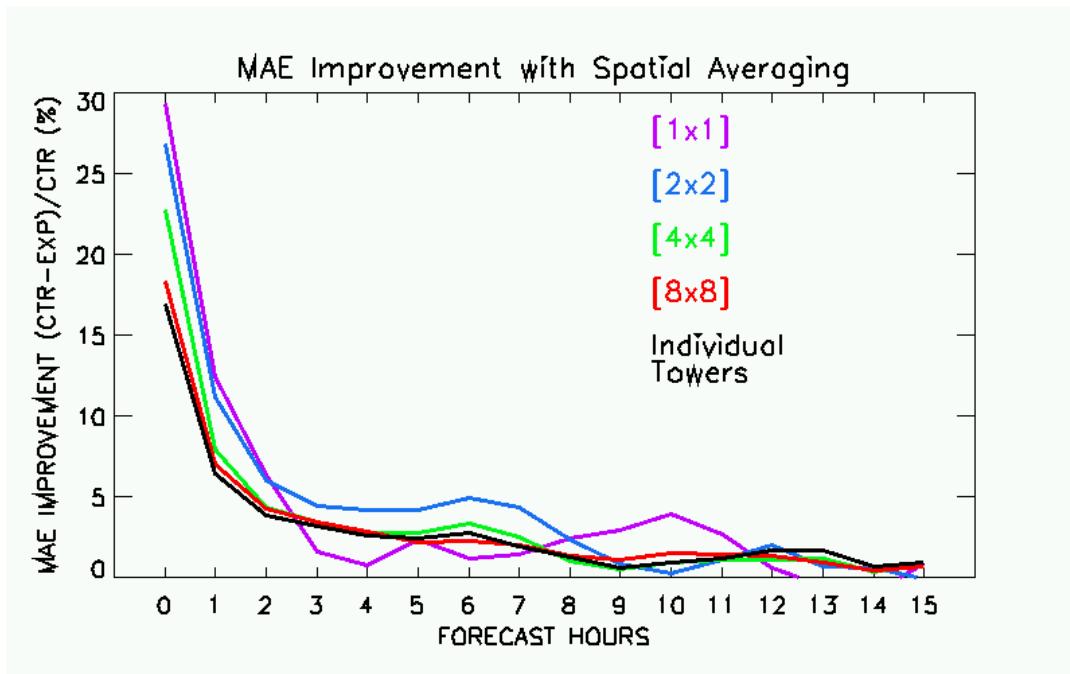


Figure 6.32. MAE percent improvement of forecast power for the various degrees of spatial averaging, for all 6 DD episodes for the NSA.

To further demonstrate the effects of spatial averaging, time series of the aggregate observed power (P_O , black lines) and forecast power (P_C for control, shown as a red line, P_E for the experimental simulations shown as a blue line) at all of the tall tower sites for the October DD episode are shown for forecast hours 00 and 03 for the NSA in Fig. 6.33, and for the SSA in Fig. 6.34. The green line shows the instantaneous MAE percent improvement (IPI), calculated as the absolute difference between the red and black lines minus the absolute difference between the blue and black lines, all normalized by the time average over the entire DD episode of the absolute difference between the red and black lines.

$$IPI = \frac{|P_C - P_O| - |P_E - P_O|}{\langle |P_C - P_O| \rangle}$$

The horizontal green line shows the average IPI over the entire DD episode, and the black horizontal line is drawn at zero MAE improvement.

In general, the aggregate control forecasts do a good job of matching the aggregate power at the model initialization time (hour 00), although significant improvements are still found by assimilating the new WFIP observations, as can be seen by the green curve and solid line, and as can be seen especially on October 20-21 (Fig. 6.33). At forecast hour 03 the aggregate forecast errors have clearly increased in both the control and experimental simulations, and large positive and negative swings in the IPI occur, but a net improvement in the IPI is still present. For the SSA the control and experimental forecasts do

not match the observations as well as in the NSA at the initialization time, and a smaller but still significant improvement in the experimental simulation exists compared to the control. At forecast hour 03 the forecast errors have increased considerably, and the *IPI* has decreased but is still positive.

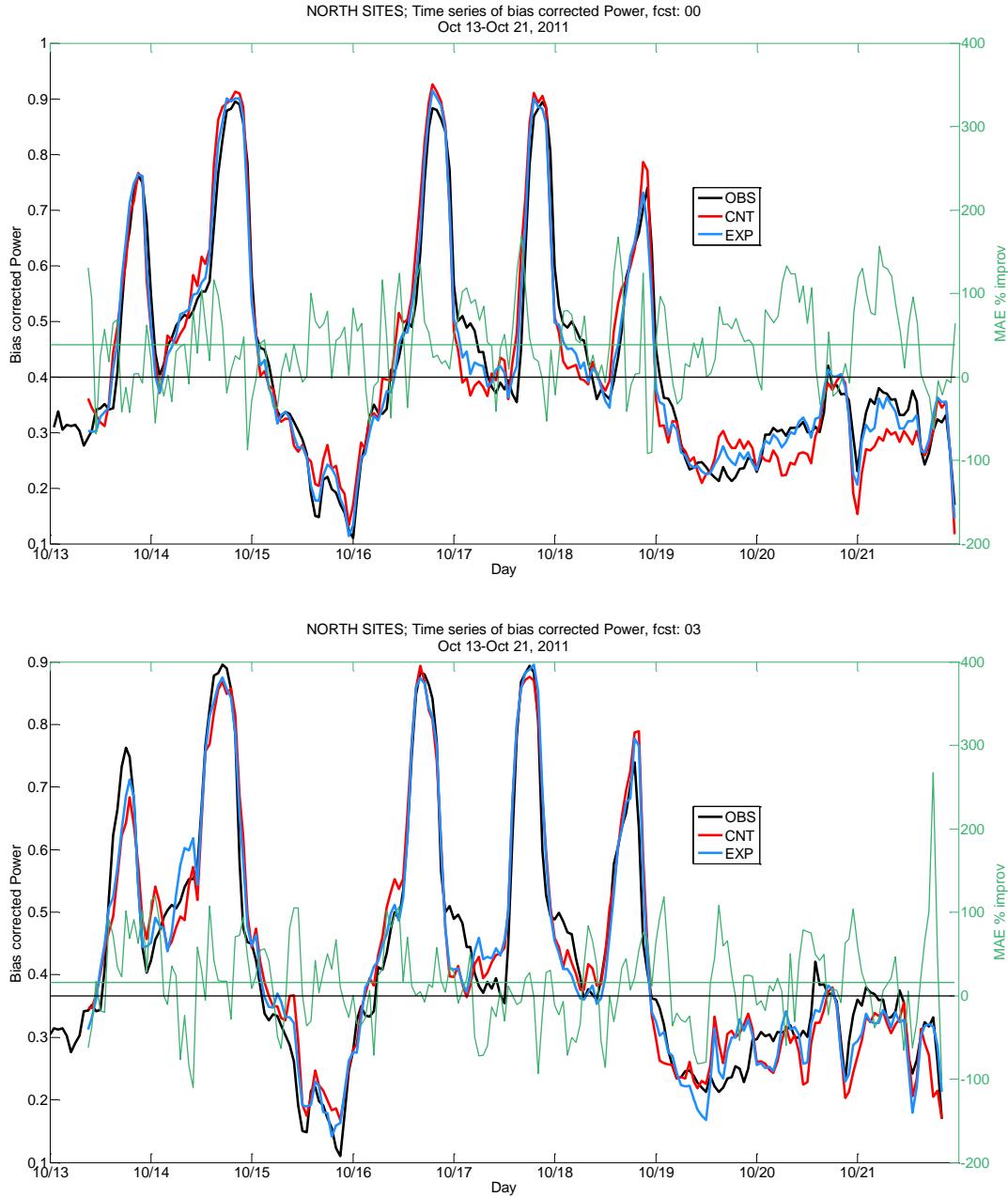


Figure 6.33. Aggregate observed power (black curve), control forecast power (red curve), experimental forecast power (blue curve), and instantaneous MAE percent improvement (green curve) at forecast hour 00 (top panel) and hour 03 (bottom panel) for the October DD episode and for the NSA. 10 min observed powers are interpolated to 15 min intervals to match the model forecast times. Horizontal green line shows the average instantaneous MAE percent improvement over the entire DD episode.

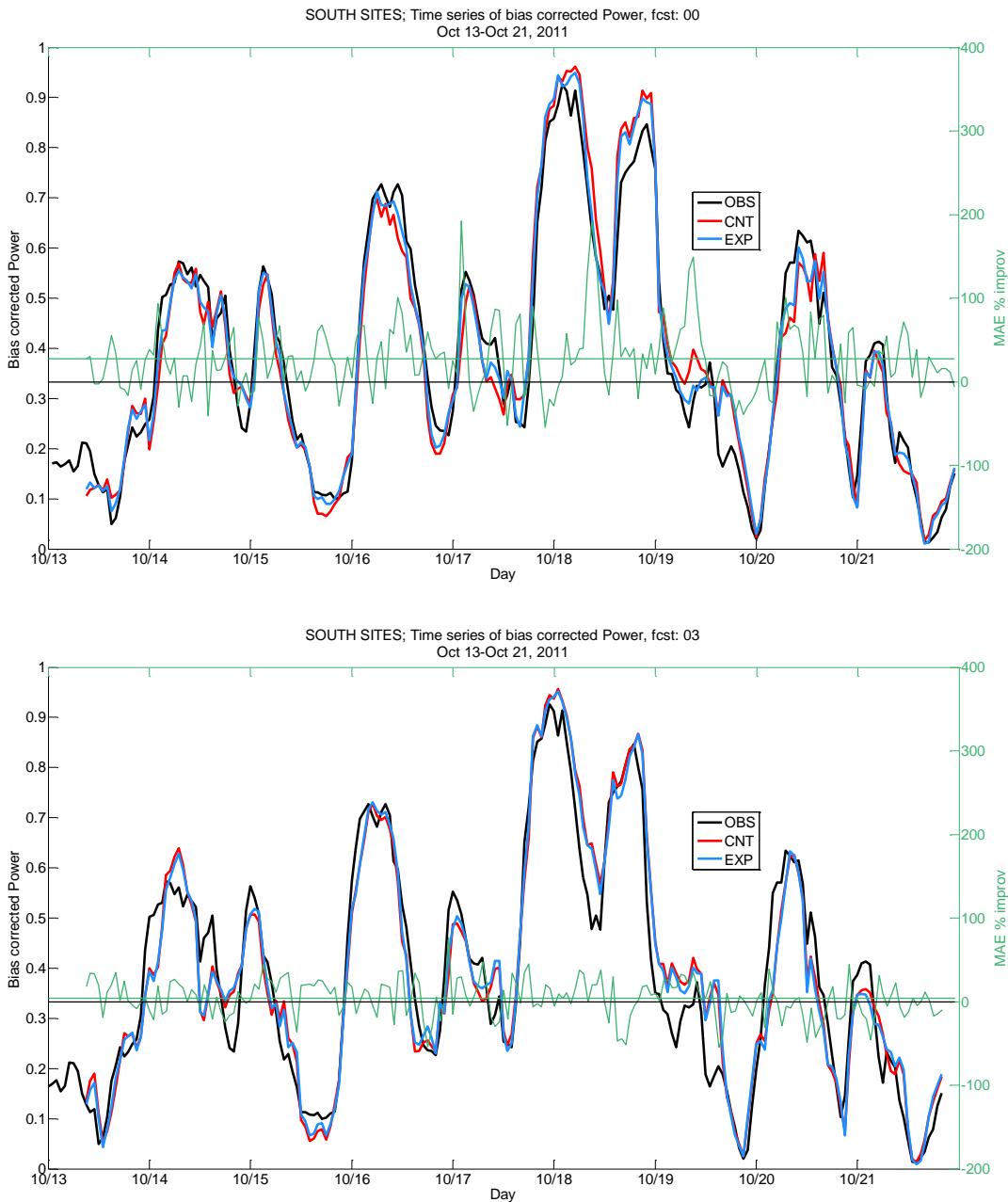


Figure 6.34. The same as Fig. 6.33 except for the SSA.

6.6.8. Geographic outlier sensitivity analysis

A significant number of the tall tower observation sites were located at a considerable distance from the remaining WFIP observations, and thus could be considered to be “geographical outliers”. Since one would expect to see a smaller impact of the bulk of the observations on a site far removed from those observations, a geographical outlier sensitivity analysis was done in which these far outlying stations

were eliminated from the percent improvement calculations. The green rectangle and circle in Fig. 6.35 show the restricted areas for the NSA and SSA tall tower evaluation, with any tall tower sites located outside of these two more restricted domains eliminated from the analysis. In the NSA 34 sites were eliminated (25 % of the total) and in the SSA 14 sites (27 % of the total) were eliminated.

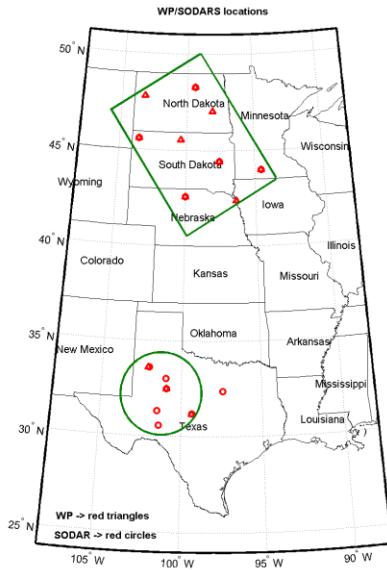


Figure 6.35. The two study areas with the more restricted domains shown by the green rectangle in the NSA and green circle in the SSA. Tall tower sites falling outside of these two areas are eliminated in the geographic outlier sensitivity analysis.

The results of eliminating these geographic outliers is illustrated in Fig. 6.36, where the dashed lines show the MAE and R^2 percent improvement for the restricted domains, and the solid lines are when using all of the tall tower sites. For both the NSA and SSA, for both MAE and R^2 , the percent improvement is found to mostly be greater when using the more restricted domains. This is especially true for the first 4-5 forecast hours, and the increase in percent improvement is more significant in the SSA than the NSA. This result demonstrates that the local density of observations influences the degree of forecast improvement: outlying sites where the density of observations assimilated is low have a smaller improvement than sites where the density of observations is greater.

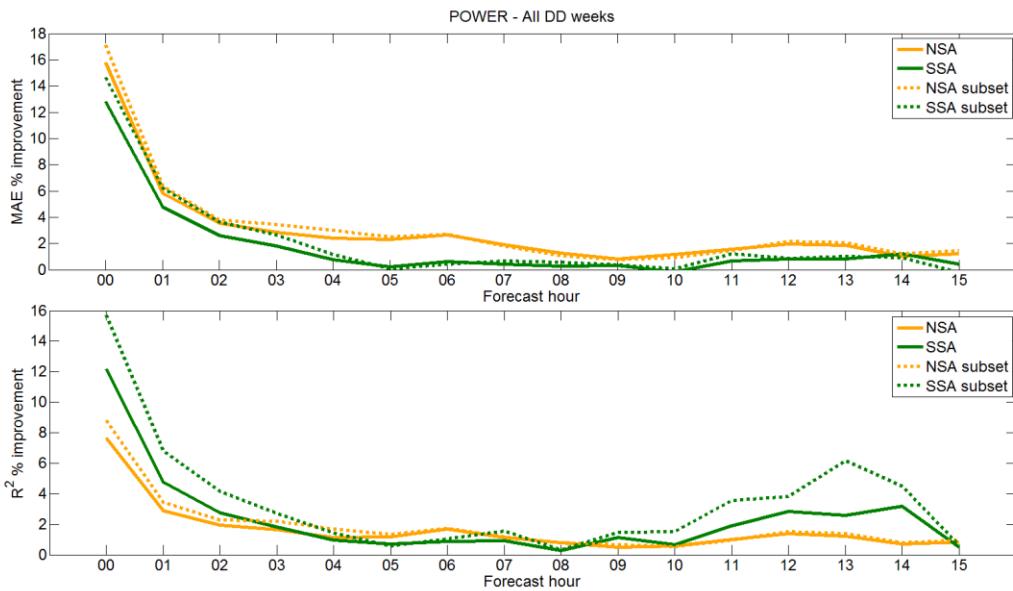


Figure 6.36. Percent improvement of MAE (top panel) and R^2 (bottom panel) for the NSA (orange) and SSA (green) when using all tower observations (solid) and when using only those tall towers that fall within the restricted analysis domains shown in Fig. 6.35.

6.7. NAM results

6.7.1 Wind Profiler and sodar verification

The impact of the new WFIP observations on the NWS/NCEP NAM model forecast skill was evaluated by comparing forecasts against WFIP profiler and SODAR observations in both the northern and southern study domains. The NAM/NDAS system completed data denial experiments for the two winter episodes only, 30 November – 6 December 2011, and 7-15 January, 2012. This was accomplished by interpolating the forecast to each observation location, taking the difference (forecast - observation), and then averaging all differences within a specified layer above ground level (AGL). For profilers this level was specified to be 0-2 km AGL while SODARS were set to 0-200 m AGL due to their much shallower vertical sampling of the wind.

Forecast verification for the NAM and CONUSnest was done using the NCEP Forecast Verification System. Statistical significance testing was done using a Monte Carlo technique with 2000 samples (Hamill, 1999). Any lines in the NAM/NDAS verification plots which lie outside the boxes are significant at the p=0.05 level. No bias correction has been applied to the NAM or CONUSnest forecasts. The control simulations for the 12 km parent and 4 km CONUSnest will be referred to as NAM and CONUSnest respectively. The experimental simulations for the 12 km and 4 km domains, which assimilated the special WFIP observations, will be referred as NAMX and CONUSnestX.

Upon initial review of profiler-based RMSE statistics for both study regions and computational domains (Figs 6.37 and 6.38), it is apparent that the addition of WFIP observations to the NAM/NDAS system provides a statistically significant benefit on the wind vector forecast for the first 4 – 6 hours of the forecast. However the southern study region showed a more lasting positive benefit through the forecast period than the northern study region, depicting reduced RMSEs out to forecast hour 8, with results that are on the edge of significance at forecast hour 7.

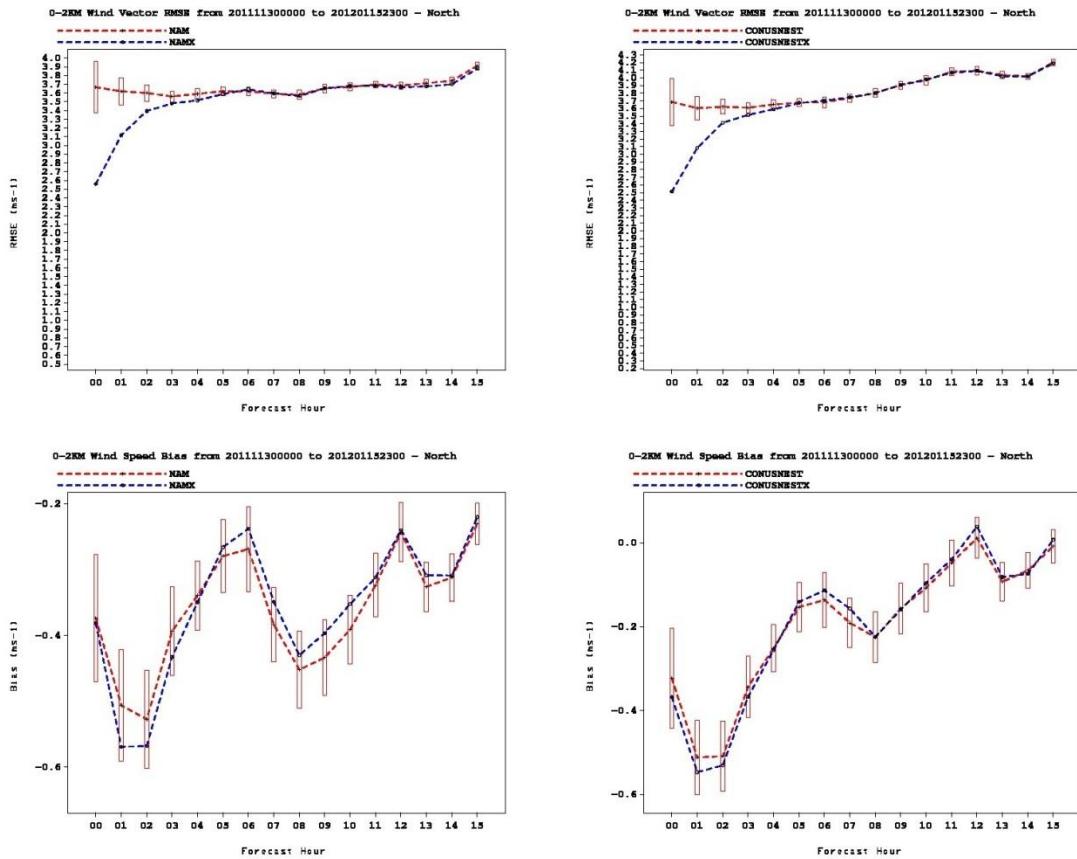


Figure 6.37. NAM vector wind RMSE (top) and wind speed bias (bottom) against WFIP wind profiler observations within the 0-2km AGL layer in the northern study region. Statistics from the 12 km parent domain occupy the left panels and forecasts from the 4 km nest domain occupy the right panels. Red traces are the control simulation and blue traces are the experimental simulation. Verification covers the two winter data denial periods.

Wind speed biases calculated against the profiler observations are generally not statistically significant (bottom row of Figs 6.37 and 6.38). However, in general, biases for both the control and experimental runs are negative, meaning that the wind speed forecasts are too slow. For the northern domain this bias is slightly worse in the early parts of both experimental forecasts but generally improves, relative to

the respective control runs, by forecast hour 5. For the southern domain the bias is generally improved, i.e. closer to zero, at all forecast hours, with the exception of forecast hour 1 with the NAMX.

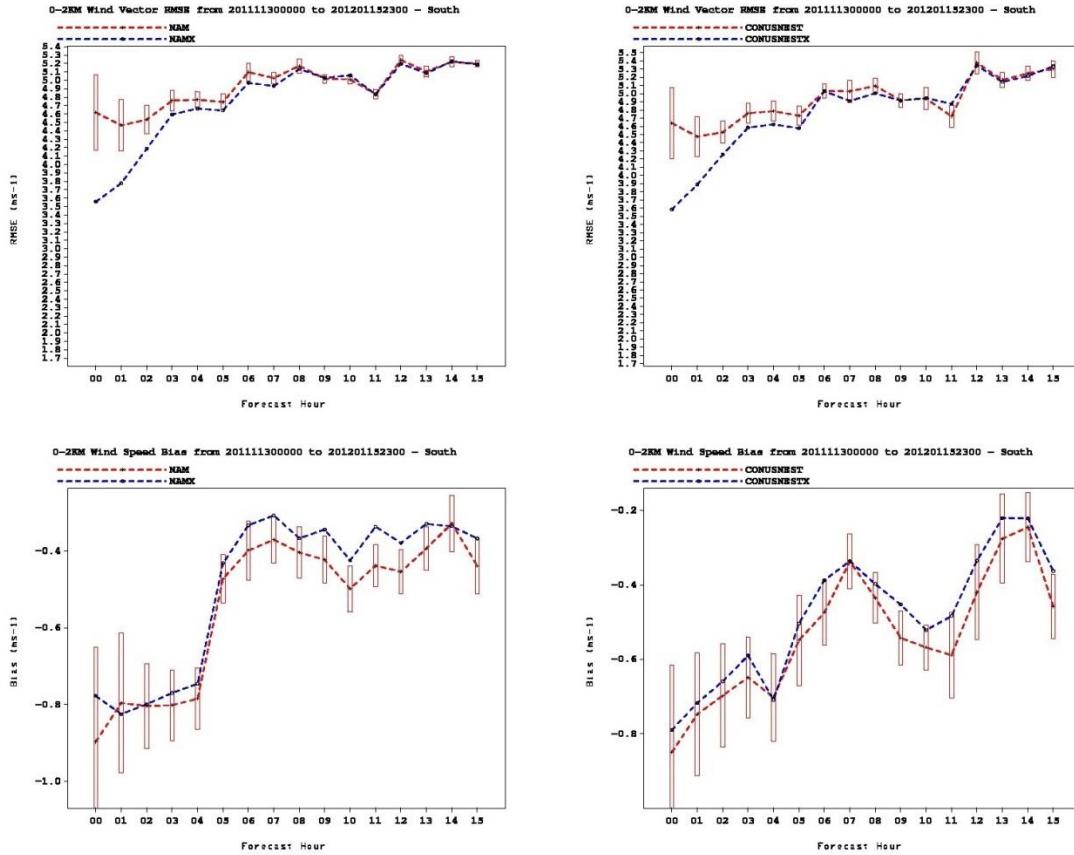


Figure 6.38. As in Fig. 6.37 except valid for the southern WFIP study region.

Figures 6.39 and 6.40 show the RMSE and bias as a function of forecast hour against SODAR observations for the northern and southern domains, respectively. For the northern study area (Fig. 6.39), both NAMX and CONUSnestX simulations show similar behavior with the profiler verification in terms of RMSE, except that the statistically significant portion of the impact has a shorter duration of about 2 hours. A reduction in RMSE may be seen out to forecast hour 8 in the NAMX – although this is not statistically significant. In the southern study area (Fig. 6.40) the interpretation of RMSE is less clear and the results fail to be statistically significant at nearly all times for both experiments. The NAMX exhibits some degradation in the RMSE for forecast hours 1 – 3, thereafter the RMSE is better than or as good as the control simulation (top left panel, Fig. 6.40). For CONUSnestX, a reduction in RMSE, relative to CONUSnest, is shown at forecast hours 0 and hours 3 – 7. Impacts are otherwise neutral, with the only negative impact occurring towards at forecast hours 12 and 13.

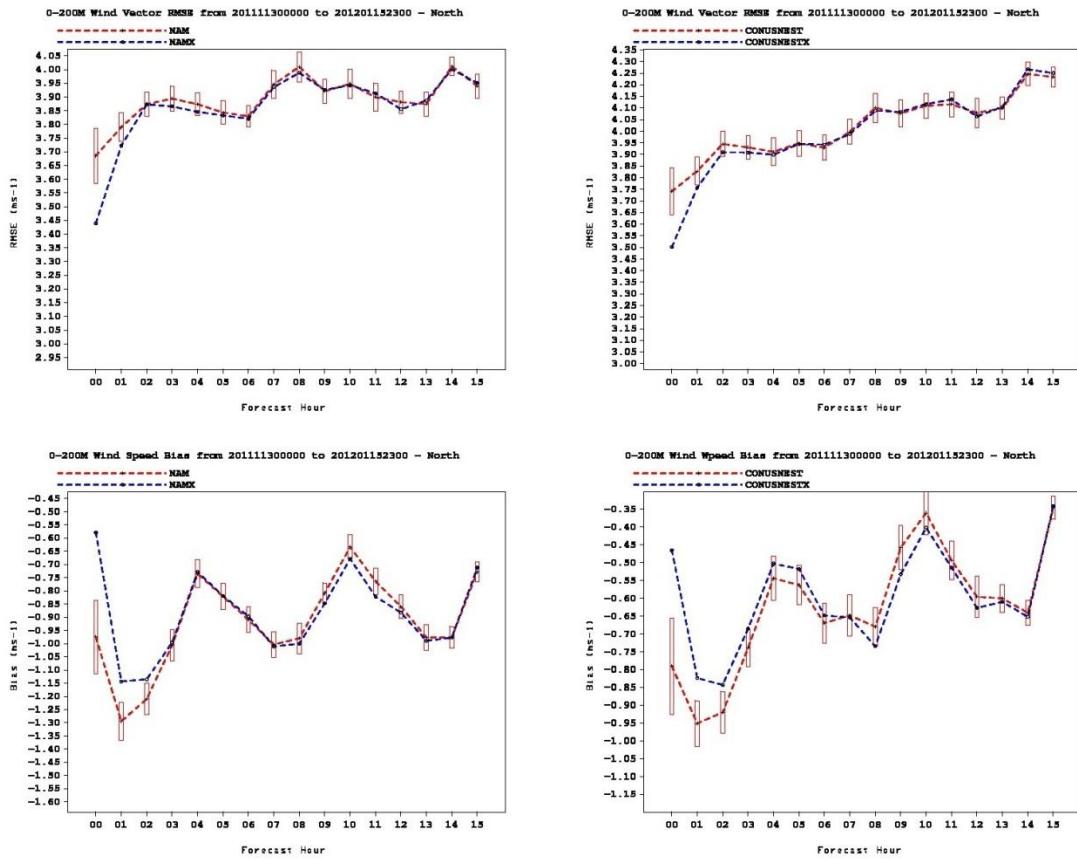


Figure 6.39. As in Fig. 6.37 except forecasts are compared to SODAR observations in the 0-200 m AGL layer. Verification is valid for the northern WFIP study domain.

The bottom panels of Figs 6.38 and 6.39 show the bias for the control and experiment simulations relative to the SODAR observations. Overall, the impacts on the wind speed bias are mixed and tend to not be statistically significant. For the northern study (Fig. 6.39) area bias improvements are seen for the first 3 hours with NAMX and up to 6 hours with CONUSnestX. After these times both experiments show degradation. In the southern study region NAMX and CONUSnestX, similar to the interpretation of the RMSE, have differing behaviors in bias (Fig. 6.40). The NAMX shows an overall increase in the wind speed bias throughout all forecast hours shown. This brings the bias closer to zero for the early and later parts of the forecast period. Otherwise this increased bias has a negative to neutral impact. The CONUSnestX shows degradation for all forecast hours except for hours 7, 9, 13, and 14.

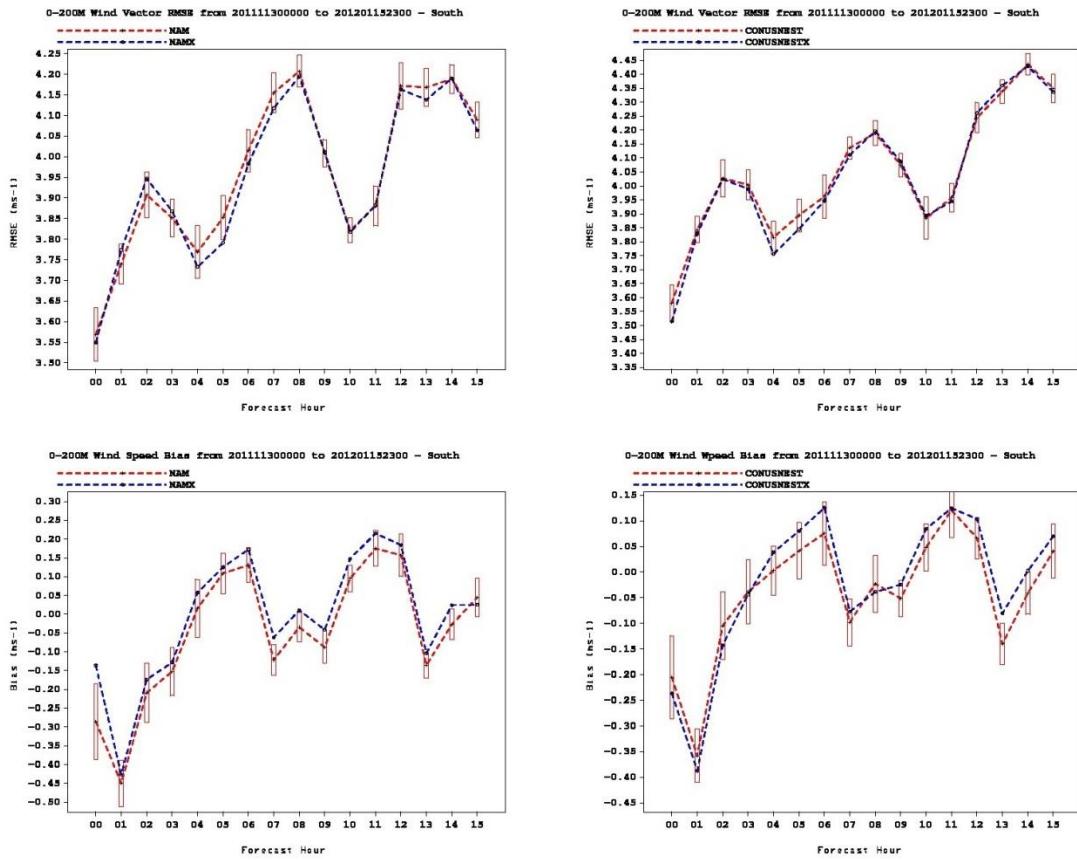


Figure 6.40. As in Fig. 6.38 except for the southern WFIP study region.

Finally, it is worth emphasizing that the results for the bias in both study regions are generally not statistically significant. In fact, this claim could likely extend to the bias associated with the profiler observations as well (bottom panels of Figs 6.37 and 6.38). This could indicate that the bias amongst all forecasts at profiler and SODAR locations has a large variance which prohibits the inference of statistical significance. While the behavior of the effect of assimilating new observations on the wind speed bias is not entirely clear, it was clear that the overall impact on the RMSEs was positive for both NAMX and CONUSnestX. The addition of WFIP observations generally yielded statistically significant improvements in the first several hours of the forecast period, especially with respect to the 0-2km AGL layer at profiler locations.

Many reasons exist which may explain the source of differences in the relative impacts of assimilating the WFIP observations between the northern and southern study regions. Such reasons include; differing observation networks, differing geography, differing local climate, and small sample size. Therefore the results presented here do not necessarily suggest the superior forecast improvement of one study region over the other, but rather that the inclusion of the additional WFIP observations yielded a positive impact in short-term, low-level wind forecasts.

6.7.1 NAM/NDAS Conventional verification over the Plains

Additional verification work was also done over the winter quarter data denial period to evaluate any potential significant results on other aspects of the forecast. Short-range forecasts were compared to conventional surface and upper-air observations. However, results from verification against upper air observations were largely inconclusive due to the spatially and temporally sparse nature of the United States upper-air network. Therefore, the focus of this brief section is with respect to the impacts of the assimilation of WFIP observations on surface observation verification in the WFIP study regions. For simplicity, verification shown here was calculated over both the northern and southern Plains regions of the United States, as defined in the NCEP Forecast Verification System (Fig. 6.41).

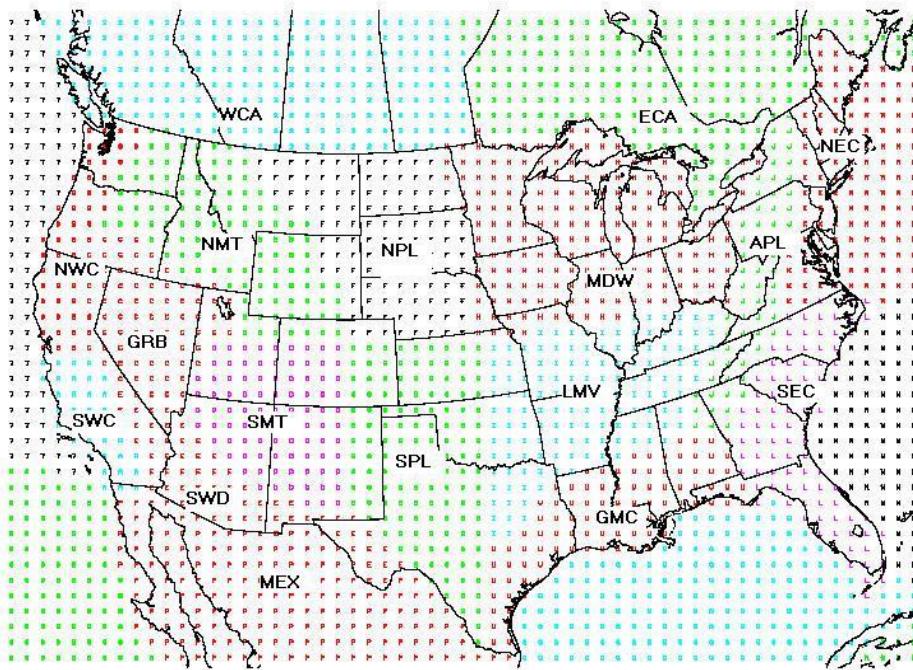


Figure 6.41. Forecast verification regions in the NCEP Forecast Verification System. For conventional verification in WFIP, regions NPL and SPL were combined to form a sub-region for the Plains states.

Verification against conventional surface observations was conducted for both NAMX and CONUSnestX over the entire winter quarter period for 2 m temperature, 2 m relative humidity, and 10 m wind. Results for 2 m relative humidity are not shown here, as the impacts were relatively neutral. For temperature, the additional WFIP observations had the effect of reducing RMSEs very slightly in the 2 – 10 hour forecast time range (~ 0.01 K, not shown) while improving biases by a statistically significant amount, for both NAMX and CONUSnestX, throughout the verification period (Fig. 6.42). The 10 m wind

verification shows a statistically significant reduction in the RMSE for the CONUSnestX at forecast hours 4 and 5, but also shows some degradation toward the end of the verification period. The NAMX depicts similar behavior, but with a slightly smaller amplitude (Fig. 6.43). Both CONUSnestX and NAMX show improved 10 m wind speed biases throughout the forecast period (Fig. 6.44) as a result of the assimilation of the additional WFIP observations.

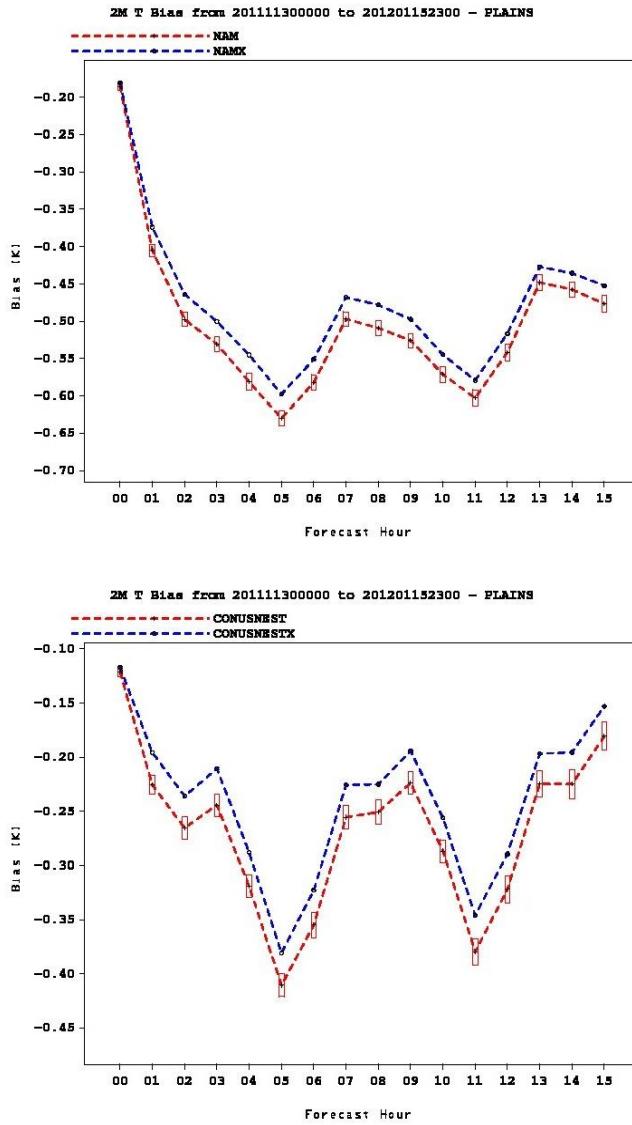


Figure 6.42. NAM/NAMX and CONUSnest/CONUSnestX 2m temperature bias over the Plains.

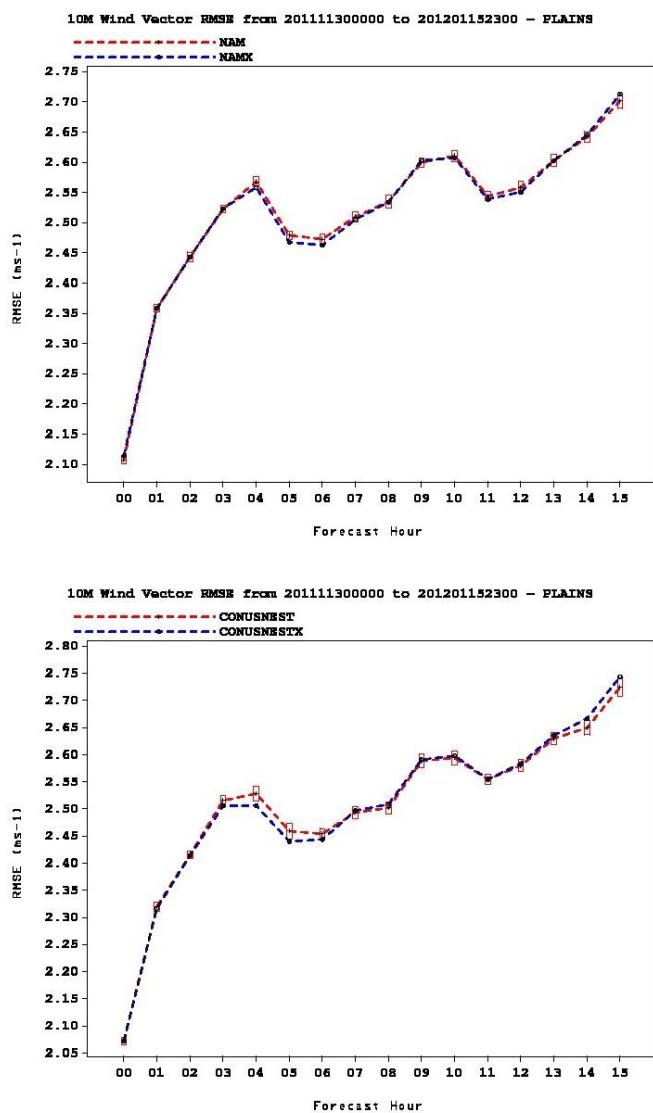


Figure 6.43. NAM/NAMX and CONUSnest/CONUSnestX 10 m wind vector RMSE over the Plains.

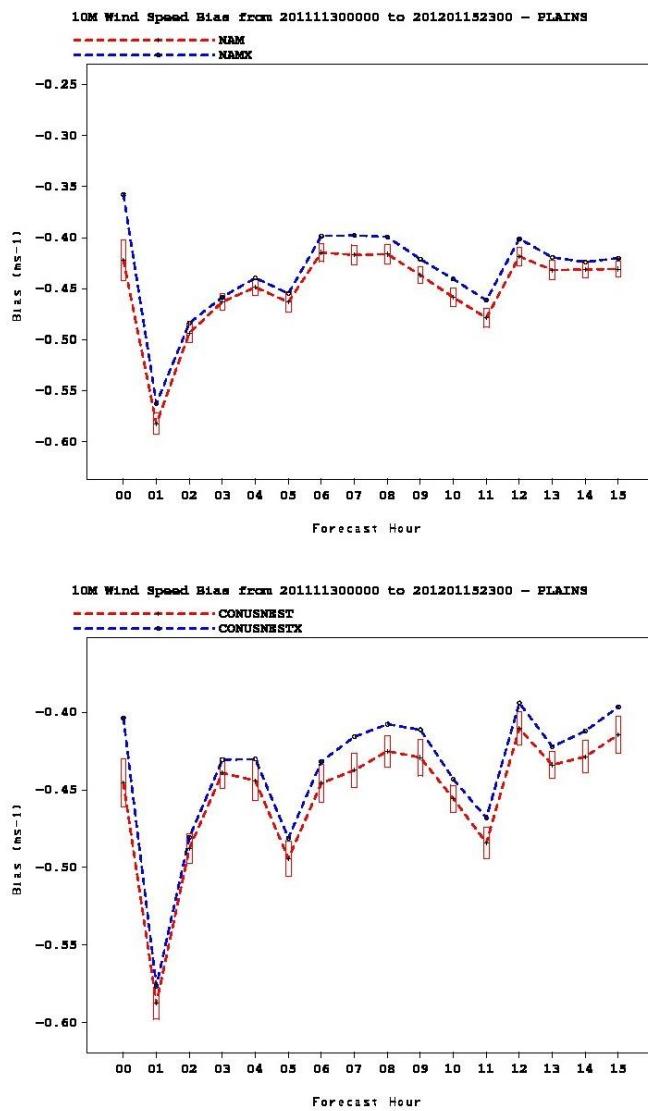


Figure 6.44. NAM/NAMX and CONUSnest/CONUSnestX 10 m wind speed bias over the Plains.

6.7.2 Tall tower and nacelle verification

The WFIP experiment allowed for the introduction and testing of two new, unique data sets provided by private sector wind energy companies, nacelle anemometer and tall tower wind observations. Until WFIP neither of these observations had been assimilated into the NDAS.

Throughout the duration of the WFIP winter quarter data denial experiments, the NAMX assimilated a total of 201,171 tall tower observations and 15,429 nacelle observations. The CONUSnestX experiment assimilated 201,144 tall tower observations and 15,429 nacelle observations. The slight discrepancy

between the NAMX and CONUSnestX in terms of the number of assimilated tall tower observations is very likely due to the gross error checking algorithm within the GSI.

If we look at the distribution of the innovation, or *observation minus forecast*, values from all analysis times within WFIP we see that both new observation types have a relatively Gaussian shape (see Figs 6.45 and 6.46 for tall tower and nacelle observations, respectively). It should also be noted that while these innovations provide information on the assimilation, they also provide information on short-term forecast errors since the background for each analysis, except at TM12, is a three hour forecast (Fig. 4.1).

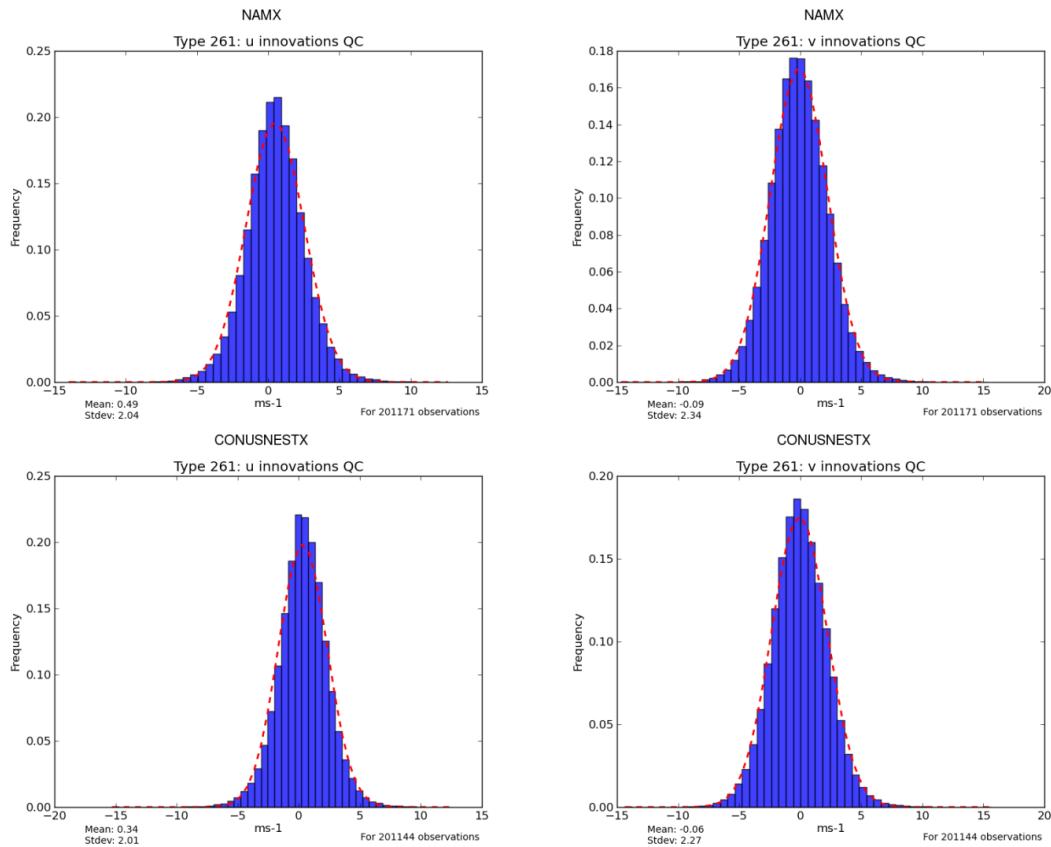


Figure 6.45. Tall tower u (left) and v (right) observation innovations (observation-forecast) from all analysis steps during the WFIP winter quarter data denial period. Distributions featured along the top are from the NAMX while distributions along the bottom row are from the CONUSnestX. The red dotted lines correspond to Gaussian distributions.

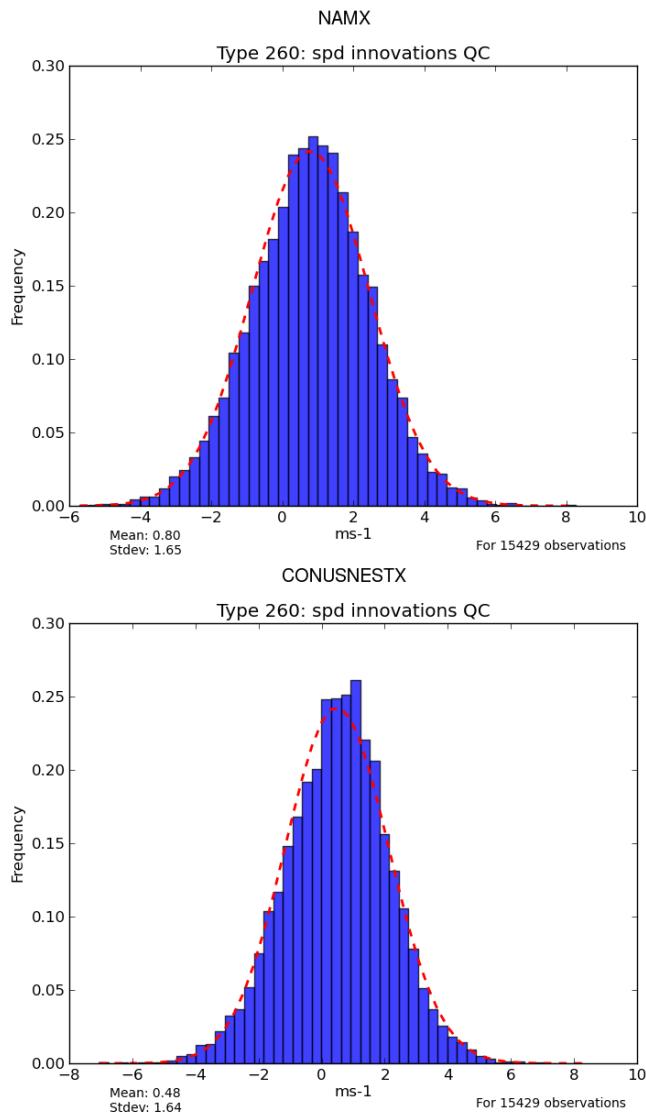


Figure 6.46. Nacelle wind speed observation innovations from all analysis steps during the WFIP winter quarter data denial period from the NAMX (top) and CONUSnestX (bottom). The red dotted lines correspond to Gaussian distributions.

The mean innovation, depicted in the lower-left of each figure, also denotes the overall mean forecast bias related to the assimilation of a particular observation. The bias for the u-component of the wind from the tall tower observations (Fig. 6.45) is 0.49 ms^{-1} for the NAMX and 0.34 for the CONUSnestX, indicating that both models may have a tendency to under-forecast the u-component of the wind. However this does not necessarily indicate a direct problem with the model, as there could also be biases present in the observations as well, as noted in the quality control section regarding the tall towers. The biases for the v-component of the winds for both NAMX and CONUSnestX are quite small, -0.09 ms^{-1} and -0.06 ms^{-1} respectively.

The mean innovations for the nacelle wind speed observations both indicate positive biases for both NAMX (0.80 ms^{-1}) and CONUSnestX (0.48 ms^{-1}) experiments (Fig. 6.46). Interestingly, the CONUSnestX has a slightly smaller bias for both the tall tower and the nacelle wind speeds which could suggest that the finer grid-spacing of the CONUSnestX allows for a more realistic representation the magnitude of the observed wind velocities in the wind turbine layer (eg; Rife et al., 2004). Guidance with high-horizontal-resolutions ($\leq 4 \text{ km}$) is already known to provide realistic and skillful representations of meso-convective phenomena (e.g.; Weisman et al., 1997; Kain et al., 2008; Schwartz et al., 2009). Nonetheless, both NAMX and CONUSnestX indicate an under-forecasting bias that may be present in the model (e.g. wind speeds which are too slow). This is an interesting result, considering the nacelle wind speed measurements are typically taken downstream of the turbine rotor blades, and thus are measuring a wind which has had some energy removed from it. Recent studies have estimated that such measured speeds may represent a 20% reduction from the true wind speed (e.g. Drechsel et al., 2012). Therefore, it was logical to expect the NAMX and CONUSnestX to over-forecast high wind speeds, however this was not the case. To investigate this bias the nacelle wind speeds and their associated innovations were decomposed into two-dimensional histograms for both NAMX and CONUSnestX simulations (Fig. 6.47). As observed wind speeds increase beyond 5 ms^{-1} , both simulations tend to produce an increasing number of positive innovations. Thus, these results suggest that the current forecast system has a *slow* speed bias as wind speeds increase beyond $5 - 8 \text{ ms}^{-1}$ compared to the nacelle winds at the turbine level (similar to what was found for the RAP model), perhaps indicating an issue of representativeness with the forecast model. More work is needed to investigate this hypothesis. One such approach to investigate this potential bias is to test adding additional vertical levels within the model boundary layer. Recent research has shown that increasing the vertical resolution within the boundary layer can improve turbine-level forecasts (Bernier and Belair, 2012). When superimposed with a sample from a WFIP SODAR platform, known for its high vertical resolution, one can see that the NAM only has a few levels which sample the wind energy generation layer at about 80 m AGL (Fig. 6.48).

Moving forward from WFIP, NCEP is actively taking the appropriate steps to ingest these observations from the private sector in real-time to assimilate into the operational NDAS/NAM for the improvement of real-time, short-range forecasts.

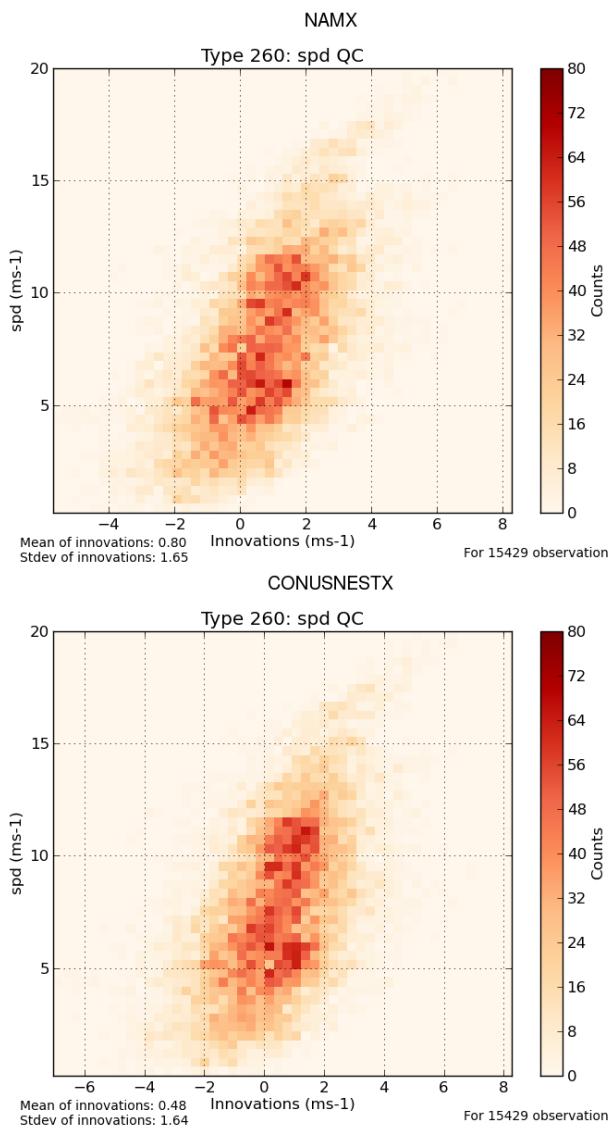


Figure 6.47. Nacelle wind speed observations and innovations depicted as a two-dimensional histogram. Plotted data are from all analysis steps during the WFIP winter quarter data denial period from the NAMX (top) and CONUSnestX (bottom).

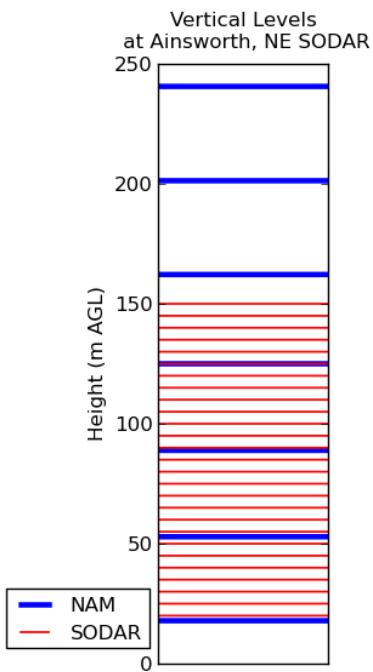


Figure 6.48. A comparison of the number of vertical levels from the NAM vs. the levels at which SODAR observations are reported at the Ainsworth, NE SODAR site.

7. Ramp Tool and Metric

7.1 Background

One of the challenges in integrating weather-dependent renewable energy onto the electric grid comes from the high temporal variability of wind energy. Wind energy production can vary greatly over short periods of time due to the inherent variability of wind speed, which is then amplified by a wind turbine's power curve, which translates the wind speed into power production. Figure 7.1 displays the time series of wind speed measured on a tall tower near turbine hub-height, and then the resulting wind power that would have been produced by a turbine when using a standard IEC2 turbine power curve (Fig. 5.2). The wind power production has long periods of time with either zero power production (for speeds below the turbine's cut-in speed) or near 100% of its maximum capacity production for high speeds. The wind power production frequently jumps rapidly between these two extremes of near zero or near 100% power, and these jumps, referred to as ramp events, can be very rapid due to the wind power increasing approximately as the cube of the wind speed in the middle portion of the turbine's power curve.

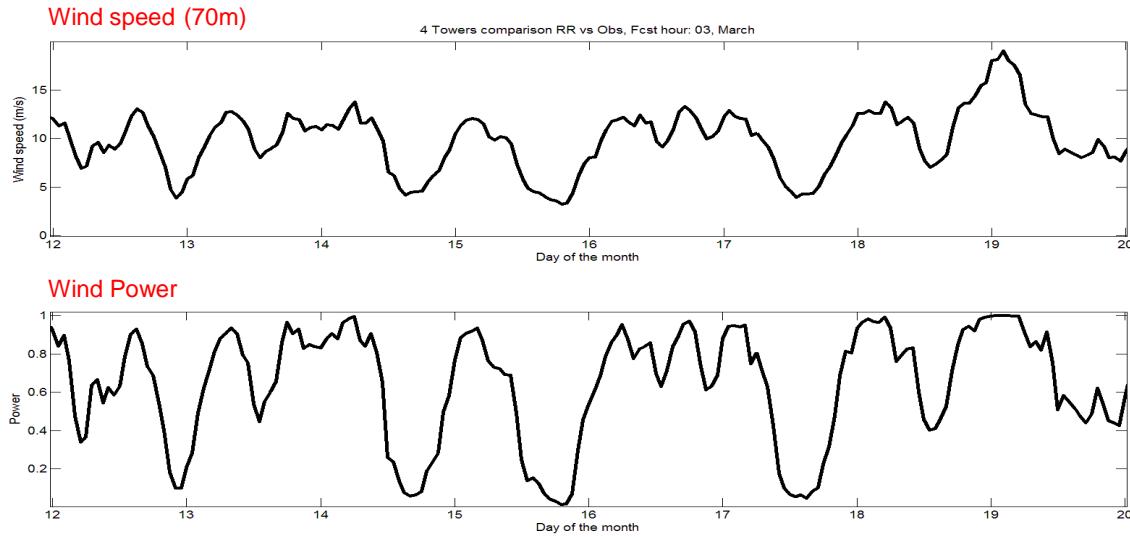


Figure 7.1. Time series of 70m AGL wind speed (top panel) and equivalent wind power (bottom panel) using an ICE2 wind power curve, averaged from the four SDSU tall tower towers in South Dakota, over an 8 day period from March 12-20, 2012.

The essence of a ramp event is a large change in power production over a short interval of time. Ramp events are important for electric grid operators to account for, as they must continuously keep power production in nearly exact balance with power demand. If large, sudden, and unanticipated changes in wind or solar power production occur, the grid operator must keep the grid in balance by making equally large and sudden changes in other conventional energy generation units. These large and sudden changes in conventional generation can increase the costs of power generation, especially if the changes are not forecast accurately, both in terms of their amplitude and their timing.

Standard metrics like MAE and RMSE equally weighted over the entire time series may not be well-suited for wind energy forecast evaluation because the wind power production can be near constant for considerable periods of time, especially near 100% or 0%, while it is the periods of transition between those two states that are most important. A wind ramp metric will provide a statistical measure of the accuracy of the model at forecasting large changes in power over short time intervals. It will weigh model agreement for these events more than model agreement that occurs during periods of near constant power. A ramp metric can be used to compare the skill of two models at forecasting ramp events, for documenting progress in improving a given model, and can also be used to provide a measure of the confidence that a forecast of a ramp of a particular magnitude and duration is correct.

Despite the importance of ramp events for renewable energy generation and grid integration, no commonly accepted definition of a ramp event exists, nor is a single strict threshold defining a ramp possible, as the thresholds for when a ramp becomes important will vary from user to user, and possibly

situation to situation. One of the key goals of WFIP was to develop a ramp tool that has the flexibility to be useful for a variety of users (wind farm developers or owners, utilities, grid operators) and that can be modified or tuned to be useful in a variety of situations. For example it could be used to develop a climatology of ramp events at a given location, or could be used to evaluate the skill of forecasting models at predicting to occurrence and characteristics of ramp events.

The ramp tool described here has three main components. The first is the identification of ramp events in a time series of power data, for which several different methodologies are developed and compared. The second and third components pertain to forecasting of ramp events. The first of these is to match observed ramp events with those predicted by a forecast model. If ramp events are defined such that they are rare events, matching is relatively simple. However, when the definition is relaxed so that ramp events become more frequent both in the observations and forecasts, matching events between the two time series can become more difficult and complicated. The final component of the ramp tool is a methodology for scoring the ability of a model to forecast ramp events. We develop a scoring metric that accounts for both the amplitude and timing (phase) errors in the forecast, and accounts for the different impacts of up and down ramp events. The particular scoring rules that we use are intended to reflect the perspective of a grid operator; however the metric itself is flexible so that it could be easily modified to reflect the needs of other participants in the energy generation system.

We note that the identification and forecast evaluation of ramp events has both similarities and differences with other meteorological phenomena, such as precipitation, air quality indices, severe convection, or droughts, and that the existing extensive set of analysis tools and techniques already developed by the meteorological community can help inform the development of a ramp tool. In this regard it is useful to consider the essential characteristics of ramp events that determine their impact on the electric grid.

The first characteristic of ramps is that they are defined as a rate of change. Since production and demand is always in balance on the electric grid, the forecast problem is to predict how production (and demand) will change from the current state, and in particular how quickly it will change. In this sense forecasting ramp events has similarities with flash flood forecasting, where not just the total amount of precipitation that falls is important, but the rate at which it falls over a given area is also essential.

Second, timing (both onset and cessation) is essential for evaluation of ramp forecasts. The importance of timing down to hourly or sub-hourly resolution is similar to several other forecasting problems, one of which is aviation forecasting. Accurate predictions of the time when an airport will be closed or reopen due to snow, fog, or thunderstorms is essential for the safe and efficient operation of the air traffic system.

Third, the direction of change of ramp events (up or down) is obviously crucial, as it is possible for a down-ramp to occur at the time when an up-ramp was forecast. This seems to have few analogs in conventional weather forecasting, although seasonal forecasting of droughts and floods has similarities.

The combination of rate of change, timing, the directionality of ramp events, and the set of responses of the grid system to those changes, makes for a unique forecasting and forecast evaluation problem.

This section is organized as follows. Section 7.2 presents three different methods for defining ramp events, and compares results from them on time-series of power. Section 7.3 discusses the issues related to ramp matching and presents a simple “closest in time” technique. Forecast skill scoring and model evaluation procedures for a single ramp are presented in Section 7.4, and for a matrix of ramps in Section 7.5. Results from the application of the ramp tool to the WFIP data set are shown in Section 7.6.

7.2 Ramp definition and identification

A ramp event is a large change in power Δp over a short time period Δt , and so can be defined by some combination of Δp , Δt , and the gradient $\Delta p / \Delta t$. There is no absolute definition of a ramp; the definition will depend on the particular application, the application will change from user to user, and possibly even over the course of time for a single user. For example, if a utility has a pumped-storage facility with full capacity, they may be concerned with ramps of one magnitude and duration, but once the reservoir is empty and they have to rely on fossil plant units for load balancing, they likely would be interested in ramps of a different duration and magnitude.

Also, users may require multiple definitions of ramps to be operative simultaneously. For example, if a ramp event has a 60% capacity change in generation over a 4 hour period, it could inform a utility that a certain type of unit needs to be brought on-line. However, if within that 4 hour period there is an embedded ramp with a 30% of capacity change in only 15 min, then a different type of unit may need to be brought online for that 15 min period within the 2 hour duration ramp. For the most efficient scheduling, all of this information needs to be available, and a ramp metric must likewise address a matrix of time and amplitude scales simultaneously to give a realistic measure of forecast skill.

Our starting assumption with the ramp definition is that the power time-series under consideration is the time series that is of importance to the user. For example, if a wind plant operator is forecasting the output from that plant, then the time series considered would be the aggregate power production from the entire plant. If on the other hand a grid operator is concerned with the aggregate power generated by wind and solar over their entire balancing area, then the time-series to be considered would be that larger aggregate power. If this appropriately aggregated time-series is the basis upon which grid operation decisions are made, then it should not be filtered or smoothed in the analysis of ramp events. Filtering routines such as the Swinging Door Algorithm (Bristol, 1990; Florita et al., 2013) have been proposed to compress a time series to a shorter series of linear segments within which smaller changes in the power are ignored as being noise. In general filtering routines such as this are not necessary for defining ramps, and they can have the detrimental effect that the filtered time-series does not contain the full range of power variations present in the original time-series.

7.2.1 Fixed Time Interval Method

Three different methods for identifying ramp events are developed and compared. The first of these methods, referred to as the Fixed Time Interval method, uses a sliding window of length WL and tests if the difference in power between the starting and ending points in the window equals or exceeds a threshold value, $\Delta p = |ps - pe| \geq \Delta p_{RD}$, where Δp_{RD} is the ramp definition threshold. If the threshold criteria is met a ramp exists. If a ramp exists and $ps < pe$ the event is an upramp, if $ps > pe$ it is a downramp. All points within the window are marked as "up" or "down" if an up or down event are found, and separate time series of both all up events and all down events are recorded. The sliding window moves forward one time step, and the process is repeated until the end of the time series is reached. Any contiguous time steps marked as being part of up events are concatenated into a single up event, and the same is done for the time-series of down events. The ramp event can therefore be longer than the window length WL, but each point in the event is part of a window of length WL that satisfies the ramp criteria that $\Delta p \geq \Delta p_{RD}$ within that window.

Each concatenated ramp event is defined by its center time, length, and total power change $\Delta P = |ps' - pe'|$, where ps' and pe' are the power values at the start and end of the concatenated ramp event. The power time-series in Figure 7.2 shows a series of ramp events that would be detected by the fixed time interval method.

Although appealing because of its simplicity, this definition of a ramp event has the possible drawbacks that 1) the selected ramp events may not intuitively look like ramps, since larger values of Δp can occur within the ramp than those defined by its endpoints, and 2) two ramp events, even up and down ramps, can be overlapping in time.

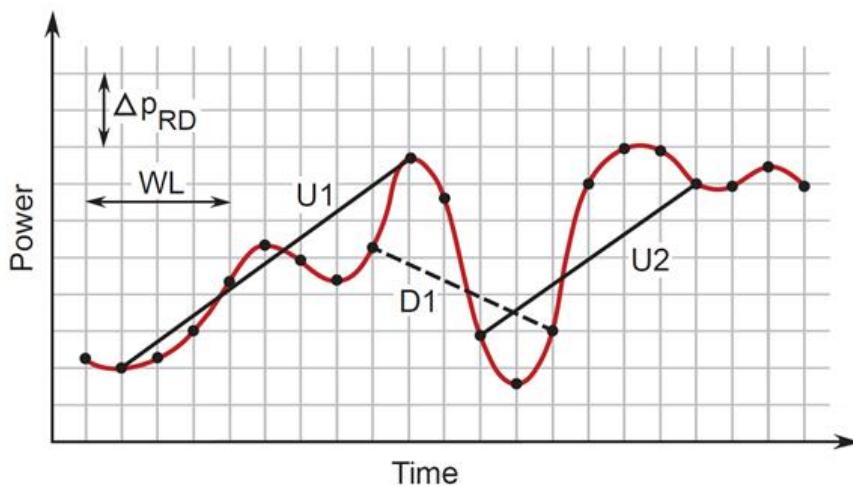


Figure 7.2 Ramps identified by the Fixed Time Interval Method for a window length WL and a ramp threshold Δp_{RD} . Up-ramps are marked using a solid line, down-ramps using a dashed line.

7.2.2 Min-Max Method

The next method, referred to as the Min-Max method, has been used previously for wind ramp detection (Cutler et al., 2007; Greaves et al., 2009; Bossavy et al., 2013), and avoids the two problems previously noted for the Fixed Time Interval method. The Min-Max method finds the maximum amplitude change in power within a sliding window of length WL, and if this change meets the criteria $\Delta p = |p_{max} - p_{min}| \geq \Delta p_{RD}$, then a ramp event occurs. If more than one pair of points within the window meets the threshold criteria, only the largest Δp is defined as a ramp. The initial ramp length is determined by the times t_{min} and t_{max} that correspond to p_{min} and p_{max} , so $\Delta t \leq WL$. The sliding window then moves forward, and a new search is made for (p_{min}, p_{max}) . If new values of either p_{min} or p_{max} are found, the ramp criteria test is applied, and if it is passed a new ramp has been detected. The magnitudes and start/end times of all up and all down ramps are stored for the entire time-series. Ramps of the same sign can occur sequentially, can overlap in time, or can be embedded within another. In these cases the ramps are combined into a single ramp that can have length greater than WL. Because of the use of the min-max values, ramps of opposite signs cannot overlap or be embedded in one another.

Figure 7.3 shows the same power time-series as in Fig. 7.2, but ramp events as detected by the Min-Max method are now displayed. We note that segments of the time series that meet the ramp power threshold criteria, but are not detected because they were not the largest ramp within the window, would be detected if a smaller ramp window WL was used. This highlights the need for a generic ramp metric to span multiple time window definitions, as discussed further in Section 7.5.

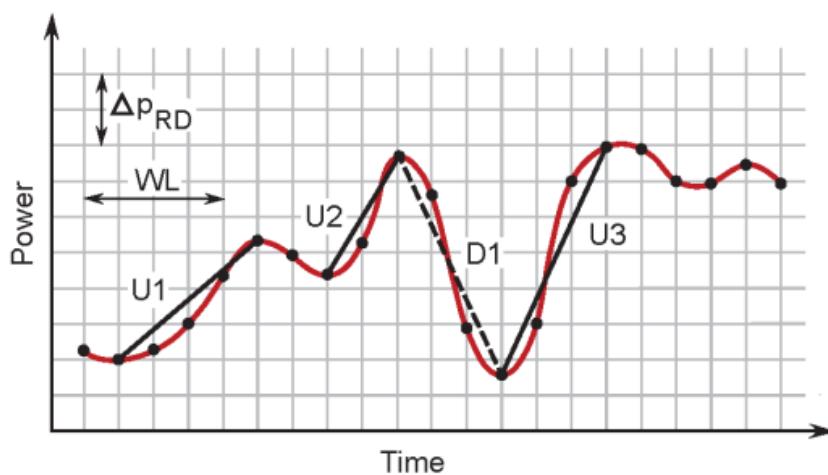


Figure 7.3. Ramps identified by the Min-Max Method for a window length WL and a ramp threshold Δp_{RD} .

7.2.3 Explicit Derivative Method

The third method that we consider is referred to as the Explicit Derivative Method. First, a smoothed time-derivative of the power $\frac{\partial \hat{p}}{\partial t}$ is derived as the slope of a linear least-squares fit to the power over a time window WL. Next, if $\left| \frac{\partial \hat{p}}{\partial t} \right| \geq \Delta p_{RD}$ a ramp exists, and if $\frac{\partial \hat{p}}{\partial t} > 0$ it is an up-ramp, if not it is a down-ramp. The beginning of an up-ramp event is found by searching for a minimum in power over the interval $\frac{1}{2} (WL)$ earlier in time than the first point where the derivative threshold is met, since those points were included in the derivative calculation. The end of an up-ramp is found searching for the maximum in power that occurs in the interval $\frac{1}{2} (WL)$ after the last point of the initial derivative ramp. Similar tests are done for the ends of a down ramp, but first searching for a maximum at the start of the ramp and a minimum at the end of the ramp.

With the derivative method it is possible for two ramps of opposite sign to be partially overlapping in time. Since this occurrence would make it difficult to compare and score model forecast and observation time-series, we modify the explicit derivative results to truncate ramps that overlap. In the period of overlap of a down ramp followed by an up ramp, the minimum value of power is chosen as the end of the down ramp and the start of the new up ramp. If more than one occurrence of the minimum value occurs, then the minimum closest to the down ramp is chosen as its end point, and the minimum closest to the up ramp is chosen as its beginning. If the period of overlap consists of an up ramp flowed by a down ramp, a similar procedure is followed, except that the maximum value of power in the region of overlap becomes the dividing point between the two ramps. Figure 7.4 displays time-series of power and the power derivative, and indicates ramps that are selected by this method. We also note that as in the other two methods, any contiguous time steps marked as being part of ramp events of the same sign are concatenated into a single event, so that the ramp event can be longer than the window length WL.

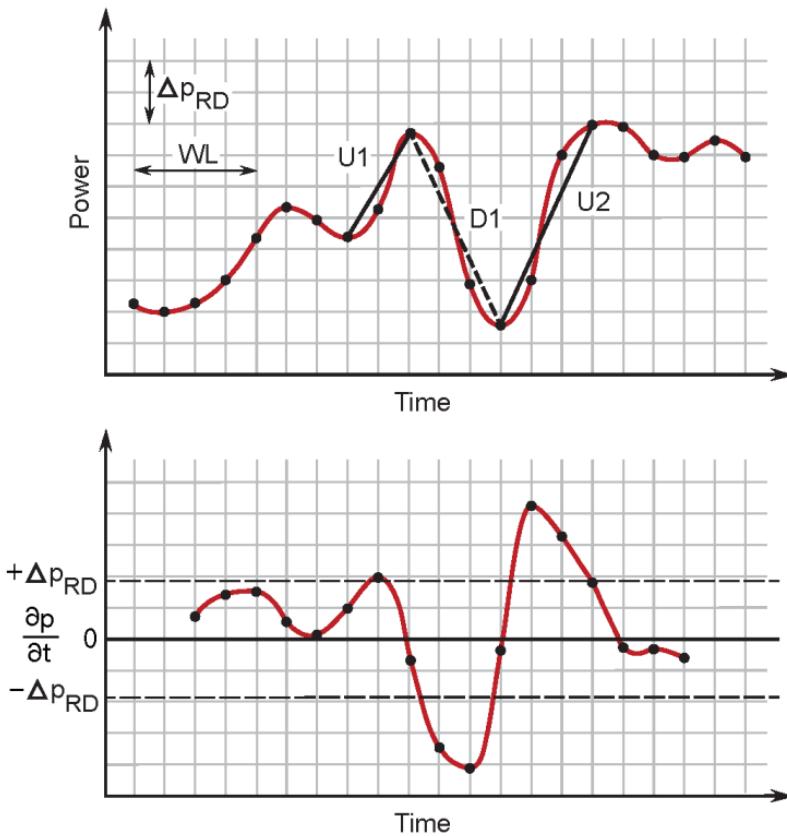


Figure 7.4. Top panel: ramps found by the explicit derivative method for a value of the smoothed derivative threshold given by Δp_{RD} and window length WL . Lower panel: the smoothed power derivative corresponding to the power data in the top panel. The dashed lines indicate the smoothed derivative thresholds defining up and down ramps.

7.3 Matching of forecast and observed ramps

To evaluate the skill of a forecast model in predicting ramp events, the next step is to develop a methodology for matching the observed and modeled events. This process can be complicated by the fact that ramps of opposite signs occur, their occurrence can be relatively frequent for some power thresholds, and phase lags (timing errors) will exist in the forecasts, which themselves can be a function of the length of the forecast (forecast horizon).

The general philosophy in our methodology is to match events that are closest in time; if multiple events have the same time separation, then those of the same sign are matched; if more than one pair of events has the same time separation and sign, those that are closest in power amplitude are matched.

In more detail, the first step is to generate comparable long time-series of model and observed ramp events. For the model, this is done by creating a time-series of forecast power for a particular forecast horizon. For example, a time-series of three hour power forecasts is created by concatenating all three-hour forecasts over a length of time T, where T is much greater than the maximum length forecast (15 h for the RAP and HRRR), and equal in length to the observed time series under consideration. All ramps within the observed and pseudo-model-time series are then found using one of the techniques described in section 7.2. The advantages of using a pseudo-time series of equal length forecasts is that first, a long time series is available, so that problems arising when part of a ramp is present at the beginning or end of a 15 h forecast are avoided. Second, statistics can be developed that are associated with a particular forecast length (e.g., a 4 hour forecast has a ramp forecast skill of X.)

The inputs to the matching algorithm are the time series of length T of ramp events from the observations and forecasts, defined by their center time, sign, and power amplitude. The number of events in the two time-series in general will not be equal. Using these inputs, a matrix of time shifts is created by comparing the center times of each model event with every observed event. A time-series of type of event is also formed for every pair of model and observed events, consisting of +1 for an up/up or down/down pair, and a -1 if it is an up/down or down/up pair. A third time series is created that consists of the difference in power between every pair of model and observed events.

The matrix of time shifts is then searched for the minimum value(s), corresponding to the model ramps that are closest in time to the observed. Frequently events with the same minimum will be found, as there are a limited number of time shifts possible. If more than one minimum exists, all of the minima are evaluated for instances when one model ramp is paired with two equally spaced (preceding and following) observed ramps. If the model value is paired with only one observation, these two events are matched, and then eliminated (in all three matrices) from any further searching. If the model ramp is paired with two observed events at the current time-shift minima, the choice of which one is matched with the model event is then made based on the “type of event” matrix and “difference in power” matrix. First, if the two observed ramps are of opposite sign, the one that is the same sign as the model ramp is selected as the match. If both observed ramps are of the same or the opposite sign as model ramp, the observed ramp with the smaller value in the “difference in power” matrix is selected as the match. Again, once a modeled and observed ramp event are matched, both are removed from all three matrices, and the search for a minimum in the time shift is applied again. This process of searching through all of the model ramps is repeated until either all of the model ramps are matched or determined to be unmatched up/null or down/null events. Next the same process is repeated for the remaining unmatched observed ramps to determine if they are unmatched null/up or null/down events.

7.4 Forecast skill scoring methodology using single ramp definition

The ramp identification and ramp matching procedures result in time series of matched pairs of ramps or unmatched events, each defined by their power gain/loss ($\Delta p_f, \Delta p_o$), length of event, and their center times (CT_f, CT_o). Using this time series of events, a forecast score is determined by

comparing the forecast and observed characteristics of each event. The ramp skill score accounts for model ramps that have been matched to observed ramps, model ramps that have not been matched with observed ramps, and observed ramps that have not been matched with model ramps. The skill score is intended to represent a utility operator's perspective for different up/down ramp scenarios, incorporating phase and amplitude errors when needed, and recognizing that up and down ramp events can have different impacts on grid operation. Additionally, the skill score is designed so that a set of random forecasts will have near zero skill. A negative skill indicates the model is worse than random, and any positive value indicates the model has value. Although a specific set of rules is applied, it is not possible to generate a single set that would be applicable for all users in all situations. These rules should be viewed as one particular realization, and are meant to be modified by users to best suit their particular circumstances.

The first step is to classify the different types of ramp scenarios that are possible, indicated in Table 7.1. This is similar to a 3x3 contingency table (Wilks, 2006) consisting of up, down and null events, except that the null-null case is not considered and does not affect the skill score. Near-zero scores are then assigned to the instances involving null (un-matched) events, which are scenarios 2, 5, 7 and 8 (Table 7.2). Although these null events could be considered to have no forecast skill, slightly positive non-zero values are applied to the instances when the poor forecast results in total power supply greater than demand (scenarios 5 and 7), which can be solved by curtailing wind energy generation. In contrast, slightly negative non-zero values are applied to the instances when the poor forecast results in total power supply being less than demand (scenarios 2 and 8), which could require a spot market power purchase that in general would be more expensive to the utility than wind curtailment. For an equal distribution of null scenarios 2, 5, 7 and 8, the model will have zero skill.

Scenario	Model	Observed
1	Up	Up
2	Up	Null
3	Up	Down
4	Down	Down
5	Down	Null
6	Down	Up
7	Null	Up
8	Null	Down

Table 7.1. Scenario definitions for matched and un-matched ramp events.

Scenario	Model	Observed	Utility Action	Score
2	Up	Null	Fossil fuel (FF) spot market purchase or Cancel planned decrease of FF units	-0.1
5	Down	Null	Wind curtailment or Cancel planned increase of FF units	+0.1
7	Null	Up	Wind curtailment or Decrease existing FF units	+0.1
8	Null	Down	FF spot market purchase	-0.1

Table 7.2. Scores for the four possible null scenarios.

For the non-Null scenarios a range of scores is possible depending on the forecast's phase and amplitude errors, as shown in Table 3. Correct up/up and down/down events have scores ranging from +1.0 (for zero phase lag and zero amplitude error) to zero (when the phase lag and/or amplitude error become large). Incorrect non-null forecasts of up/down or down/up have mostly negative scores. However, these are modified on the smaller end of their range by applying the constraint that as the phase error becomes large for scenarios 1, 3, 4 and 6, the score asymptotes to that for the corresponding pair of null events.

$$\begin{aligned}
 \text{MinScoreScenario 1} &\rightarrow \text{ScoreScenario 2} + \text{ScoreScenario 7} \\
 \text{MaxScoreScenario 3} &\rightarrow \text{ScoreScenario 2} + \text{ScoreScenario 8} \\
 \text{MinScoreScenario 4} &\rightarrow \text{ScoreScenario 5} + \text{ScoreScenario 8} \\
 \text{MaxScoreScenario 6} &\rightarrow \text{ScoreScenario 5} + \text{ScoreScenario 7}
 \end{aligned}$$

For example, as the time separation between a forecast/observed up/up event (scenario 1) becomes large, the model skill asymptotes to the sum of an up/null and a null/up event (scenario 2 + scenario 7). This constraint forces the scores for scenarios 3 and 6 to approach values of -0.2 and +0.2 for large phase lags, given the previous assumptions of small but non-zero score values for scenarios 2, 5, 7 and 8 given in Table 7.2. On the more extreme end of the skill range, the up/down event (scenario 3) also has a worse score (-1.2) than a down/up event (scenario 6). This is consistent with the fact that an up/down ramp event, with any phase lag, will have a more costly impact on grid operations than a down/up event, as the former will require a spot market power purchase, while the later can be solved by wind curtailment.

Scenario	Model	Observed	Score
1	Up	Up	+1.0 to 0.0
2	Up	Null	-0.1
3	Up	Down	-1.2 to -0.2
4	Down	Down	+1.0 to 0.0
5	Down	Null	+0.1
6	Down	Up	-0.8 to +0.2
7	Null	Up	+0.1
8	Null	Down	-0.1

Table 7.3. Range of scores possible for all 8 event scenarios.

Equations are derived to compute the scores for the non-Null scenarios (1, 3, 4, and 6) that take into account the timing and amplitude of the forecast ramp compared to the observed ramp, and can include penalties when the model over-predicts the amplitude or predicts its occurrence later than observed. The score for an individual ramp event in one of these four scenarios is given by:

$$Score\# = MaxAmpScore\# * (\sqrt{\alpha * \tau}) + MinAmpScore\# * (1 - \sqrt{\alpha * \tau})$$

where the # sign refers to the 4 non-null ramp event scenarios, and MaxAmpScore is the first (most extreme) value and MinAmpScore the second value of the range of scores shown in Table 7.3 for each of the non-null scenarios. The forecast timing and amplitude skill parameters τ and α are defined as the linear equations:

$$\tau\# = \left[1 - \frac{|CT_f - CT_o|}{WL} \right] * F_T\#$$

$$\alpha_{1,4} = [1 - |\Delta p_f - \Delta p_o|] * F_A\#$$

$$\alpha_{3,6} = \left[\frac{|\Delta p_f - \Delta p_o|}{2} \right] * F_A\#$$

F_T and F_A are functions that are unity except in the cases when a late prediction penalty or a sign prediction penalty occurs (discussed below) in which case $0 \leq F_T, F_A \leq 1$. With these definitions the timing skill falls in the range $0 \leq \tau \leq 1$, and the amplitude skill falls in the range $\Delta p_{RD} \leq \alpha \leq 1$. The amplitude skill α is unity if the observed ramp amplitude identically matches the forecast amplitude and

both are greater than the threshold p_{RD} . The minimum value of α occurs when one of either the forecast or observed ramp amplitudes is Δp_{RD} and the other is 1. The minimum amplitude skill cannot become less than Δp_{RD} if an observed and forecast ramp are found that both meet the minimum threshold criteria Δp_{RD} .

The score equation has the properties that for up/up and down/down events, when the forecast has no timing error *and* no amplitude error ($\alpha = 1, \tau = 1$) the score is equal to the most extreme value (MaxAmpValue) for that scenario in Table 7.3, and when the timing error equals its maximum allowed value (the window length WL) score is equal to its value closest to zero (MinAmpValue) in Table 7.3 value for that scenario. For an up/down or down/up event, the largest the amplitude error can be is 2, $\alpha_{3,6} = 1$, and the score becomes MaxAmpValue in Table 3, while the smallest the amplitude error can be is $2 \Delta p_{RD}$, in which case $\alpha_{3,6} = \Delta p_{RD}$, and the score becomes $\Delta p_{RD} * \text{MinAmpValue}$.

The last option in the scoring is to include values different than unity for the Late Prediction Penalty (LPP) or Under/Over Prediction Penalty (UOPP) functions F_T, F_A . The LPP occurs only for those scenarios when a late forecast event is worse for a utility than an early forecast because it implies a spot market power purchase rather than wind curtailment. Likewise, the UOPP occurs only for those scenarios when the sign of the difference between forecast and observed ramp amplitudes negatively impacts grid operation.

The Late Prediction Penalty for the 4 non-null cases is given by

$$F_T\# = \left[1 - \frac{(CT_f - CT_o) * \text{WeightLPP}\#}{WL} \right], \quad \text{iff } CT_f > CT_o$$

otherwise $F_T\# = 1$.

The Under/Over Prediction Penalty is given by

$$F_A\# = 1 - \frac{(\Delta p_f - \Delta p_o) * \text{WeightUOPP}\#}{1 - \Delta p_{RD}}, \quad \text{iff } \Delta p_f > \Delta p_o$$

otherwise $F_A\# = 1$.

Scenario	Model	Observed	Weight LPP	Weight UOPP
1	Up	Up	0	0.15
3	Up	Down	0	0
4	Down	Down	0.15	0.15
6	Down	Up	0	0

Table 7.4 Weights applied for ramp Late Prediction Penalties and Under/Over Prediction Penalties.

There is no LPP for scenario 1 since a late forecast will incur curtailment of wind power but no spot market power purchase. The UOPP for Scenario 1 is applied (WeightUOPP1 = 0.15) when the forecast over-predicts the ramp amplitude because this will require a spot market power purchase to cover the shortage of wind power actually produced. If the wind ramp power amplitude is under-predicted, the wind simply will be curtailed and $F_A\# = 1$.

For Scenario 3 (up/down) there is no LPP since a late forecast will be less onerous for a grid operator than one that predicts an up event at the precise time that an observed down ramp occurs. There is no additional UOPP for Scenario 3 since the forecast is always over-predicting the power.

The LPP for Scenario 4 (down/down) is applied (WeightLPP4 = 0.15) when the forecast is late because this will require a spot market power purchase to cover the shortage of wind power actually produced. If on the other hand the down ramp is forecast early, then wind will need to be curtailed or fossil fuel generation scaled back if possible and $F_T\# = 1$. The UOPP for Scenario 4 is applied (WeightUOPP4 = 0.15) when the forecast ramp power amplitude is smaller than the observed because this will require a spot market power purchase to cover the shortage of wind power actually produced. If the ramp amplitude is over-predicted, the wind simply will be curtailed or fossil fuel generation will be turned down if possible and $F_A\# = 1$.

The final non-null scenario, Scenario 6 (down/up) has no LPP since a late forecast will incur curtailment of wind power but no spot market power purchase. Also, there is no UOPP for scenario 6 since the forecast is always under-predicting the power.

We note that when the LPP and UOPP are applied (WeightLPP = 0.15 and WeightUOPP = 0.15) τ and α decrease quadratically rather than linearly as the timing error and amplitude errors increase. Also, the tunable penalty weights here have been set to nominal values of 0.15 for both LPP and UOPP, but any value could be used.

7.5 Forecast skill scoring: Matrix of skill values

The ramp metric developed above applies to ramps defined by a single power amplitude threshold and window length. By itself, however, the forecast skill for this single definition of a ramp does not provide the full measure of the value of the forecast to a utility or grid operator. For example, perfect forecasts of a 30% power capacity up-ramp over two hours and a 90% up-ramp over two hours will both have forecast skills of 1.0, yet forecasting the larger ramp will have more value than the smaller ramp.

Also, if one is trying to answer to identify which of several models has the highest skill at forecasting ramps, the answer may change depending on which power threshold and which window length are used. Ideally, one would like to know which model is best for a range of power thresholds and window

lengths that could reasonably be of importance for grid operations, and to then average the model's skill over this range of values.

As a means to condense forecast skill for a range of ramp thresholds, we consider a matrix of ramp skills. Figure 7.5 illustrates the concept of a matrix of ramp skills in a schematic form. Skill scores as defined in Section 7.4 for particular values of power threshold and window length are calculated and placed into each matrix element. Skill scores using extreme ramp definitions, with the largest power thresholds and the shortest window lengths, are placed in the top left corner of the matrix. Skill scores for more frequently occurring and weaker ramp events, defined using lower power thresholds and longer window lengths, will be placed in the bottom right corner of the matrix.

In place of averaging the ramp skill scores in all of the matrix elements equally, a weighting function is applied before averaging that accounts for the fact that the skill scores for the more extreme events will have a greater impact on grid operations than the weaker ramps.

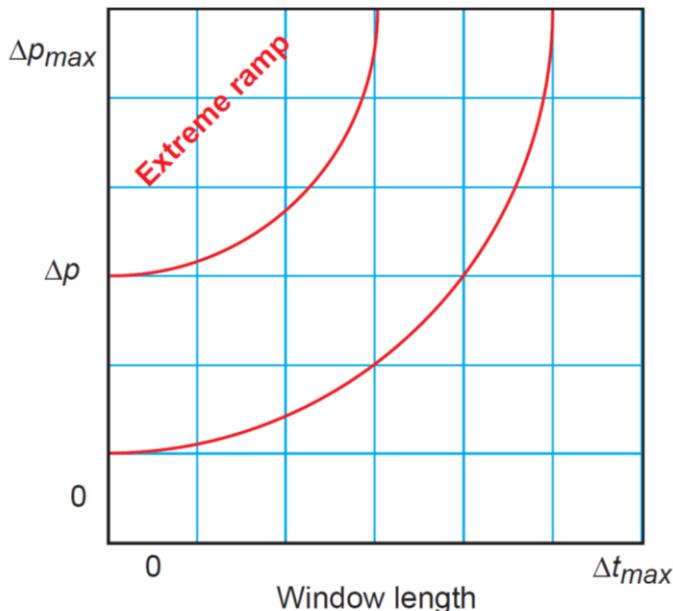


Figure 7.5 Schematic diagram of a weighted ramp matrix. Extreme ramps, with large changes in power over short time intervals, are placed in the top-left corner, while low amplitude ramps of longer duration are placed in the bottom right corner. A weighting function, denoted by the red isopleths, is then applied to each matrix element, before averaging into a single overall model skill score.

7.6 Results from the WFIP data denial experiments

The ramp metric tool has been applied to the WFIP observations and model forecasts from the 6 control (no WFIP observations assimilated) and experimental (with WFIP observations assimilated) simulations in the data denial experiments, for all three of the ramp definition methods. An example of ramp identification using the Min-Max method from the September DD episode, applied to observations from

the SDSU tower FAH, is displayed in Fig. 7.6. The top panel shows the computed power from the tower 80m anemometer, the middle panel the power derived from the RAP control simulation, and the bottom panel the power from the RAP experimental simulation that assimilates the special WFIP observations. The model values are shown at the model initialization time, hour 00. The model values are at 15 min resolution, and the original 10 min observed values have been interpolated to the same 15 min intervals of the model. Using a ramp definition of a power change greater than 50% over a nominal 2 h period, in the observations 6 up ramps are found (shown in red) and 8 down ramps (shown in green). In contrast, the control simulation finds only 2 up and 2 down ramps, while the experimental simulation finds 5 of each, with the additional ramps generally matching up well with the observed ramps in the top panel.

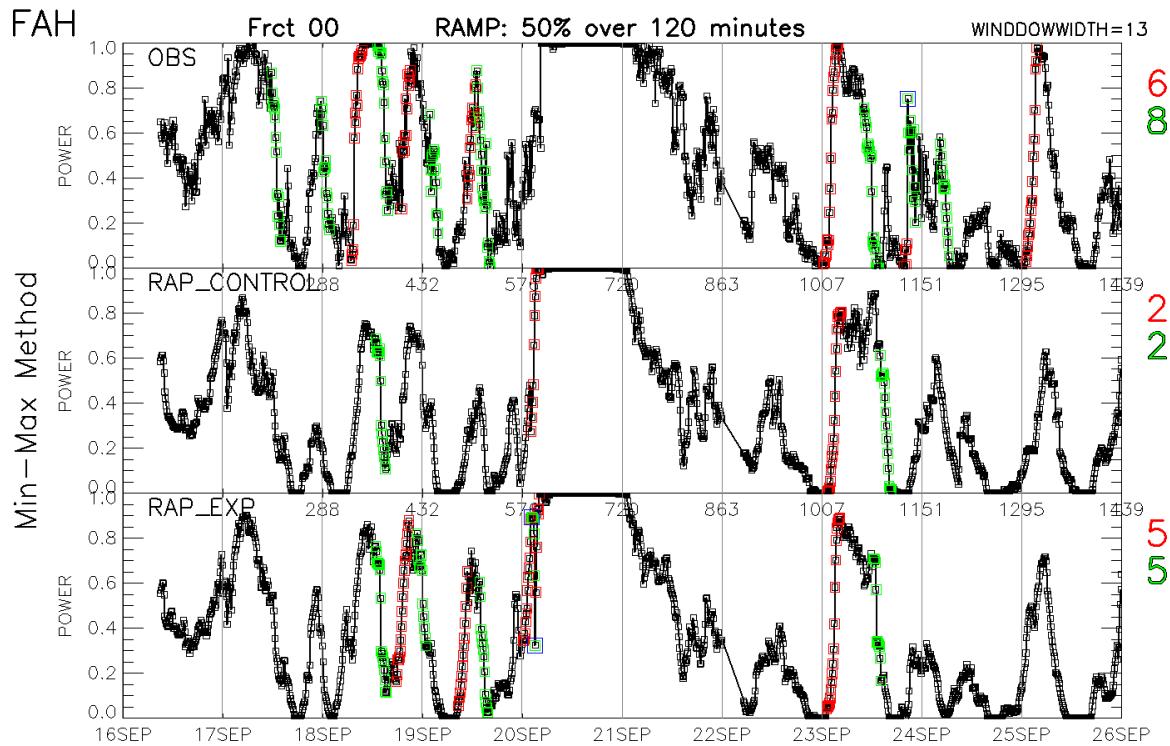


Figure 7.6 Time series of power estimated from anemometer measurements on a tall tower and from the RAP control model. Ramp events for the three ramp definition methods are shown using a 50% power change threshold over 2 hours, red for up ramps and green for down. The numbers of ramps found in each time series are shown on the right.

Figure 7.7 displays the number of occurrences of ramp events found using the Min-Max ramp definition method, for forecast hour 00 (initialization time, left panels) and hour 06 (right panels), for the RAP control (top panels) and experimental (bottom panels) simulations, and for a range of ramp power thresholds from 30 to 70%, and window lengths of 30, 60, 120 and 180 minutes. Relatively few extreme ramp events are found (top left corner of each panel) while many small amplitude and long ramps are found (bottom right corner of each panel). A slightly larger number of events is found at forecast hour

06 compared to the initialization time, indicating that the RAP's initialization procedure leads to slightly smoother wind fields at hour 00. The differences between the number of events in the control and experimental simulations are also relatively small, indicating that although the assimilation of the special observations may change the strength and timing of the ramp events, it does not significantly change their overall numbers. Summing the number of events in the matrices at forecast hour 00 gives 12.1 for both the control and experimental simulations, while at forecast hour 06 there are 15.0 for the control and 15.3 for the experimental simulation. In contrast to the similarity between the four panels of model counts, the number of occurrences in the observations is much larger (note the different scale on color bar). This is because individual towers are used in the analysis, and at 13 km resolution the model has considerably smoother fields than the observed point location power time series.

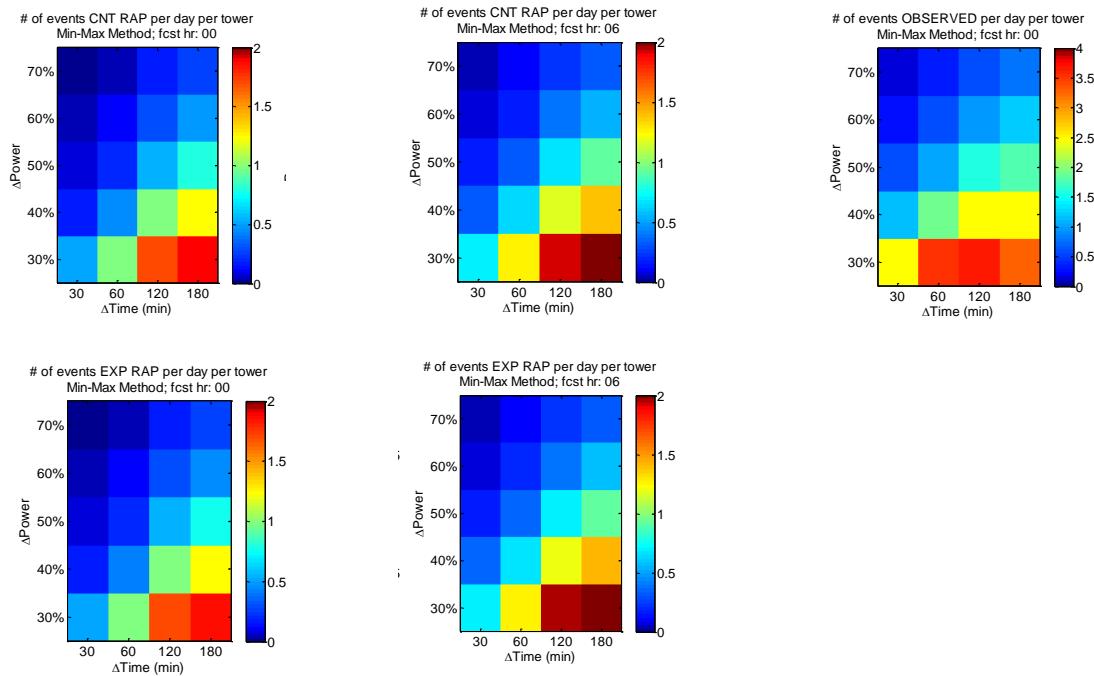


Figure 7.7. The average number of occurrences of ramp events that fall into each matrix bin per DD episode using the Min-Max Method, for the forecast initialization time (hour 00, left panels) and forecast hour 06 (middle panels), for the control (top) and experimental run (bottom) of the ESRL RAP model, NSA and SSA combined. The top right panel is the same but for the tall tower observations. The ramp definition power threshold ranges from 30 to 70%, and the window length from 30 to 180 minutes.

Skill score matrices for the RAP control simulations are shown in Fig. 7.8 for the initialization time and hourly forecasts out to 7 hours, averaged for the 6 DD episodes, NSA and SSA combined. In combining the two study areas, each area is also weighted by the number of tower locations in that area that were used to compute the statistics. The skill is largest for long window lengths (180 minutes) but is quite uniform across all amplitudes of ramps. The skill is greatest at the initialization time, and slowly decays

with forecast length. Skill scores for the experimental RAP simulations look qualitatively similar and are not shown.

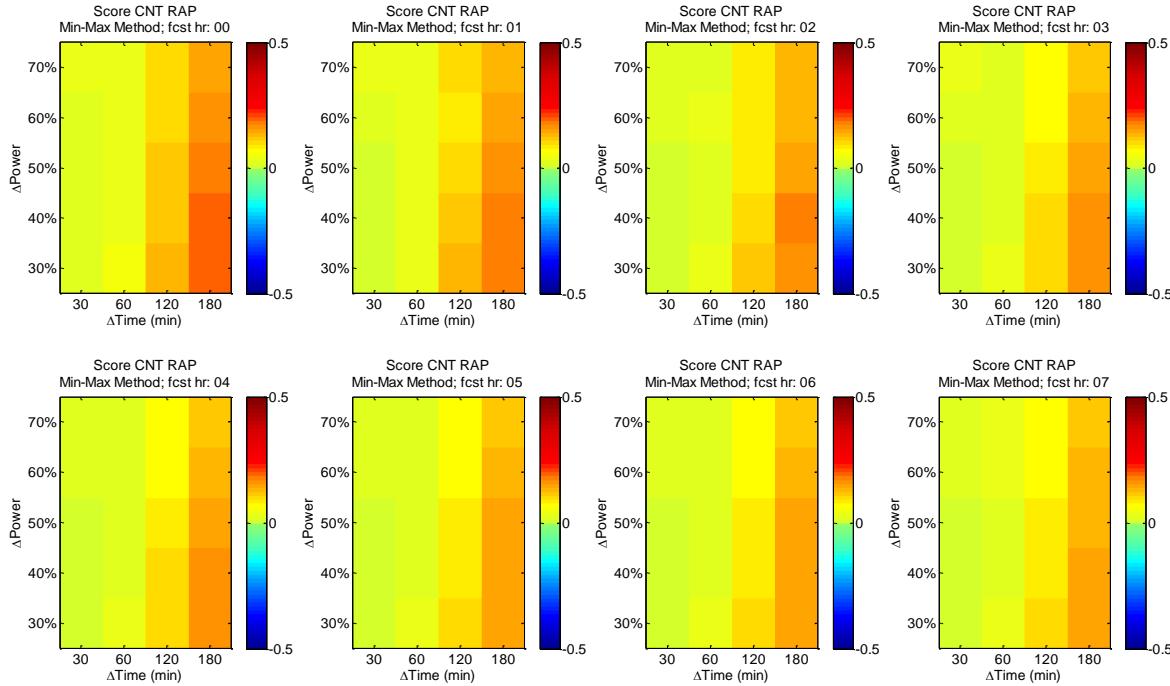


Figure 7.8. The matrix of skill scores for the control ESRL RAP forecasts averaged for all 6 DD episodes, NSA and SSA combined.

In order to derive a single skill score for the model, a weighting matrix is required. The weighting matrix that we have used is simply to start with a weight of 1.0 in the top left corner (most extreme ramp) and to decrease the weight by 10% for each 10% change in ramp power threshold and each increment in window length, as shown in Fig. 7.9.

Δp_{max}	1.0	0.9	0.8	0.7
Δp	0.9	0.8	0.7	0.6
	0.8	0.7	0.6	0.5
	0.7	0.6	0.5	0.4
	0.6	0.5	0.4	0.3
Window length	WL ₁	WL ₂	WL ₃	WL ₄

Figure 7.9. The default weighting scores applied to the matrices of skill scores shown in Fig. 7.8.

The average score across the entire matrix is shown in the top panel of Figs. 7.10 for the fixed time-interval method, using both an equal weighting of all the matrix elements (blue lines), and when using the weighting matrix (red lines). The solid lines are for the RAP control simulation, while the dashed lines are for the experimental simulation. The un-weighted averaged skill score is greater than the weighted skills core, because the model has less skill at forecasting the most extreme ramps with the largest power changes over short window lengths. The experimental simulation (dashed line) is seen to have greater skill than the control simulation for the first 9 forecast hours, after which the differences between the two are negligible. The lower panel of Fig. 7.10 shows the percent improvement of the experimental simulation compared to the control. The weighted improvement averages to approximately 7% from forecast hours 1-9, with peak improvements reaching 11%.

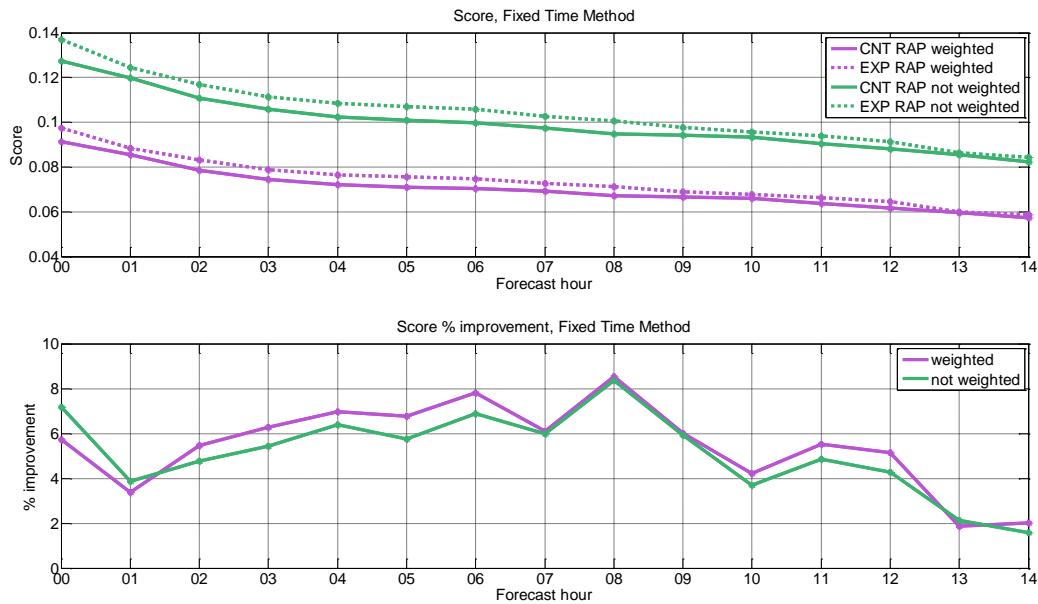


Figure 7.10. Skill score results for the Fixed-Time ramp identification method, averaged for all DD episodes, NSA and SSA combined. Top panel: Skill scores for the Control (solid) and Experimental (dashed) data denial simulations, for forecast hours 0-14. Green is the skill score with equal weighting of all matrix elements, purple is when using the weighting matrix shown in Fig. 7.9. Bottom panel: the percent improvement in the experimental forecasts over the control, for the un-weighted (green) and weighted (purple) matrices of skill scores.

The averaged skill scores for the RAP data denial simulations for the Min-Max and Explicit Derivative methods are shown in Figs. 7.11 and 7.12. For the Min-Max method the weighted improvement averages to approximately 9% for forecast hours 1-9, with peak improvements reaching 14%, while for the Explicit Derivative method the weighted improvement averages to approximately 7% for forecast hours 1-9, with peak improvements reaching 10%. Notably, the improvement in skill score does not decrease as quickly with forecast length as do the standard bulk statistics shown in Section 5, but is much more constant over the first 9 forecast hours.

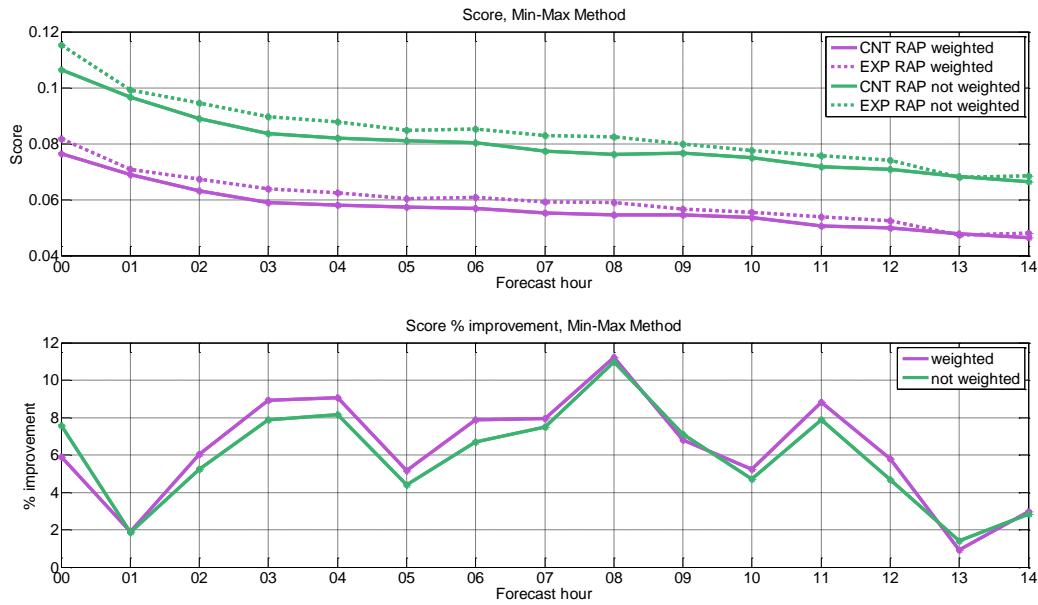


Figure 7.11. The same as Fig. 7.10, except for the Min-Max ramp detection method.

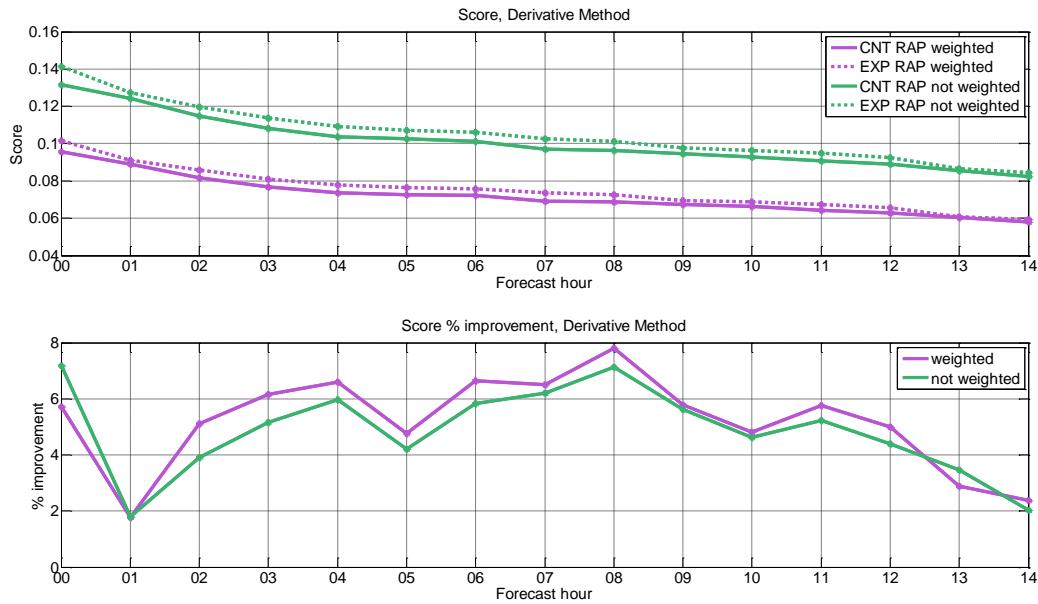


Figure 7.12. The same as Fig. 7.10, except for the explicit derivative ramp detection method.

A breakdown of the ramp events into each of the 8 different possible scenarios is shown in Fig. 7.13. The largest number of events by far occur for the two model null-event scenarios (7 and 8), which reflects the fact that the observational data has much more temporal variability than the smoother model time series, so that many observed ramps have no match in the model time series. If only

longer duration ramps were included in the ramp definition, or if the ramp metric were applied to spatially averaged data, many fewer model null occurrences (scenario's 7 and 8) would occur. The next most common events are scenarios 1 (up/up) and 4 (down/down). Scenarios 2 (up/null) and 5 (down/null) are the next most common, but have much lower rates of occurrence than scenarios 1 and 4. The least common of all are scenarios 3 (up/down) and 6 (down/up). Again the number of these inverse events would decrease greatly with longer duration ramp definitions and with any degree of spatial averaging. Assimilation of the WFIP observations results in more up/up (scenario 1) and down/down (scenario 2) events, as well as fewer null/up (scenario 7) and null/down (scenario 8) missed forecast events. For the up/down (scenario 3) and down/up (scenario 6) obverse events, assimilation of the WFIP observations makes little difference on average across all the forecast hours shown. The up/null (scenario 2) and down/null (scenario 5) are the only events where assimilation of the WFIP observations does not improve the forecasts. The reason for this is not clear.

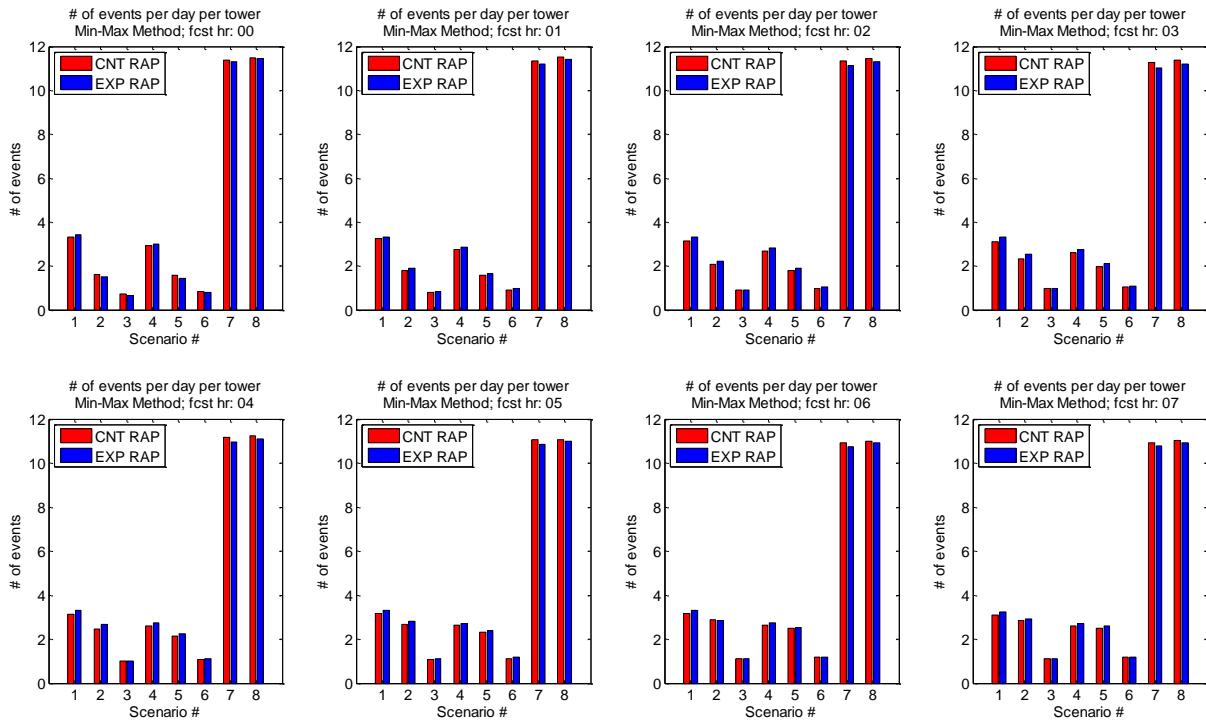


Figure 7.13. The number of matched ramp events per day and per tower site, in each of the 8 ramp scenarios using the Min-Max method for all 6 DD episodes, NSA and SSA combined. Red is for the control simulation, and blue for the experimental. All 20 Δp and Δt combinations shown in Fig. 7.8 are summed to form the number of events.

The ramp skill scores can also be broken down into each scenario category. Fig. 7.14 shows the sum of the matrix-weighted scores average for all 6 DD episodes, NSA and SSA combined. Considering first the control simulation, the positive contribution to skill score comes from scenarios 1 and 4, when the

forecast accurately predicts up ramps when they are observed, and down ramps when they are observed. Scenarios 7 and 8 (model forecast null events when an observed up or down ramp occurs) have on average no net effect, as do scenarios 2 and 5 (observed null events when a forecast up or down ramp occurs). Scenarios 3 and 6 (ramp forecasts with the opposite sign than is observed) have a negative contribution that is much smaller than the positive contribution of scenarios 1 and 4.

The improvement in forecast skill in the experimental simulations also comes solely from scenarios 1 and 4, with greater positive summed scores for the experimental simulations (blue bars) than the control (red bars). There is no net change in summed scores for scenarios 7 and 8, while for scenarios 2 and 5, and 3 and 6, the net impact is very small or slightly negative. It is somewhat surprising that assimilation of the WFIP observations improves the skill of correctly forecast events (scenarios 1 and 2) but does not improve the scores of inversely forecast events (scenarios 3 and 6). One possible reason might be if the observed ramps that contribute to the inverse events are very local, occurring at a single tower site or over very small geographic areas, assimilation of that single tower with may not have sufficient weight to alter the forecast initialization, as it shouldn't since it is such a localized event.

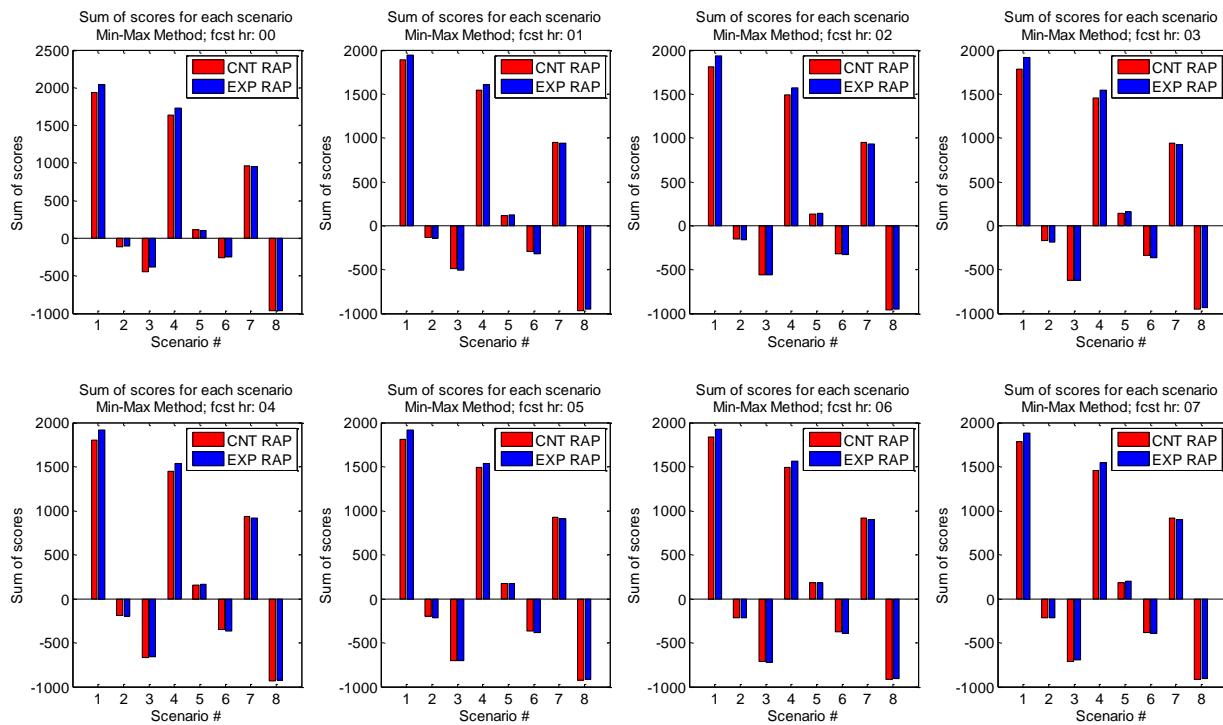


Figure 7.14. Sum of the matrix-weighted scores for each ramp scenario type, for all 6 DD episodes, NSA and SSA combined.

The next component of the ramp analysis for WFIP is to examine the differences in ramp forecast skill between the different DD episodes and between the NSA and SSA. Figure 7.15 shows the percent

improvement in ramp forecast skill averaged over forecast hours 1-9, for each of the 6 DD episodes, and for the NSA (blue) and SSA (red). Improvements in the NSA are larger than the SSA, averaging 10.6% for the NSA versus 3.6% for the SSA (consistent with the bulk statistics MAE improvement found in Section 6). Also, the improvement in the NSA is more uniform across the 5 DD episodes than it is in the south. Determination of the reason for the differences between the NSA and SSA will require additional research. Hypotheses include, first, that the new WFIP observations were spread over a wider geographic area in the NSA than in the SSA, allowing for the model initial field improvements to be more robust and affect a wider area, thereby having a more lasting positive impact before advecting away from the area of interest. Second, the NSA had more tall tower observations, more wind profiler observations, and the addition of nacelle anemometer observations; the greater numbers of observations seems likely to have also contributed to the greater improvement in ramp forecast skill.

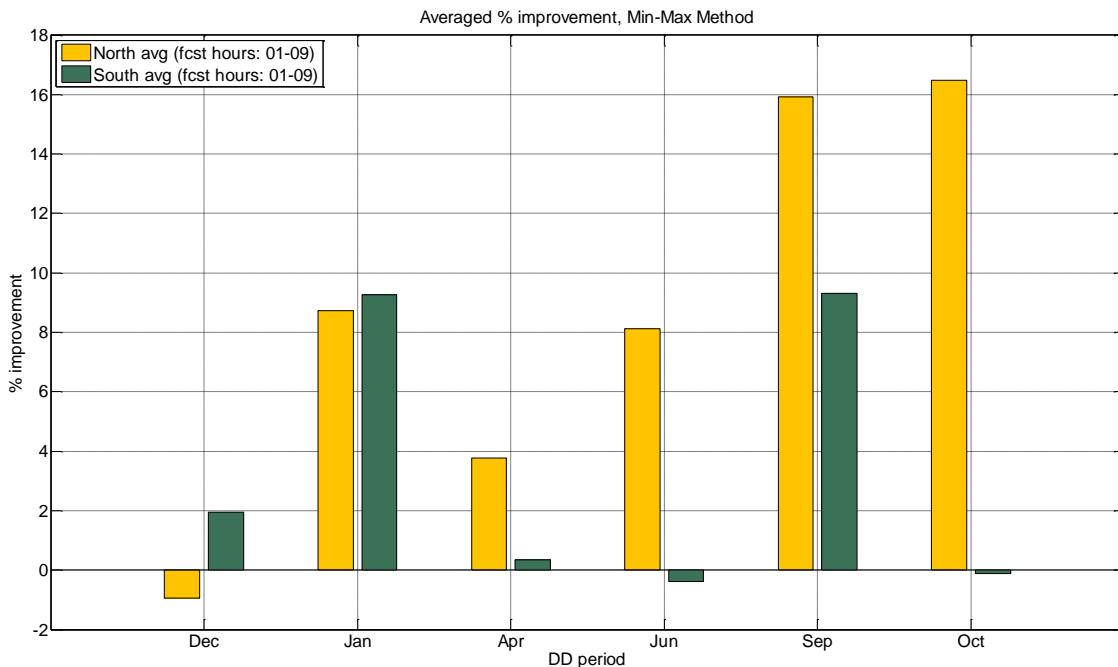


Figure 7.15 Percent improvement in ramp forecast skill using the Min-Max ramp definition, averaged for the first 9 forecast hours, for each of the 5 DD episodes, and for the NSA (orange) and SSA (green).

The final component of the ramp analysis is to investigate the model skill at up versus down ramps, and the impact of the WFIP data assimilation on both. To more fairly compare the model skill for up and down ramps, a simplified symmetric set of scoring rules was applied, as listed in Table 7.5, with no late prediction or over/under prediction penalties.

Scenario	Model	Observed	Score
1	Up	Up	+1.0 to 0.0
2	Up	Null	0.0
3	Up	Down	-1.0 to 0.0
4	Down	Down	+1.0 to 0.0
5	Down	Null	0.0
6	Down	Up	-1.0 to 0.0
7	Null	Up	0.0
8	Null	Down	0.0

Table 7.5 *Range of scores possible for all 8 event scenarios for a simplified, symmetric up and down ramp scoring scheme.*

As can be seen in Fig. 7.16, down ramps (red curves) are more difficult to predict than up ramps (blue curves). Also, the solid lines are for the control that does not assimilate in the WFIP obs, while the dashed line is for the experimental simulations that do assimilate the WFIP observations. Assimilation of the WFIP obs increases the forecast skill, and this increase can be seen for the first 12 forecast hours.

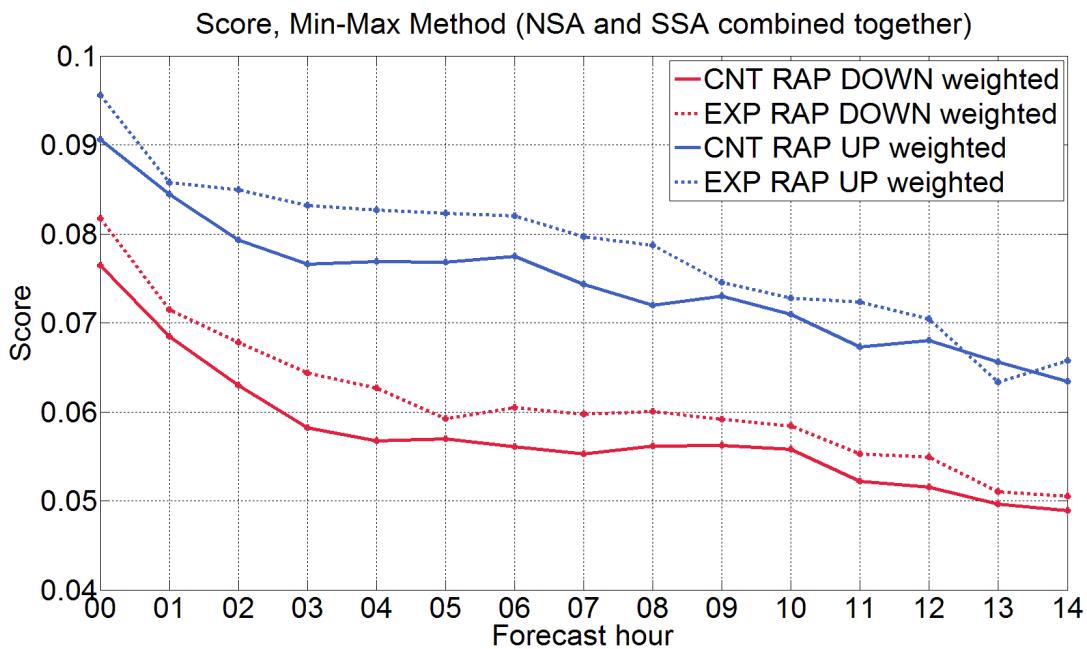


Figure 7.16. Ramp forecast skill for up (blue) and down (red) ramps using a symmetric set of scoring scenarios, using the weighted matrix of events. Solid lines are for the RAP control, and dashed lines for the RAP experimental simulations assimilating the WFIP observations.

8. Surface flux and wind profile observations

Sonic anemometer measurements were made by the Field Research Division of the Air Resources Laboratory of NOAA at the Brady airport (BDY), Colorado City airport (COC), and Jayton (JTN) National Profiler Network sites, all in Texas. The purpose of these measurements was to examine the relationship between measured turbulent fluxes near the surface and wind speed profiles extending to heights where wind turbines operate. Some highlights of these measurements will be presented here.

The sonic anemometer at BDY was located at a height of 3.19 m. Wind speed profiles for the estimation of displacement height d and roughness length z_0 were obtained using the sonic anemometer, a cup and vane anemometer at a height of 10 m on the same tower, and a nearby Atmospheric Systems Corporation (ASC) sodar measuring wind speeds in 10 m intervals from 30 m to 200 m. Most of the fetch for these measurements consisted of closely cropped and often dry grass with patchy scattered shrubs in the distance. Groves of trees several meters in height were present in some approaches within 100 m of the measurements (between approximately 90° and clockwise to 225°). A method was developed for the estimation of d and z_0 in neutral conditions excluding wind directions between 90° and 225°. The detailed description of the method for determining d and z_0 is beyond the scope of this summary, but it is based on Bayesian statistical methods. Neutral conditions were defined as the absolute value of the

kinematic heat flux less than $0.01 \text{ m s}^{-1} \text{ }^{\circ}\text{C}$ and wind speeds greater than 7 m s^{-1} . Values of 3.6 and 2.0 cm were determined for d and z_0 , respectively.

The sonic anemometer at COC was located at a height of 3.25 m. Wind speed profiles for the estimation of d and z_0 were obtained using the sonic anemometer and a nearby ASC sodar measuring wind speeds in 10 m intervals beginning at 30 m continuously up to 200 m. Most of the fetch between 90° and 270° (southerly wind components) consisted of short grasses, airport runway, and a few scattered larger objects. Between 270° and 90° (northerly components) the fetch consisted of a heterogeneous mix of grasses, weeds, and larger shrubs. Most of the neutral cases for the determination of d and z_0 had winds from the south and southeast. d and z_0 were determined to be zero and 5 cm, respectively.

The sonic anemometer at JTN was located atop a tower at a height of 10 m. Wind speed profiles were obtained using the sonic anemometer and a nearby Atmospheric Research Technology sodar measuring wind speeds in 10 m intervals beginning at 30 m continuously up to 200 m. The site was located in a relatively open forest consisting primarily of mesquite trees about 4-7 m in height. The fetch was relatively homogeneous in all directions although the trees were somewhat sparser to the north. Displacement height and z_0 were determined to be 1.5 m and 45 cm, respectively.

Figure 8.1 shows the diurnal averages for friction velocity and sensible heat flux during the winter, spring, summer and autumn seasons as represented by the months of January, April, July, and October. The sensible heat fluxes were calculated from the kinematic heat fluxes assuming an air density of 1.2 kg m^{-3} and 0.95 atmospheres. The friction velocities at BDY and JTN are similar whereas JTN has much larger values reflecting the presences of an open forest canopy. JTN also has significantly larger daytime sensible heat fluxes than the other sites, with the largest difference of roughly 50% occurring in the spring.

Figure 8.2 shows the fractional differences between the wind speeds predicted by the standard logarithmic wind speed profile relationship and the wind speeds observed by the sonic anemometers and sodars. At each site the values of d and z_0 specified above were used in the logarithmic profiles together with the dimensionless ψ functions recommended by Businger *et al.* (1971). Because of potential flow obstructions, the results shown are for wind directions of 225° clockwise to 90° for BDY, 90° clockwise to 270° for COC, and all directions for JTN. The fractional differences are broken down into ranges of Obukhov length (L , m) and wind speed (WS , m s^{-1}). The range of L is given in brackets [] and is either open ended (e.g., $L [-\infty, -20]$ represents all L between negative infinity and -20 m) or a closed range (e.g., $L [-20, 0]$ represents all L between -20 and 0 m).

The most apparent feature of these plots are the huge differences (over predictions) that were associated with the most strongly stable conditions ($0 \text{ m} < L < 20 \text{ m}$) at all 3 sites. This result is not surprising given that the logarithmic profile is derived for the “constant flux” surface layer (Garrett, 1992), which is very shallow in stable conditions. In these cases the winds at turbine level may be effectively decoupled from the surface. For all other ranges of L the fractional errors commonly have

magnitudes of less than 20%. In unstable conditions there is a tendency for the fractional error magnitudes to be a little larger with higher wind speeds ($WS>4$) compared to all wind speeds ($WS>0$) for a given range of L , which is consistent with the wind speed dependent bias discussed in Section 6. All 3 sites exhibited significantly larger negative fractional errors (under predictions) for the most unstable conditions ($-20 < L < 0$) and highest wind speeds ($WS>4$). With the exception of the most stable conditions, the fractional errors in stable conditions were generally less than 10-20% with a slight tendency toward under prediction. These errors tended to decrease as wind speed increased.

An engineering alternative to using logarithmic wind profiles is the use of power laws to describe the variation of wind speed with height. Figure 8.3 shows the best-fit values determined for the exponent in the wind speed power law relationship at 80 m height expressed as a function of both the stability parameter z/L and time of day. The 10-minute average wind speed values measured by the sodar at 80 m height were used with reference to the 10-minute wind speeds measured by the corresponding sonic anemometer in the determinations. The range of values allowed in the iteration to the best fit value was between 0 and 1. While there is a great deal of scatter, the results are similar for the 3 sites. Nighttime values were commonly in the range of 0.2 to 0.6, a little less at BDY, a little more at JTN. Daytime values were generally between 0 and 0.2 with values commonly ranging up to about 0.3 at COC and JTN. As expected, the value of the exponent was a strong function of z/L .

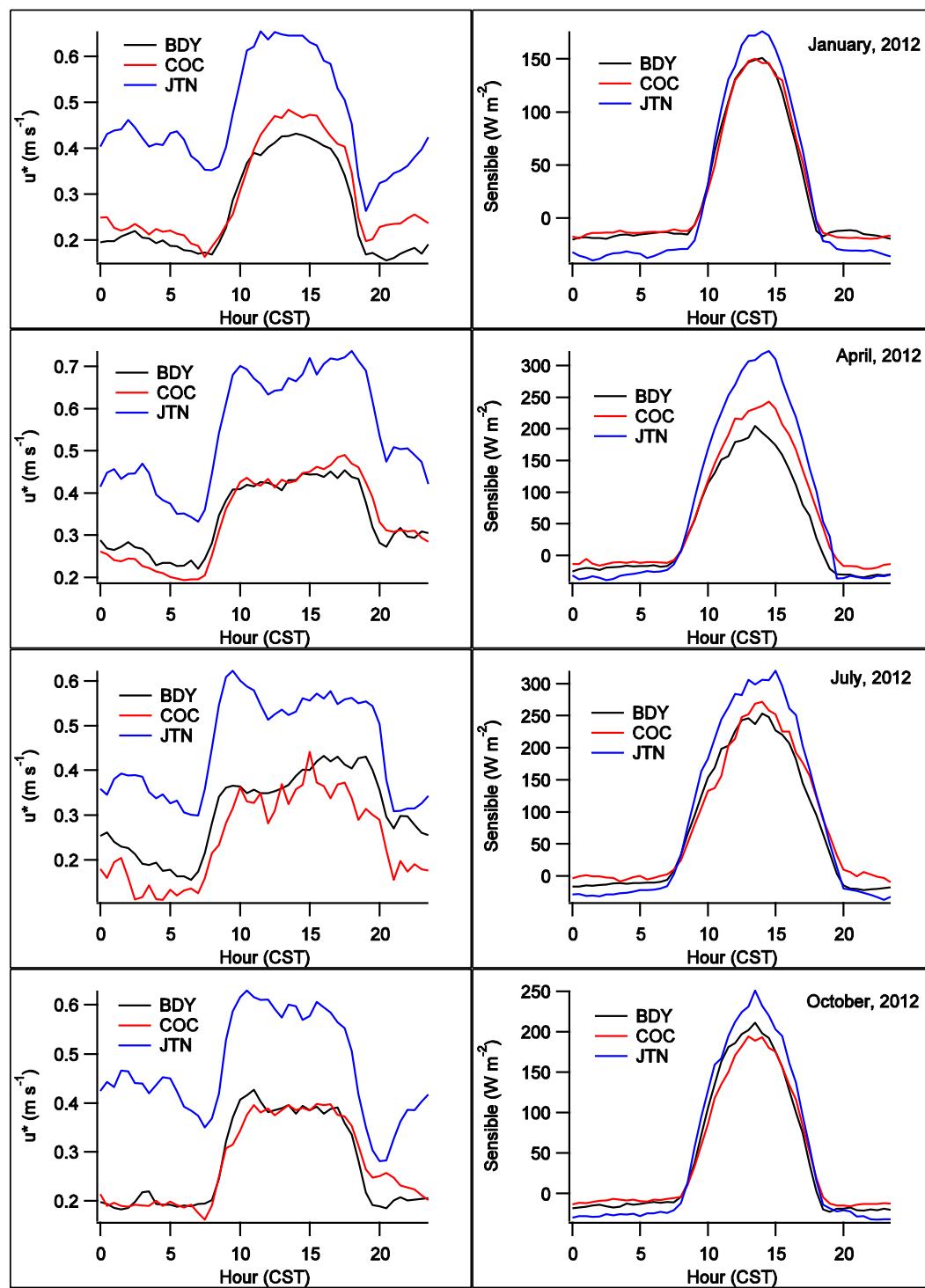


Figure 8.1. Diurnal averages for friction velocity (u^*) and sensible heat flux for winter, spring, summer, and fall, represented by the months of January, April, July, and October for the Brady (BDY), Colorado City (COC), and Jayton (JTN) sites. CST is Central Standard Time.

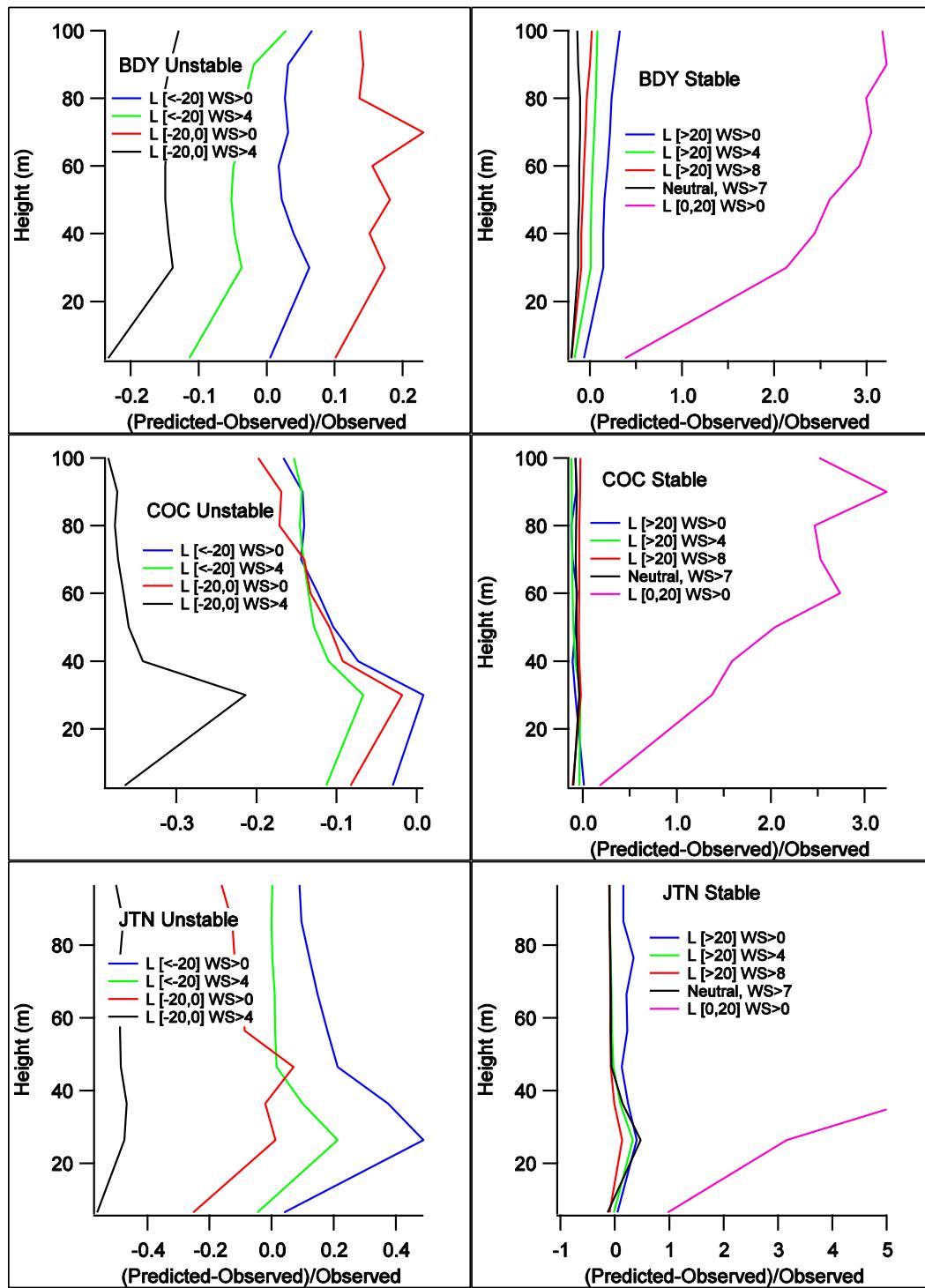


Figure 8.2. Fractional errors comparing winds predicted by the logarithmic wind speed and observed winds at the surface (sonic anemometer) and aloft (sodar) for the 3 sites. The notation used in the legend is explained in the accompanying text.

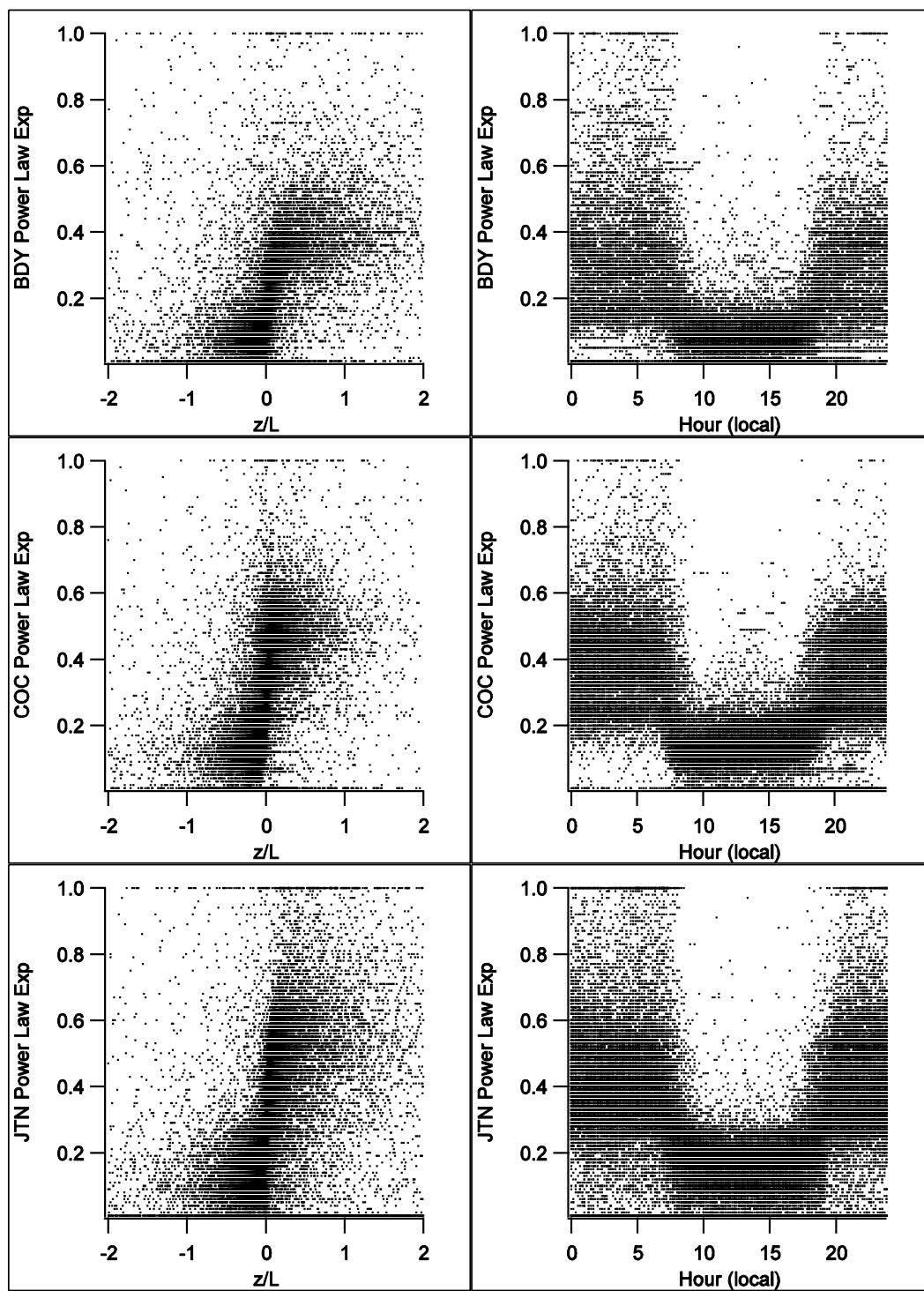


Figure 8.3. Best-fit values for the exponent to the wind speed power law by stability parameter z/L and time of day. Determinations used the 10-minute averages for the sonic anemometer near the surface and sodar at the 80 m height.

9. Summary and Conclusion

WFIP is a DOE sponsored research project whose overarching goals are to improve the accuracy of short-term wind energy forecasts, and to demonstrate the economic value of these improvements. WFIP participants included several DOE national laboratories; two NOAA research laboratories; the NOAA National Weather Service/NWS; and two teams of partners from the private sector and university communities, led by AWS Truepower and WindLogics.

Prior to WFIP, NOAA did not have a focused program to improve its foundational wind forecasts for the wind energy industry. WFIP offered the opportunity for NOAA to jump-start its efforts at improving forecast model skill for this industry, as well as the opportunity to work directly with experts in wind energy. It also offered the WFIP private sector partners (WindLogics inc., and AWS Truepower) the opportunity to advance their own forecasting capabilities either through use of the improved NOAA forecasts or through their own forecasting systems.

WFIP considered two avenues for improving wind energy forecasts. The first was through the assimilation of new meteorological observations into numerical weather forecast models. Additional observations allow for a more precise depiction of the model's initial state of the atmosphere, potentially resulting in more accurate forecasts. The intent of the WFIP instrumentation networks were to provide observations through a deep layer of the atmosphere, and over a sufficiently broad area, to influence NWP forecasts out to 6 hours lead time. New instrumentation was deployed or acquired during concurrent year-long field campaigns in two high wind energy resource areas of the U.S. The first was in the upper Great Plains, including North Dakota, South Dakota, Nebraska, Minnesota and Iowa, where DOE and NOAA partnered with the WindLogics team. The second field campaign was centered in west Texas, where DOE and NOAA partnered with the AWS Truepower team. The WFIP observing systems included 12 wind profiling radars, 12 sodars, and several lidars. In addition, WFIP allowed for NOAA to collect and assimilate for the first time proprietary tall tower (184 locations) and wind turbine nacelle anemometer (411 locations) meteorological observations from the wind energy industry. A key component of WFIP was to develop improved quality control (QC) procedures to ensure that the assimilated observations were as accurate as possible, as a few erroneous observations can easily negate the positive impact of many accurate observations when assimilated into a NWP model. With proper data QC algorithms applied, good agreement was found between the co-located sodar, wind profiling radar, and lidar observed wind speeds.

The second avenue for improving wind energy forecasts was to improve the NWP models directly. Midway through the WFIP field program, NOAA/NWS upgraded its operational hourly-updated NWP forecast model from the Rapid Update Cycle (RUC) model to the Rapid Refresh (RAP) model, and the impacts of this upgrade have been evaluated using WFIP observations. Also, during the course of WFIP NOAA/ESRL made further improvements to the RAP and HRRR models, incorporating more advanced model physics and numerics, new data types assimilated, and better data assimilation procedures, including for the assimilation of wind profiling radar data. The WFIP observations allowed for a

quantification of the improvement of the HRRR and RAP, and allowed the NOAA/NWS to evaluate its North American Mesoscale (NAM) model skill at forecasting hub-height winds for the first time.

Due to the great volumes of data generated by the HRRR, raw model output from it was not publically available prior to WFIP. With WFIP funding NOAA was to obtain the computer infrastructure to make this data available in real-time to both the two private sector teams, as well as to any other party on the wind energy industry.

Pseudo-power forecasts were evaluated by converting tall tower (mostly 60-80m) and model wind speeds to equivalent power using a standard IEC2 power curve. Most of the WFIP forecast skill analysis compares model forecast at point tower locations, appropriate for an individual wind plant that fits within a single model grid cell. For some applications one would instead be interested in comparing spatially averaged power forecasts with spatially averaged model forecasts, for example if a number of dispersed wind plants were feeding power into a transmission line, and the overall power flowing through that transmission line is the quantity of interest. Spatially averaged forecast skill has been investigated for the NOAA RAP model. Also, as part of WFIP, NOAA in collaboration with DOE and private industry partners, developed a ramp metric tool that identifies wind ramp events, matches forecast and observed ramps, and calculates a skill score for the forecasts. Finally, a physical process study was carried out to investigate the relationship of hub-height winds on surface heat and momentum fluxes, and to evaluate the applicability of flux-dependent wind profile laws at replicating the wind profiler through the wind turbine rotor layer.

A summary of the specific scientific results from WFIP follows.

Percent MAE improvements between the NWS RUC operational hourly-updated forecast model and the real-time NOAA/ESRL RAP hourly-updated forecast model, calculated over the first 6 months of the WFIP field campaign, were significant. In the Northern Study Area (NSA) a 13% power improvement at forecast hour 01 was found, decreasing to a minimum improvement of 6-7% for forecast hours 7-15. In the Southern Study Area (SSA) a 15% power improvement at forecast hour 01 was observed, decreasing to a minimum improvement of 5% at forecast hour 15. This improvement reflects the combined effects of the better RAP model versus the RUC model, as well as the contribution from assimilation of the WFIP observations into the research RAP model.

To quantify the impact of assimilation of the additional WFIP observations only, data denial (DD) experiments were run with the RAP and NAM models. In these experiments a set of control simulations was run that did not assimilate any of the special WFIP observations, which was then compared to an experimental simulation that did assimilate the WFIP observations. Six DD episodes were run, each from 7-12 days long, spanning all four seasons of the year. Using conventional statistical analysis with the tall tower data sets for verification, the experimental simulations were found to improve the average MAE power forecast skill at the 95% confidence level for the first seven forecast hours in the NSA, and through forecast hour 03 in the SSA. This improvement ranged from 8% at forecast hour 1 to

3% at forecast hour 6 in the NSA, and from 6% at forecast hour 1 to 1% at forecast hour 6 in the SSA. Positive forecast skill improvement remained until the last forecast hour 15 in both study areas, but at levels less than 2%. Although the NAM DD simulations were only run for two episodes (December and January) the results are fully consistent with the findings from the RAP model over the larger data set. The forecast skill improvement due to assimilation of the new WFIP observations was also found to be dependent on the location of the verifying site. Verifying tower sites that were on the periphery of the NSA and SSA domains had smaller improvements than those located within the core observing network area, demonstrating the increased benefit of having more observations spread over a larger geographic area.

The dependence of the forecast percent improvement on season, verification time of day, and observed power was investigated. Although the magnitude of improvement varied considerably between the 6 DD episodes, no clear seasonal trends across both study areas was evident. This suggests that the variability was more related to sampling issues than to meteorological characteristics of the different seasons. In contrast, the forecast improvement was found to depend strongly on the hour of the day that the forecast was verified at. In the NSA the largest improvements were observed during the daytime hours, with considerably smaller improvements during the nighttime hours. In the SSA the diurnal variation of the improvement was less clear, with a suggestion of two maxima, one also in the early daytime hours, with the second in the night. The power MAE itself was also found to have a strong diurnal signature. In both study areas the lowest MAE was associated with forecasts that were initialized and verified during the daytime hours. MAE during the nighttime hours was significantly greater (up to a factor 2), reflecting the fact that the stable boundary layer and evolution of the nocturnal low level jet is poorly understood and modeled, and is an area in need of further study and investigation. In terms of dependence on the observed power, the power forecast improvement had at most a small variation, with slightly larger improvement for larger observed power.

The dependence of the forecast improvement on the size of the forecast error was also investigated. For positive forecast errors (when the model forecasts more power than later materializes) no obvious dependence on forecast error is found. For negative forecast errors (when the model under-forecasts the power), the improvement is greatest for smaller forecast errors, decreases with increasing size of the error, and becomes negative for the most negative errors. The reasons for the negative impact of the assimilated data on the largest power under-forecasts are not understood, and require further investigation, including analysis of the types of meteorological phenomena associated with these events.

The degree of spatial averaging of the forecasts and observations before they are compared is found to have a profound impact on the skill of the forecast, with the power MAE decreasing by more than a factor of 2 as the spatial averaging goes to the maximum possible. This demonstrates the advantage to ISO's of having spatially distributed generation, not only because it provides less variability in generation, but also because the generation that is produced can be better forecast. Surprisingly, the

impact of assimilation of the new WFIP observations measured as a percent improvement stays constant or even increases with the degree of spatial averaging, up to domains on the order of 400 km x 600km.

The skill of the RAP model at forecasting ramp events was studied with the ramp tool developed for WFIP, using data from 6 DD episodes for which 15 min model output was available. The model was found to have greater forecast skill for longer duration ramps, but the skill was only marginally dependent on the magnitude of the ramps. The lack of skill at forecasting short duration ramps is likely due to the fact that these events also span a small spatial scale, making it difficult for the 3dvar data assimilation scheme to represent them in the model's initialization. Research into more advanced data assimilation techniques may provide greater skill for these small scale/short duration ramp events.

Assimilation of the special WFIP observations was found to improve the ramp forecast skill, averaged over the first 9 forecast hours, by more than 10% in the NSA, but only 3.5% in the SSA. The difference between the two study areas is consistent with that found for conventional MAE statistics. Reasons for the greater impact of the special WFIP observations in the NSA than in the SSA are, first, the NSA had more tall tower observations, more wind profiler observations, and the addition of nacelle anemometer observations; the greater numbers of observations is likely to have contributed to the greater improvement in both conventional MAE and ramp forecast skill. Second, the new observations were spread over a wider geographic area in the NSA than in the SSA, allowing for the model initial field improvements to be more robust and affect a wider area, thereby having a more lasting positive impact before advecting out of the study area. Third, a larger number of synoptic scale systems in the NSA may also have contributed to the larger impact of the new observations in the NSA relative to the SSA.

The degree of ramp forecast skill improvement also varied considerably between DD episodes, especially in the SSA. Most of the improvement was found to come from correctly forecasting up ramp events and down ramp events with greater accuracy, as opposed to decreasing the penalty from forecasting a ramp when one of the opposite sign occurs. The lack of improvement for these opposite sign forecasts may be because these events are of short duration and small spatial scale, which makes them difficult to assimilate into the models, so that the models have little skill in forecasting them.

In the final component of the analysis, estimating hub height wind speeds using stability dependent flux-profile relationships was found to be problematic in stable conditions, when hub-height winds can decouple from surface forcing. It seems likely that there is a fundamental limit to the accurate extrapolation of near surface speeds to hub-height in stable conditions without other supporting information on the depth of the stable boundary layer and height and strength of the low-level jet.

In summary, it is clear that significant improvements to hub-height wind forecasts have been achieved during WFIP from both improved weather forecast models, and by assimilating additional observations into those models. The development of the ability to assimilate nacelle and tall tower observations from the wind energy community is a significant benefit from the WFIP project, given that more wind

plant operators are routinely making their data available to NOAA. The WFIP project should help accelerate the ingest of these observation types into the operational RAP and NAM models.

Looking forward, key areas of research that WFIP has identified are first, to improve the accuracy of meteorological observations from existing instruments, and to develop inexpensive sensors that can provide the required measurements and be deployed in wide networks. Future research also will be necessary to evaluate the impact of new observations in complex terrain or coastal areas: will they have a greater or lesser effect than in the Great Plains? Also, research is needed to determine which type of instrument has the largest impact, and what deployment density of observation is optimal. If the new observations span ever larger geographic domains, how large of an improvement can they eventually contribute and over what length of forecast? In terms of model improvements, the stable nocturnal boundary layer remains a forecast weakness. Low-level jets contribute to the high wind resource of the Great Plains, and the inability to forecast these well contributes to the larger model errors found at night. Improving forecasts within the nocturnal boundary layer will require new physical parameterization schemes of fundamental processes such as turbulent mixing, as well as better model initial conditions. Better model initial conditions will require not only better observations, but perhaps equally important, better methods to assimilate observations in the stable boundary layer, so that the full benefit of new observations can be achieved.

10. References

- Alpert, J., 2004: Sub-grid scale mountain blocking at NCEP, *Proc. 20th Conference on Weather on Analysis and Forecasting/17th Conference on Numerical Weather Prediction*, American Meteorological Society. Seattle, WA. P2.4. [Available online at <https://ams.confex.com/ams/pdfpapers/71011.pdf>].
- Benjamin, S.G., B.E. Schwartz, E.J. Szoke, and S.E. Koch, 2004a: The value of wind profiler data in U.S. weather forecasting. *Bull. Amer. Meteor. Soc.*, **85**, 1871-1886.
- Benjamin, S. G., and Coauthors, 2004b: An Hourly Assimilation–Forecast Cycle: The RUC, *Mon. Wea. Rev.*, **132**, 495–518.
- Benjamin, S. G., S. S. Weygandt, D. Devenyi, J. M. Brown, G. A. Manikin, T. L. Smith, and T. G. Smirnova, 2004c: Improved moisture and PBL initialization in the RUC using METAR data, Extended Abstracts, *22nd Conference on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc. 17.3.
- Benjamin, S.G., B.D. Jamison, W.R. Moninger, S. R. Sahm, B. Schwartz, T.W. Schlatter, 2010: Relative short-range forecast impact from aircraft, profiler, radiosonde, VAD, GPS-PW, METAR, and mesonet observations via the RUC hourly assimilation cycle. *Mon. Wea. Rev.*, **138**, 1319-1343
- Bernier, N. B., S. Belair, 2012: High Horizontal and Vertical Resolution Limited-Area Model: Near-Surface and Wind Energy Forecast Applications, *J. Appl. Meteor. Climatol.*, **51**, 1061–1078.
- Bianco, L., D. Gottas, and J. M. Wilczak, 2013: Implementation of a Gabor transform data quality control algorithm for UHF wind profiling radars, *J. Atmos. Ocean. Tech.*, **30**, 2697-2703, doi: 10.1175/JTECH-D-13-00089.1.
- Bossaavy, A., R. Girard, and G. Kariniotakis, 2013: A novel method for comparison of different wind power ramp characterization approaches. European Wind Energy Association annual meeting, Vienna Austria.
- Bristol, E., 1990: Swinging door trending: adaptive trend recording. International Standards for Automation.
- Burk, S. D., and W. T. Thompson, 1989: A vertically nested regional numerical prediction model with second-order closure physics, *Mon. Wea. Rev.*, **117**, 2305–2324.
- Businger, J.A., J.C. Wyngaard, C. Izumi, and E.F. Bradley (1971). Flux profile relationships in the atmospheric surface layer, *J. Atmos. Sci.*, **28**, 181-189.

Chou, Ming-Dah, and Max J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models, *NASA Tech. Memo*, 84 pp.

Cutler N.J., Kay M., Jacka K., Nielsen T.S., 2007: Detecting, categorizing and forecasting large ramps in wind farm power output using meteorological observations and WPPT. *Wind Energy*, 10(5):453–470.

Daley, R., 1991: *Atmospheric Data Analysis*, Cambridge University Press, 457 pp.

Derber, J., and A. Rosati, 1989: A global oceanic data assimilation system, *J. Phys. Oceanogr.*, **19**, 1333–1347.

Desroziers, G. and S. Ivanov, 2001: Diagnosis and adaptive tuning of observation error parameters in a variational assimilation, *Quart. J. Roy. Meteor. Soc.*, **127**, 1433–1452.

Devenyi, D., and S. G. Benjamin, 2003: A variational assimilation technique in a hybrid isentropic-sigma coordinate, *Meteor. Atmos. Phys.*, **82**, 245–257.

Drechsel, S., G. J. Mayr, J. W. Messner, R. Stauffer, 2012: Wind Speeds at Heights Crucial for Wind Energy: Measurements and Verification of Forecasts, *J. Appl. Meteor. Climatol.*, **51**, 1602–1617.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, *J. Geophys. Res.*, **108** (D22), 16, doi:10.1029/2002JD003296.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model. Preprints, *15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 280–283.

Ferrier, B. S., W. Wang, and E. Colon, 2011: Evaluating cloud microphysics schemes in nested NMMB forecasts. *24th Conf. on Weather Analysis and Forecasting/20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc.

Florita, A., B.-M. Hodge, and K. Orwig, 2013: Identifying wind and solar ramps. 2013 IEEE Green Technologies Conference, DOI 10.1109/GreenTech.2013.30.

Garratt, J.R., 1992: *The Atmospheric Boundary Layer*, Cambridge University Press, 316 pp.

- Gandin, L. S., 1988: Complex Quality Control of Meteorological Observations, *Mon. Wea. Rev.*, **116**, 1137–1156.
- Greaves B., Collins J., Parkes J., Tindal A, 2009: Temporal forecast uncertainty for ramp events. *Wind Engineering*, 33(11):309–319.
- Grell, G. and D. Devenyi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques, *Geophys. Res. Lett.*, **29**, 1693, doi:10.1029/2002GL015311.
- Griesser, T., and H. Richner, 1998: Multiple peak processing algorithm for identification of atmospheric signal in Doppler radar wind profiler spectra, *Meteor. Z.*, **7**, 292–302.
- Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts, *Wea. Forecasting*, **14**, 155–167.
- Houtekamer, P. L., L. Lefavre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction, *Mon. Wea. Rev.*, **124**, 1225–1242.
- Holm, E., E. Andersson, A. Beljaars, P. Lopez, J.-F. Mahfouf, A. Simmons, and J.-N. Thepaut, 2002: Assimilation and modelling of the hydrologic cycle: ECMWF's status and plans, *ECMWF Tech. Memo.*, 55 pp. [383].
- Huang, X.-Y. and P. Lynch, 1993: Diabatic Digital-Filtering Initialization: Application to the HIRLAM Model, *Mon. Wea. Rev.*, **121**, 589–603.
- Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models, *J. Geophys. Res.*, **113**, D13103, doi:10.1029/2008JD009944.
- Janjic, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes, *Mon. Wea. Rev.*, **122**, 927–945.
- Janjic, Z. I., 2001: Nonsingular implementation of the Mellor-Yamada level 2.5 scheme in the NCEP meso model, *NCEP Office Note*, 91 pp. [437].
- Janjic, Z. I., 2003: A nonhydrostatic model based on a new approach, *Meteorology and Atmospheric Physics*, **82**, 271–285.

- Janjic, Z. I., 2005: A unified model approach from meso to global scales, *Geophysical Research Abstracts*, General Assembly, Vienna, Austria, European Geosciences Union, **7**, 05 582, SRef-ID: 1607–7962/gra/EGU05–A–05 582.
- Janjic, Z. I. and T. L. Black, 2007: An ESMF unified model for a broad range of spatial and temporal scales, *Geophysical Research Abstracts*, General Assembly, Vienna, Austria, European Geosciences Union, **9**, 05 025, 2007, SRef-ID: 1607–7962/gra/EGU2007–A–05 025.
- Janjic, Z. I. and R. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part I Dynamics, *NCAR Tech. Note*, NCAR/TN- 489 +STR. 74 pp.
- Kain, J. S., et al., 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP, *Wea. Forecasting*, **23**, 931–952.
- Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation, and Predictability*, Cambridge University Press, 341 pp.
- Kristensen, L., 1998: Cup anemometer behavior in turbulent environments, *J. Atmos. Ocean. Technol.*, **15**, 5–17.
- Kristensen, L., 1999: The perennial cup anemometer, *Wind Energy*, **2**, 59–75. doi: 10.1002/(SICI)1099-1824(199901/03)2:1<59::AID-WE18>3.0.CO;2-R
- Lehmann, V., 2012: Optimal Gabor-Frame-Expansion-Based Intermittent-Clutter-Filtering method for radar wind profiler, *J. Atmos. Ocean. Technol.*, **29**, 141–158.
- Lorenc, A. C., 1986: Analysis methods for numerical weather prediction, *Quart. J. Roy. Meteor. Soc.*, **112**, 1177–1194.
- Lorenz, E. N., 1963: Deterministic nonperiodic flow, *J. Atmos. Sci.*, **20**, 130–141.
- Lynch, P., D. Giard, and V. Ivanovici, 1997: Improving the Efficiency of a Digital Filtering Scheme for Diabatic Initialization, *Mon. Wea. Rev.*, **125**, 1976–1982.
- Merritt, D. A., 1995: A statistical averaging method for wind profiler Doppler spectra, *J. Atmos. Oceanic Technol.*, **12**, 985–995.
- Mittermaier, M. P., 2013: A strategy for verifying near-convection-resolving model forecasts at observing sites, *Wea. Forecasting*, In press.

- Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave, *J. Geophys. Res.*, **102**, 16 663–16 682.
- Pan, Z.-T., S. G. Benjamin, J. M. Brown, and T. G. Smirnova, 1994: Comparative experiments with MAPS on different parameterization schemes for surface moisture flux and boundary-layer processes, *Mon. Wea. Rev.*, **122**, 449–470.
- Parrish, D. F. and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system, *Mon. Wea. Rev.*, **120**, 1747–1763.
- Parrish, D. F., J. C. Derber, R. J. Purser, W.-S. Wu, and Z.-X. Pu, 1997: The NCEP global analysis system: Recent improvements and future plans, *J. Meteor. Soc. Japan*, **75**, 359–365.
- Purser, R. J., W.-S. Wu, D. F. Parrish, and N. M. Roberts, 2003: Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances, *Mon. Wea. Rev.*, **131**, 1524–1535.
- Reisner, J., R. M. Rasmussen, and R. T. Bruintjes, 1998: Explicit forecasting of supercooled liquid water in winter storms using the MM5 mesoscale model, *Quart. J. Roy. Meteor. Soc.*, **124**, 1071–1107.
- Rife, D. L., C. A. Davis, Y. Liu, T. T. Warner, 2004: Predictability of Low-Level Winds by Mesoscale Meteorological Models, *Mon. Wea. Rev.*, **132**, 2553–2569.
- Rogers, E., and Coauthors, 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans. Preprints, *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 2A.4. [Available online at <http://ams.confex.com/ams/pdfpapers/154114.pdf>].
- Schwartz, C. S., et al., 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing, *Mon. Wea. Rev.*, **137**, 3351–3372.
- Skamarock, W. C. and Coauthors, 2008. A description of the Advanced Research WRF version 3, *NCAR Tech. Note NCAR/TN-475STR*, 125 pp.
- Smirnova, T. G., J. M. Brown, and S. G. Benjamin, 1997: Evolution of soil moisture and temperature in the MAPS/RUC assimilation cycle, *Preprints, 13th Conf. on Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 172–175.

Smirnova, T. G., S. G. Benjamin, J. M. Brown, B. Schwartz, and D. Kim, 2000: Validation of long-term precipitation and evolved soil moisture and temperature fields in MAPS, *Preprints, 15th Conf. on Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 43–46.

Talagrand, O., 1997: Assimilation of observations, an introduction, *J. Meteor. Soc. Japan*, **75**, 191-209.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk micro-physics scheme. Part I: Description and sensitivity analysis, *Mon. Wea. Rev.*, **132**, 519–542.

Thompson, G., P. R. Field, R. M. Rasmussen, W. D. Hall, 2008: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization, *Mon. Wea. Rev.*, **136**, 5095–5115.

Weber, B. L., D. B. Wuertz, D. C. Welsh, R. McPeek, 1993: Quality Controls for Profiler Measurements of Winds and RASS Temperatures, *J. Atmos. Ocean. Tech.*, **10**, 452–464. doi: [http://dx.doi.org/10.1175/1520-0426\(1993\)010<0452:QCFPMO>2.0.CO;2](http://dx.doi.org/10.1175/1520-0426(1993)010<0452:QCFPMO>2.0.CO;2)

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems, *Mon. Wea. Rev.*, **125**, 527–548.

Wilczak, J. M., and Coauthors, 1995: Contamination of wind profiler data by migrating birds: Characteristics of corrupted data and potential solutions, *J. Atmos. Ocean. Tech.*, **12**, 449–467.

Wilczak, J. M., E. E. Gossard, W. D. Neff, and W. L. Eberhard, 1996: Ground-based remote sensing of the atmospheric boundary layer: 25 years of progress, *Boundary-Layer Meteorol.*, **78**, 321-349.

Wu, W.-S., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances, *Mon. Wea. Rev.*, **130**, 2905–2916.

Appendices

Appendix 1. Automated Quality Control Algorithms

Ground clutter interference is caused by physical elements such as power lines or trees that shake in the wind and are measured most often by the side-lobes of the radar antennae, giving small but non-zero speeds that will create a low bias in the profiler wind speeds. Ground clutter is generally worse in the lowest gates, its effect decreasing with height. Ground clutter was identified based on the assumption that near the surface the true wind speed should increase with height. Using the 10m wind speed measured by a prop-vane at each profiler site as a reference, periods when both $U(z) < (A * U_{10})$ and $U_{10} > 2.0 \text{ ms}^{-1}$ were identified as likely contamination and eliminated, where $U(z)$ is the radar speed at height z , and $A = 1.3$. The lower wind speed threshold was applied because in near-calm conditions the true wind profile may not increase with height by much, if at all, and clutter is less likely to occur. This algorithm was implemented at all of the WPR radar sites. One site, VLC, had extremely bad clutter, and for it a lower wind speed threshold of 1.5 ms^{-1} was used, and a value of $A= 1.2$.

Radio frequency interference is typically caused by cell-phone transmission that occurs near the 915 MHz wind profiler frequency. Like ground clutter, RFI typically causes low wind speeds, however contaminated speeds generally start above the lowest few gates and can affect any level. This algorithm compares speeds between two levels, starting with the lowest two gates. If the speed in the upper of the two gates was less than 5 ms^{-1} , and the speed decreased with height by more than 1 ms^{-1} between the two gates, the upper gate was identified as being contaminated by RFI and eliminated. The procedure was then repeated always comparing the next highest gate with the last gate that was deemed to have good data. In relatively rare cases where RFI affected the first gate, those bad data were eliminated by the ground clutter algorithm.

RFI can also contaminate the RASS temperature measurements. The characteristics of contaminated RASS temperatures are that they have a nearly constant value at multiple heights within a single measured profile. An algorithm was developed that searched for contaminated values by binning the gated RASS temperatures within a profile, where each bin had a narrow range of 0.2 C . If more than 4 values occurred in any single bin, those 4 values were eliminated from the profile.

Appendix 2. Instrument Inter-comparisons

Scatter plot inter-comparisons are included in this Appendix for each of the 6 data denial episodes, using the fully QC'd data that was assimilated into the models.

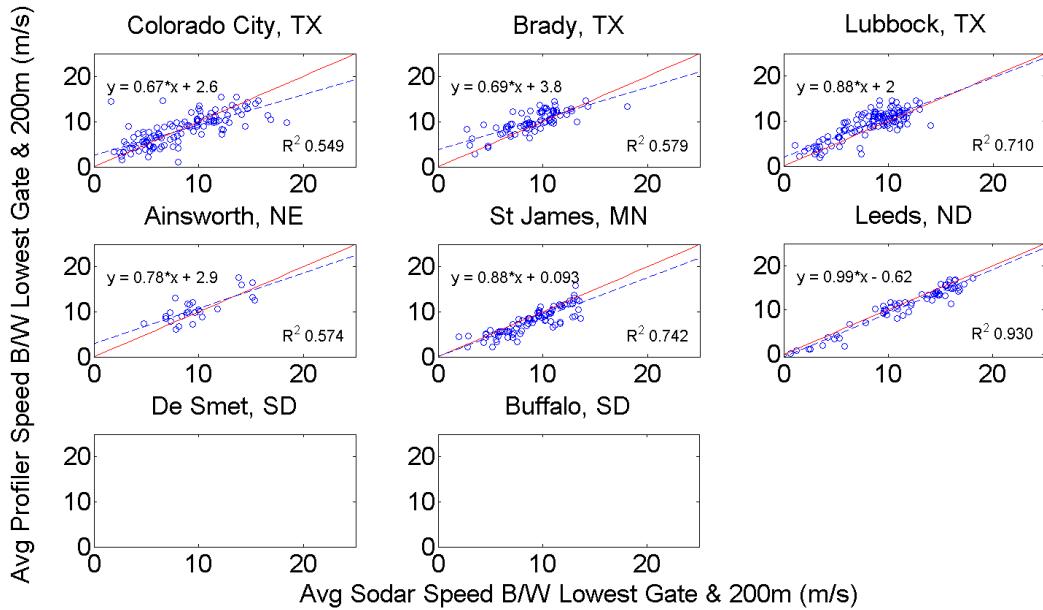


Figure A1. Wind profiling radar and sodar inter-comparisons for the Nov 30 – Dec 06, 2011 Data Denial episode.

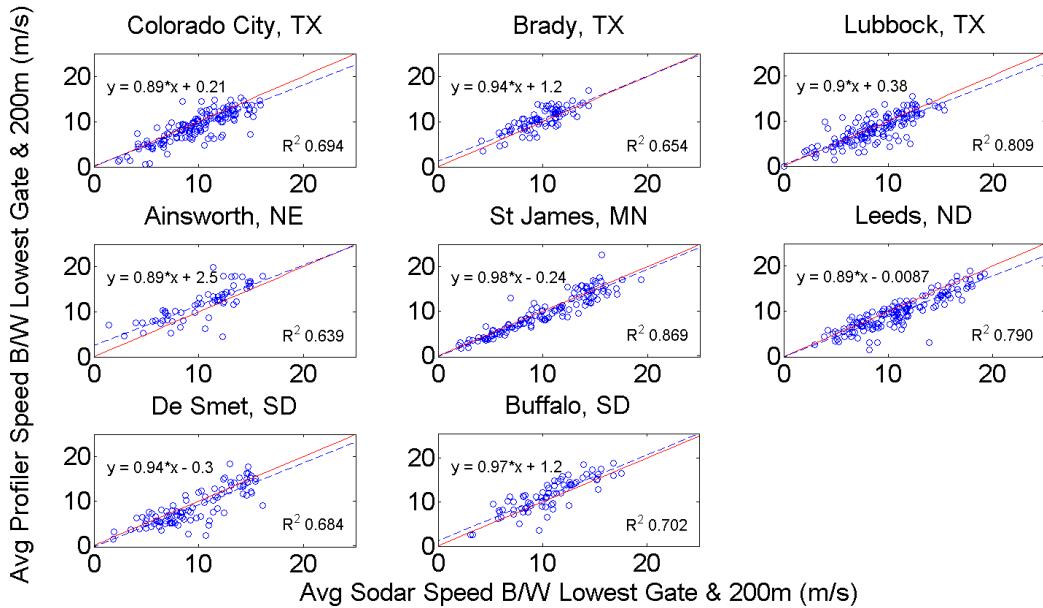


Figure A2. Wind profiling radar and sodar inter-comparisons for the Jan 07 – Jan 15, 2012 Data Denial episode.

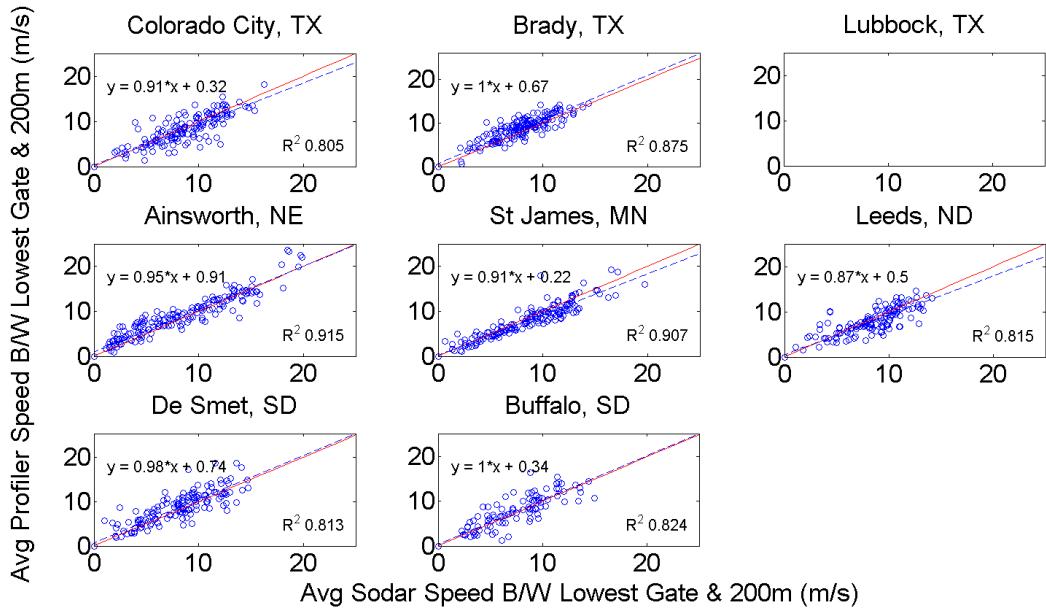


Figure A3. Wind profiling radar and sodar inter-comparisons for the April 14 – April 25, 2012 Data Denial episode.

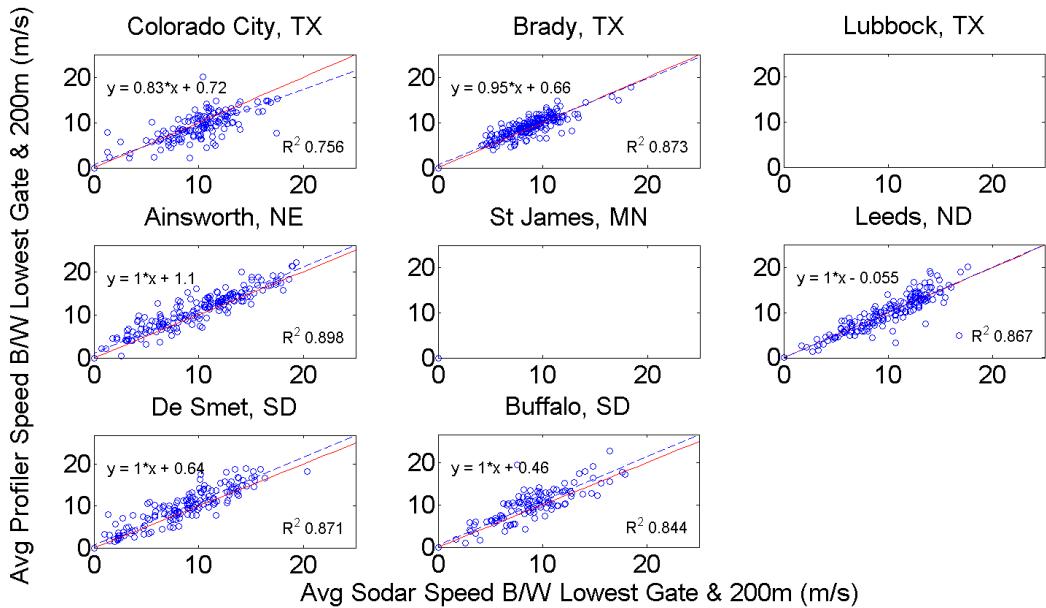


Figure A4. Wind profiling radar and sodar inter-comparisons for the Jun 09 – Jun 18, 2012 Data Denial episode.

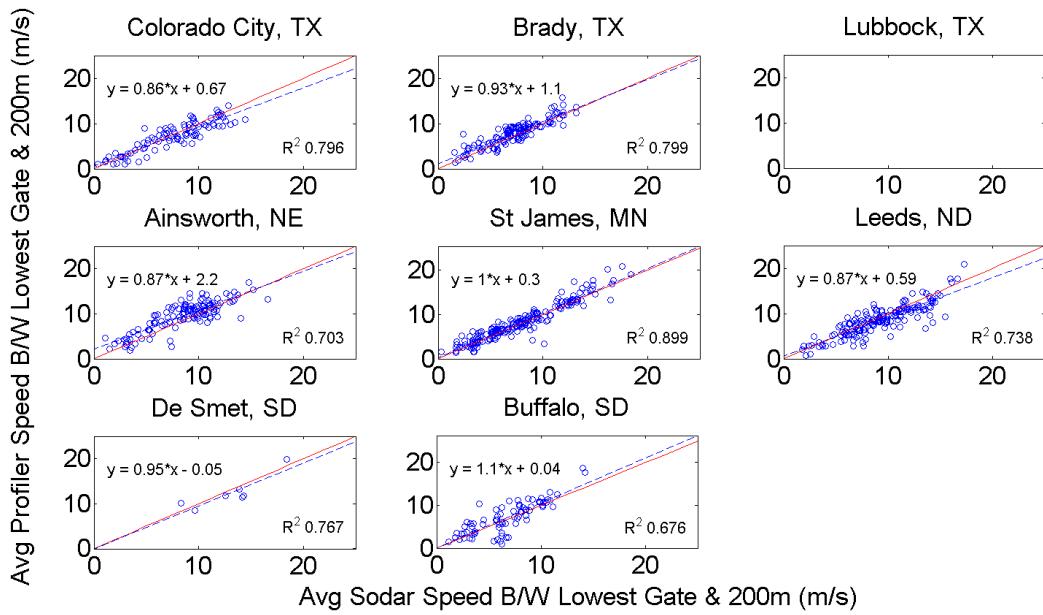


Figure A5. Wind profiling radar and sodar inter-comparisons for the Sept 16 – Sept 25, 2012 Data Denial episode.

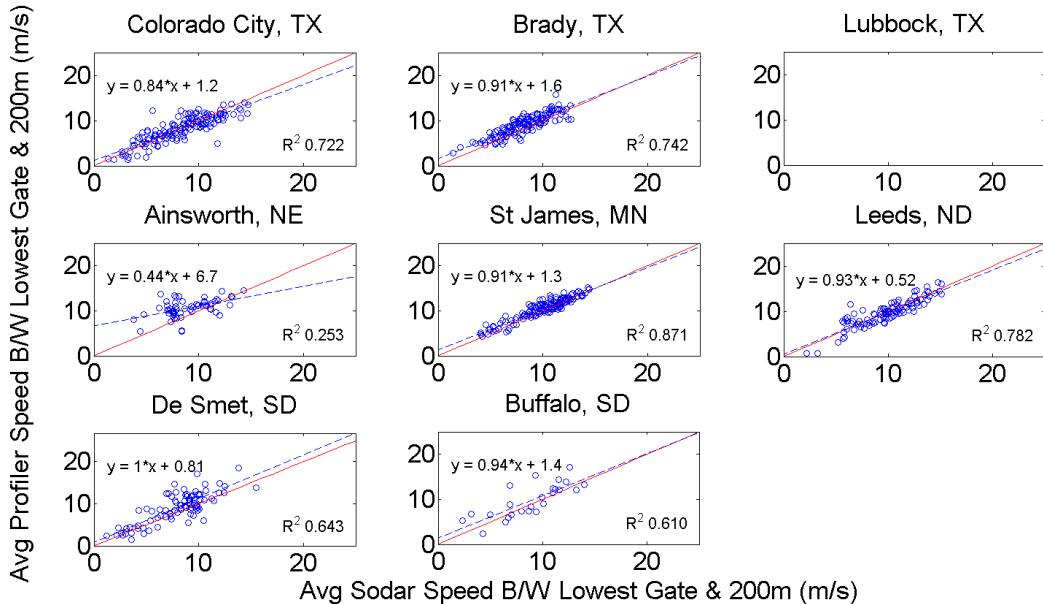


Figure A6. Wind profiling radar and sodar inter-comparisons for the Oct. 13-20, 2012 Data Denial episode.

List of Figures

<i>Figure 1.1. Geographic domain of the Northern Study Area.....</i>	9
<i>Figure 1.2. Geographic domain of the Southern Study Area.....</i>	10
<i>Figure 2.1. The WFIP 915 MHz wind profiling radar at Saint James, MN.....</i>	13
<i>Figure 2.2. 24 hour time-height cross-section of hourly averaged winds from the 915 MHz Brady Texas Wind Profiling Radar.....</i>	14
<i>Figure 2.3. The WFIP 449 MHz wind profiling radar located at Buffalo, ND.....</i>	15
<i>Figure 2.4. 24 hour time-height cross-section of hourly averaged winds from the 449 MHz Buffalo ND Wind Profiling Radar.....</i>	16
<i>Figure 2.5. 24 hour of hourly-sampled RASS virtual temperature.....</i>	17
<i>Figure 2.6. 24 hour wind time-height cross-section from the Reagan TX sodar.....</i>	17
<i>Figure 2.7. 24 hour wind time-height cross-section from the DOE/PNNL lidar deployed at DeSmet SD....</i>	18
<i>Figure 2.8. 24 hour wind time-height cross-section from the Leosphere WindCube8 lidar.....</i>	19
<i>Figure 2.9. 915-MHz wind profiler time-height cross sections of hourly winds computed by a standard consensus procedure and the combined Gabor, thresholding, and Weber-Wuertz pattern recognition scheme.....</i>	23
<i>Figure 2.10. Hourly averaged winds from the WFIP Watford City, ND 915 MHz wind profiler. Top panel: winds produced by a standard consensus algorithm. Bottom panel: winds produced using the Weber-Wuertz pattern recognition processing algorithm.....</i>	25
<i>Figure 2.11. 10-day time series of wind directions from the RAP model (red curves) and observations (black curves).....</i>	27
<i>Figure 2.12. Idealized wind direction histograms for a tower with a large mean bias (MB) error and narrow distribution (blue curve), and a smaller mean direction bias with a broader distribution (red curve).....</i>	28
<i>Figure 2.13. The dimensionless direction error factor DF for each of the towers/levels.....</i>	29
<i>Figure 2.14. Histogram of the mean direction bias of all of the towers/levels.....</i>	30
<i>Figure 2.15 Inter-comparisons of WPR, sodar, and lidar data from the June 9-18, 2012 data denial time period.....</i>	32
<i>Figure 2.16 Scatter plots of the real-time, hourly averaged WPR and sodar data for six sites that had co-located systems, for the period Oct. 13-20, 2011.....</i>	33
<i>Figure 2.17 Scatter plots of the hourly averaged WPR and sodar data for six sites that had co-located systems after additional QC was applied, for the period Oct. 13-20, 2011.....</i>	34

<i>Figure 3.1. Model domains for the 13 km Rapid Refresh (blue), 13 km RUC (red) and the 3 km HRRR (green).....</i>	35
<i>Figure 3.2. NAM Parent 12 km (black) and 4 km CONUSnest (red) computational domains used during WFIP.....</i>	38
<i>Figure 4.1. NAM/NDAS data assimilation cycling diagram.....</i>	47
<i>Figure 5.1 A screen-shot of the main page of the model/observation evaluation web page.....</i>	49
<i>Figure 5.2. An IEC class 2 wind turbine power curve, which shows the expected wind power produced as a function of wind speed for this class of turbine.....</i>	50
<i>Figure 5.3. MAE percent improvement of the ESRL/RAP model over the NCEP/RUC model for the vector wind as a function of forecast length in the Northern Study Area.....</i>	51
<i>Figure 5.4. The same as for Fig. 5.3, except for coefficient of determination R^2 and MAE percent improvement of wind power.....</i>	52
<i>Figure 5.5. MAE percent improvement of the ESRL/RAP model over the NCEP/RUC model for the vector wind as a function of forecast length in the Southern Study Area.....</i>	53
<i>Figure 5.6. The same as for Fig. 5.5, except for coefficient of determination R^2 and MAE percent improvement of wind power.....</i>	54
<i>Figure 5.7. RMSE improvement of the vector wind for the HRRR model over the ESRL/RAP model, for the NSA.....</i>	55
<i>Figure 6.1. RAP control simulation wind profiler radar biases (model-observation) as a function of height, for each of the 6 DD episodes, averaged for all 15 forecast hours.....</i>	58
<i>Figure 6.2. RAP control bias for as a function of length of forecast, averaged for all 6 DD episodes and over the layer 0-500m AGL, for the NSA (orange) and SSA (green).....</i>	58
<i>Figure 6.3. RAP control bias using the wind profiler speed observations, as a function of forecast verification time (UTC), for all 6 DD experiments.....</i>	59
<i>Figure 6.4 RAP control bias at forecast hour 00 using the wind profiling radar observations, as a function of observed wind speed, averaged over the lowest 500m AGL and over all 6 DD episodes.....</i>	60
<i>Figure 6.5. RAP control simulation sodar biases as a function of height, for each of the 6 DD episodes, averaged for all 15 forecast hours.....</i>	61
<i>Figure 6.6. RAP control sodar bias for as a function of length of forecast, averaged for all 6 DD episodes and over the layer 0-200m AGL, for the NSA (orange) and SSA (green).....</i>	61
<i>Figure 6.7. RAP control bias using the sodar speed observations, as a function of forecast verification time (UTC), for all 6 DD experiments, averaged over the layer 0-200m AGL.....</i>	62
<i>Figure 6.8 RAP control wind speed bias at forecast hour 00 using the sodar wind speed observations, as a function of the observed wind speed, averaged over all 6 DD episodes.....</i>	63

<i>Figure 6.9. RAP bias using the tall tower observations, as a function of length of forecast, averaged for all 6 DD episodes, for the NSA (orange) and SSA (green).....</i>	63
<i>Figure 6.10. As in Fig. 6.9, except showing the biases for individual DD episodes.....</i>	64
<i>Figure 6.11. RAP control speed bias using the tall tower observations as a function of forecast verification time (UTC, x-axis), and forecast length (y-axis), averaged for all 6 DD experiments.....</i>	65
<i>Figure 6.12 RAP control bias using the tall tower observations as a function of observed wind speed, averaged over all 6 DD episodes, for the NSA (top) and SSA (bottom), for three different forecast lengths.....</i>	66
<i>Fig. 6.13. Vertical profiles of vector wind RMSE averaged over all 12 WPR sites and all 55 DD RAP simulation days, at the model initialization time and three hour forecast length increments.....</i>	67
<i>Fig. 6.14. As in Fig. 6.13 for the 9 NSA sites (top set of panels) and 3 SSA sites (bottom set of panels)....</i>	69
<i>Fig 6.15. Wind profiling radar layer averaged (0-2000m AGL) RMSE, averaged for the 6 control DD episode simulations (red) and the experimental DD simulations (blue).....</i>	70
<i>Figure 6.16. The same as Fig. 6.15 except using sodar observations, with the layer average between 0-200m.....</i>	72
<i>Figure 6.17. MAE and R^2 percent improvement of the experimental RAP simulation assimilating the special WFIP observations over a control that does assimilate the WFIP observations, for no bias correction, and the mean, diurnal, and speed dependent bias corrections.....</i>	74
<i>Figure 6.18. RAP tall tower-derived RMSE, averaged for the 6 control DD episode simulations (red) and the experimental DD simulations (blue). Top 4 panels are for vector wind, and bottom 4 panels are for power. Left panels are for the 9 NSA and right panels are for the SSA. Panels with black curves show difference between control and experimental simulation RMSE's in the corresponding panel above, and error bars indicate 95% confidence intervals.....</i>	76
<i>Figure 6.19. MAE percent improvement for the vector wind, for the NSA (orange curve) and SSA (green curve) for all 55 DD episode days.....</i>	77
<i>Figure 6.20. The same as Fig. 6.19 except for power.....</i>	77
<i>Figure 6.21 MAE precent improvement in the vector wind broken out by DD episode, for all 55 DD episode days, for the NSA (top panel) and SSA (bottom panel).....</i>	78
<i>Figure 6.22. The same as Fig. 6.21 except for power, and for both MAE and R^2.....</i>	79
<i>Figure 6.23. MAE of power in the NSA as a function of forecast length (y-axis) and forecast validation time (x-axis)</i>	80
<i>Figure 6.24. Similar to Fig. 6.23 except for the SSA power MAE.....</i>	81

<i>Figure 6.25. Power MAE percent improvement of forecast power in the NSA as a function of forecast length (y-axis) and forecast validation time (x-axis).....</i>	82
<i>Figure 6.26. Similar to Fig. 6.25 except for the SSA.....</i>	82
<i>Figure 6.27. Percent improvement in MAE for four forecast lengths as a function of observed power, averaged for all 6 DD episodes, for the NSA (top panel) and for the SSA (bottom panel).....</i>	83
<i>Figure 6.28. Top panel: scatter plot of the control and experimental simulation power errors at each tower site using 15 min data, for all 6 DD episodes, NSA and SSA combined, at forecast hour 00. The 1:1 line is shown in magenta, and the teal line is the best fit to the data. Dashed red and blue lines define the large error thresholds, in this example at 80% power capacity.....</i>	84
<i>Figure 6.29. Power MAE percent improvement for control simulation errors larger than the threshold bin error size, for all 6 DD episodes, NSA and SSA combined.....</i>	85
<i>Figure 6.30. Spatial averaging boxes using an 8x8 grid for the NSA. The 4x4 grid is formed by combining 4 neighboring cells in the 8x8 grid, and the 2x2 grid by combining 16 neighboring cells.....</i>	86
<i>Figure 6.31. Power MAE with different degrees of spatial averaging for all 6 DD episodes for the NSA.....</i>	87
<i>Figure 6.32. MAE percent improvement of forecast power for the various degrees of spatial averaging, for all 6 DD episodes for the NSA.</i>	88
<i>Figure 6.33. Aggregate observed power (black curve), control forecast power (red curve), experimental forecast power (blue curve), and instantaneous MAE percent improvement (green curve).....</i>	89
<i>Figure 6.34. The same as Fig. 6.33 except for the SSA.....</i>	90
<i>Figure 6.35. The two study areas with the more restricted domains shown by the red rectangle in the NSA and circle in the SSA. Tall tower sites falling outside of these two areas are eliminated in the geographic outlier sensitivity analysis.....</i>	91
<i>Figure 6.36. Percent improvement of MAE (top panel) and R^2 (bottom panel) for the NSA (orange) and SSA (green) when using all tower observations (solid) and when using only those tall towers that fall within the restricted analysis domains shown in Fig. 6.35.....</i>	92
<i>Figure 6.37. NAM vector wind RMSE (top) and wind speed bias (bottom) against WFIP wind profiler observations within the 0-2km AGL layer in the northern study region.....</i>	93
<i>Figure 6.38. As in Fig. 6.37 except valid for the southern WFIP study region.....</i>	94
<i>Figure 6.39. As in Fig. 6.37 except forecasts are compared to SODAR observations in the 0-200 m AGL layer. Verification is valid for the northern WFIP study domain.....</i>	95
<i>Figure 6.40. As in Fig. 6.39 except for the southern WFIP study region.....</i>	96
<i>Figure 6.41. Forecast verification regions in the NCEP Forecast Verification System. For conventional verification in WFIP, regions NPL and SPL were combined to form a sub-region for the Plains states.....</i>	97

<i>Figure 6.42. NAM/NAMX and CONUSnest/CONUSnestX 2m temperature bias over the Plains.....</i>	98
<i>Figure 6.43. NAM/NAMX and CONUSnest/CONUSnestX 10 m wind vector RMSE over the Plains.....</i>	99
<i>Figure 6.44. NAM/NAMX and CONUSnest/CONUSnestX 10 m wind speed bias over the Plains.....</i>	100
<i>Figure 6.45. Tall tower u (left) and v (right) observation innovations (observation-forecast) from all analysis steps during the WFIP winter quarter data denial period. Distributions featured along the top are from the NAMX while distributions along the bottom row are from the CONUSnestX. The red dotted lines correspond to Gaussian distributions.....</i>	101
<i>Figure 6.46. Nacelle wind speed observation innovations from all analysis steps during the WFIP winter quarter data denial period from the NAMX (top) and CONUSnestX (bottom). The red dotted lines correspond to Gaussian distributions.....</i>	102
<i>Figure 6.47. Nacelle wind speed observations and innovations depicted as a two-dimensional histogram. Plotted data are from all analysis steps during the WFIP winter quarter data denial period from the NAMX (top) and CONUSnestX (bottom).....</i>	104
<i>Figure 6.48. A comparison of the number of vertical levels from the NAM vs. the levels at which SODAR observations are reported at the Ainsworth, NE SODAR site.....</i>	105
<i>Figure 7.1. Time series of 70m AGL wind speed (top panel) and equivalent wind power (bottom panel) using an ICE2 wind power curve, averaged from the four SDSU tall tower towers in South Dakota, over an 8 day period from March 12-20, 2012.....</i>	106
<i>Figure 7.2 Ramps identified by the Fixed Time Interval Method for a window length WL and a ramp threshold Δp_{RD}.....</i>	109
<i>Figure 7.3. Ramps identified by the Min-Max Method for a window length WL and a ramp threshold Δp_{RD}.....</i>	110
<i>Figure 7.4. Top panel: ramps found by the explicit derivative method for a value of the smoothed derivative threshold given by Δp_{RD} and window length WL. Lower panel: the smoothed power derivative corresponding to the power data in the top panel.....</i>	112
<i>Figure 7.5 Schematic diagram of a weighted ramp matrix. Extreme ramps, with large changes in power over short time intervals, are placed in the top-left corner, while low amplitude ramps of longer duration are placed in the bottom right corner. A weighting function, denoted by the red isopleths, is then applied to each matrix element, before averaging into a single overall model skill score.....</i>	119
<i>Figure 7.6 Time series of power estimated from anemometer measurements on a tall tower and from the RAP control model. Ramp events for the three ramp definition methods are shown using a 50% power change threshold over 2 hours, red for up ramps and green for down. The numbers of ramps found in each time series are shown on the right.....</i>	120
<i>Figure 7.7. The average number of occurrences of ramp events that fall into each matrix bin per DD episode using the Min-Max Method, for the forecast initialization time (hour 00, left panels) and forecast hour 06 (middle panels), for the control (top) and experimental run (bottom) of the ESRL RAP model, NSA</i>	

<i>and SSA combined. The top right panel is the same but for the tall tower observations. The ramp definition power threshold ranges from 30 to 70%, and the window length from 30 to 180 minutes.....</i>	121
<i>Figure 7.8. The matrix of skill scores for the control ESRL RAP forecasts averaged for all 6 DD episodes, NSA and SSA combined.....</i>	122
<i>Figure 7.9. The default weighting scores applied to the matrices of skill scores shown in Fig. 7.8.....</i>	122
<i>Figure 7.10. Skill score results for the Fixed-Time ramp identification method, averaged for all DD episodes, NSA and SSA combined. Top panel: Skill scores for the Control (solid) and Experimental (dashed) data denial simulations, for forecast hours 0-14. Green is the skill score with equal weighting of all matrix elements, purple is when using the weighting matrix shown in Fig. 7.9. Bottom panel: the percent improvement in the experimental forecasts over the control, for the un-weighted (green) and weighted (purple) matrices of skill scores.....</i>	123
<i>Figure 7.11. The same as Fig. 7.10, except for the Min-Max ramp detection method.....</i>	124
<i>Figure 7.12. The same as Fig. 7.10, except for the explicit derivative ramp detection method.....</i>	124
<i>Figure 7.13. The number of matched ramp events per day and per tower site, in each of the 8 ramp scenarios using the Min-Max method for all 5 DD episodes, NSA and SSA combined. Red is for the control simulation, and blue for the experimental. All 20 Δp and Δt combinations shown in Fig. 7.8 are summed to form the number of events.....</i>	125
<i>Figure 7.14. Sum of the matrix-weighted scores for each ramp scenario type, for all 5 DD episodes, NSA and SSA combined.....</i>	126
<i>Figure 7.15 Percent improvement in ramp forecast skill using the Min-Max ramp definition, averaged for the first 9 forecast hours, for each of the 5 DD episodes, and for the NSA (orange) and SSA (green).....</i>	127
<i>Figure 7.16. Ramp forecast skill for up (blue) and down (red) ramps using a symmetric set of scoring scenarios, using the weighted matrix of events. Solid lines are for the RAP control, and dashed lines for the RAP experimental simulations assimilating the WFIP observations.....</i>	129
<i>Figure 8.1. Diurnal averages for friction velocity (u^*) and sensible heat flux for winter, spring, summer, and fall, represented by the months of January, April, July, and October for the Brady (BDY), Colorado City (COC), and Jayton (JTN) sites. CST is Central Standard Time.....</i>	132
<i>Figure 8.2. Fractional errors comparing winds predicted by the logarithmic wind speed and observed winds at the surface (sonic anemometer) and aloft (sodar) for the 3 sites. The notation used in the legend is explained in the accompanying text.....</i>	133
<i>Figure 8.3. Best-fit values for the exponent to the wind speed power law by stability parameter z/L and time of day. Determinations used the 10-minute averages for the sonic anemometer near the surface and sodar at the 80 m height.....</i>	134
<i>Figure A1. Wind profiling radar and sodar inter-comparisons for the Nov 30 – Dec 06, 2011 Data Denial episode.....</i>	147

<i>Figure A2. Wind profiling radar and sodar inter-comparisons for the Jan 07 – Jan 15, 2012 Data Denial episode.....</i>	147
<i>Figure A3. Wind profiling radar and sodar inter-comparisons for the April 14 – April 25, 2012 Data Denial episode.....</i>	148
<i>Figure A4. Wind profiling radar and sodar inter-comparisons for the Jun 09 – Jun 18, 2012 Data Denial episode.....</i>	148
<i>Figure A5. Wind profiling radar and sodar inter-comparisons for the Sept 16 – Sept 25, 2012 Data Denial episode.....</i>	149
<i>Figure A6. Wind profiling radar and sodar inter-comparisons for the Oct. 13-20, 2012 Data Denial episode.....</i>	149

List of Tables

<i>Table 1.1 The types and numbers of meteorological observing instruments.....</i>	11
<i>Table 2.1. List of instrument types, numbers, and providers.....</i>	12
<i>Table 3.1. 13 km RUC domain configuration.....</i>	38
<i>Table 3.2. 13 km Rapid Refresh domain configuration.....</i>	39
<i>Table 3.3. 3 km HRRR domain configuration.....</i>	39
<i>Table 3.4. 12 km NAM domain configuration.....</i>	39
<i>Table 3.5. The NAM 4 km CONUSnest domain configuration.....</i>	40
<i>Table 3.6. Improvements made to the NOAA/ESRL RAP and HRRR research models in 2012.....</i>	41
<i>Table 4.1. Real-time forecast GSI values of observation error and gross error for the assimilated WFIP instrumentation types.....</i>	46
<i>Table 4.2. Data denial simulation GSI values of observation error and gross error for the assimilated WFIP instrumentation types.....</i>	46
<i>Table 6.1 Data types and quantities assimilated in the data denial simulation experiments for both the Northern and Southern Study Areas.....</i>	56
<i>Table 6.2 Dates for six data denial studies.....</i>	57
<i>Table 7.1. Scenario definitions for matched and un-matched ramp events.....</i>	114
<i>Table 7.2. Scores for the four possible null scenarios.....</i>	115
<i>Table 7.3. Range of scores possible for all 8 event scenarios.....</i>	116
<i>Table 7.4 Weights applied for ramp Late Prediction Penalties and Under/Over Prediction Penalties.....</i>	117
<i>Table 7.5 Range of scores possible for all 8 event scenarios for a simplified, symmetric up and down ramp scoring scheme.....</i>	128

List of Acronyms

AGL – above ground level
ANL – DOE/Argonne National Laboratory
ARL – NOAA/Air Resources Laboratory
ASOS – Automated Surface Observing System
COUNS - contiguous United States
CONUSnest -NAM contiguous United States 4 km nest
CONUSnestX -NAM CONUSnest experimental simulations assimilating WFIP observations
CST – Central Standard Time
DD - data denial
DOE - U. S. Department of Energy
ERCOT - Energy Reliability Council of Texas
ESRL – NOAA/Earth Systems Research Laboratory
FTPS - File Transfer Protocol-Secure Sockets Layer
GDAS – Global Data Assimilation System
GFS – Global Forecast System
GSI – Gridpoint Statistical Interpolation
HPC - High Performance Computing
HRRR – High Resolution Rapid Refresh model
IEC2 -International Electrotechnical Commission Class 2
IP - Internet Protocol
LLNL – DOE/Lawrence Livermore National Laboratory
LPP - Late Prediction Penalty
LSM - Land Surface Model
MADIS - Meteorological Assimilation Data Ingest System
MAE – Mean Absolute Error
MHz – Mega-Hertz
MISO - Midwest Independent System Operator
MPP – Multi-Peak Picking
MSL – height above mean sea level
NAM – North American Mesoscale forecast system
NAMX – NAM simulations assimilating the WFIP observations
NCEP – NOAA/NWS/National Centers for Environmental Prediction
NDAS – NAM Data Assimilation System
NMMB – Non-hydrostatic Multiscale Model on the B grid
NOAA - National Oceanic and Atmospheric Administration
NWP – Numerical Weather Prediction
NSA – Northern Study Area (WindLogics domain)
NWS – NOAA/National Weather Service
NREL – DOE/National Renewable Energy Laboratory
PNNL – DOE/Pacific Northwest National Laboratory
prepBUFR - prepared Binary Universal Form for the Representation of meteorological data
RAP – Rapid Refresh model (run at NOAA/NCEP)
RASS – Radio Acoustic Sounding System

RFI – Radio Frequency Interference

RMSE – Root Mean Squared Error

RR – Rapid Refresh model (run at NOAA/ESRL)

RRTM - Rapid Radiative Transfer Model

RUC – Rapid Update Cycle model

SCP - Secure Copy Protocol

SDSU – South Dakota State University

SNR – Signal-to-Noise Ratio

SSA – Southern Study Area (AWS Truepower domain)

TCP - Transmission Control Protocol

TM - TMXX is the model initialization time minus XX hours

TTU – Texas Tech University

UOPP - Under/Over Prediction Penalty

UTC - Universal Time Coordinate, or Greenwich Mean Time

QC – Quality Control

WFIP – Wind Forecast Improvement Project

WL – ramp window length

WPR – Wind Profiling Radar

WRF-ARW - Weather Research and Forecasting model, Advanced Research Weather version

3DVar – Three-dimensional variational data assimilation