THE WIND FORECAST IMPROVEMENT PROJECT (WFIP)

A Public–Private Partnership Addressing Wind Energy Forecast Needs

BY JAMES WILCZAK, CATHY FINLEY, JEFF FREEDMAN, JOEL CLINE, LAURA BIANCO, JOSEPH OLSON, IRINA DJALALOVA, LINDSAY SHERIDAN, MARK AHLSTROM, JOHN MANOBIANCO, JOHN ZACK, JACOB R. CARLEY, STAN BENJAMIN, RICHARD COULTER, LARRY K. BERG, JEFFREY MIROCHA, KIRK CLAWSON, EDWARD NATENBERG, AND MELINDA MARQUIS

An observational, data assimilation, and modeling study demonstrates improvements in the accuracy of wind forecasts for wind energy.

Wind power is a variable energy source, dependent on weather conditions. Electric grid operators keep the grid stable by balancing variable generation resources (e.g., wind and solar) and conventional generation (e.g., coal, gas, and nuclear) with energy demand. Having accurate advance knowledge of the amount of wind power available through reliable weather forecasts can lead to improvements in the efficiency of the entire electrical grid system, including the operation of fossil fuel plants, resulting in lower costs as well as lower CO_2 emissions (Marquis et al. 2011; GE Energy 2010; EnerNex 2011). 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, thereby mitigating anthropogenically forced climate change while also reducing air pollution.

The U.S. Department of Energy (DOE) sponsored the Wind Forecast Improvement Project (WFIP) with the goal of advancing the integration of wind power and reducing the cost of energy by improving short-term wind energy forecasts, including forecasts of ramp events (large changes in wind power production over short time intervals). WFIP was a public–private partnership with two private sector teams led by forecasting companies WindLogics and AWS Truepower that collaborated with DOE and National Oceanic and Atmospheric Administration (NOAA) laboratories and with the National Weather Service. The core of WFIP was composed of two concurrent year-long field programs in high wind energy resource areas of the United States (the upper Great Plains and Texas).

WFIP employed two avenues for improving wind energy forecasts: enhanced measurement networks and numerical weather prediction (NWP) model system advancements. The former included networks of in situ and remote sensing instruments deployed in the two study areas, including proprietary tall tower and turbine nacelle (i.e., the housing containing the generator and gearbox) anemometer observations from the wind energy industry, and for the first time assimilating the proprietary data into NOAA's 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 was to provide observations through a deep layer of the atmosphere, and over a sufficiently broad area, to influence NWP forecasts out to at least a 6-h lead time. These observations were assimilated into real-time forecasts as well as retrospective simulations spanning the WFIP field campaign year to allow for an evaluation of seasonal differences in the skill of the models and the impact of the observations.

Wind energy forecasting efforts have often focused on day-ahead forecasting (typically 18–42 h ahead) because some conventional power plants, such as coal plants, are normally scheduled that far in advance. But additional opportunities exist on the short-term (0–6 h) time frame to adjust schedules, start natural gas generators, minimize scheduling errors in the energy markets, or use bilateral trading to take advantage of improved forecasts. Even when a wind power plant has been scheduled a day ahead, a more accurate forecast for the next few hours is important for balancing day-ahead forecast errors, minimizing penalties, and maximizing revenues. This short-term time frame was the focus of WFIP.

The second avenue for enhancing wind energy forecasts was to improve the NWP forecast systems used by all partners directly and to produce a broad assessment of model skill specifically validated with *turbine-height* wind observations. Midway through the WFIP field program, the NOAA/National Weather Service (NWS) upgraded its operational hourly updated NWP forecast model from the Rapid Update Cycle (RUC) model to the Rapid Refresh (RAP) model, which included improvements resulting from WFIP, and the impacts of this upgrade were evaluated using WFIP observations. In addition, improvements to the research version of RAP and to NOAA's High-Resolution Rapid Refresh (HRRR) research model were continuously made during WFIP. WindLogics incorporated forecasts from these improved NOAA models into machine learning energy prediction algorithms, while the AWS Truepower team developed and operated in real time an experimental nine-member optimized ensemble forecast system for WFIP that utilized RAP and HRRR for initial and boundary conditions. Also, with WFIP funding, NOAA developed a data-dissemination capability to make the large amounts of raw model output from the HRRR model available in real time to the two private sector teams and to the entire wind energy industry.

Complementary research to the work discussed here addressed the accuracy of using stability-dependent wind profile relationships to estimate the wind speed at turbine-hub height, the development of a community ramp tool and metric, development of a gap-filling and quality control algorithm for remote sensing data, and various sensitivity studies examining the role of observation type and data assimilation techniques in model performance.

The WFIP program builds upon a long research effort within the meteorological community aimed at providing better environmental information to support wind energy. The first computer models devoted to wind power forecasting were developed during the 1980s, an outgrowth of a Pacific Northwest National Laboratory (PNNL) working group (Wendell et al. 1978; Bossanyi 1985). Throughout the 1990s, a variety of statistical approaches were employed to improve forecast skill. In 1999, eWind, the predecessor of the forecasting system used in the WFIP southern study area (SSA), was developed by AWS Truepower (AWST). In the early 2000s, the California Independent System Operator (CAISO) developed a centralized wind power forecasting system (Makarov et al. 2010). Since then, a large number of ISOs, utilities, and balancing authorities have deployed wind power forecasting systems in the United States, all of which are dependent on national-scale or global wind forecasts and observations provided by NOAA or other national forecasting centers. With deeper

AFFILIATIONS: Wilczak, Benjamin, and Marquis—NOAA/Earth System Research Laboratory, Boulder, Colorado; Finley, Sheridan, and Ahlstrom—WindLogics Inc., St. Paul, Minnesota; Freedman*—AWS Truepower, Albany, New York; Cline—DOE/Energy Efficiency and Renewable Energy, Washington, D.C.; Bianco, Olson, and Djalalova—University of Colorado/CIRES, Boulder, Colorado; Manobianco, Zack, and Natenberg—MESO, Inc., Troy, New York; Carley—IM Systems Group, NOAA/National Weather Service, College Park, Maryland; Coulter—DOE/Argonne National Laboratory, Lemont, Illinois; Berg—DOE/Pacific Northwest National Laboratory, Richland, Washington; Mirocha—DOE/Lawrence Livermore National Laboratory, Livermore, California; Clawson—NOAA/Air Resources Laboratory, Idaho Falls, Idaho

*** ADDITIONAL AFFILIATION:** Atmospheric Sciences Research Center, University at Albany, State University of New York, Albany, New York

CORRESPONDING AUTHOR: Dr. James M. Wilczak, NOAA/Earth System Research Laboratory, 325 Broadway, Mail Stop PSD3, Boulder, CO 80305
E-mail: james.m.wilczak@noaa.gov

The abstract for this article can be found in this issue, following the table of contents.

DOI:10.1175/BAMS-D-14-00107.1
In final form 31 October 2014

penetration of wind energy, accurately forecasting the wind is ever more critical for developing and managing the modern electrical grid (Monteiro et al. 2009; Mahoney et al. 2012; Giebel and Kariniotakis 2007). Thus, the WFIP research effort seeks to complement and continue the evolution of wind energy forecasting, further facilitating the development and operation of wind power production in the United States.

PARTNERSHIPS. A strength of WFIP was its collaborative framework, bringing together federally funded laboratories and centers, private sector companies, and universities. Participants included several Department of Energy (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 System Research Laboratory (ESRL) and Air Resources Laboratory (ARL)], the NWS, and two teams of partners from the private sector and university communities. One team, led by AWS Truepower, included the Electric Reliability Council of Texas (ERCOT), which operates the electric grid in most of Texas. A second team was led by WindLogics, Inc. (a subsidiary of NextEra Energy), and included the Midcontinent Independent System Operator (MISO), which operates the electric grid in the upper Midwest states. Additional observations were also made available in-kind by Iberdrola USA, Leosphere, and West Texas A&M University (WTAMU). A list of team partners is provided in Table 1.

OBSERVATIONS. New instrumentation was deployed or acquired in two high wind energy resource areas of the United States during concurrent year-long field campaigns that ran from September 2011 to September 2012. The first area was in the upper Great Plains (Fig. 1), or the northern study area (NSA), where DOE and NOAA partnered with the WindLogics team. The second field campaign was centered over the SSA in western and central Texas (Fig. 2), where DOE and NOAA partnered with the AWS Truepower team. A vital and ultimately successful aspect of WFIP was the collaboration between the wind energy industry and NOAA in acquiring and assimilating for the first time into NOAA's models proprietary data from tall tower (mostly 60 m) and wind turbine nacelle-mounted anemometers (Table 2). All 405 nacelle anemometer sites and a subset of the tall tower sites were available for assimilation in real time (41 out of 133 in the NSA, 27 out of 51 in the SSA), while the remaining industry tall towers were available after a several-day delay for use in retrospective modeling studies. The WFIP observing systems also included 12 wind profiling radars, 12 sodars, several lidars, and 71 surface meteorological stations. Observations from the radar wind profilers and sodars, as well as surface meteorological stations, also were assimilated into the NWP models used to make wind power forecasts. The primary observations used for evaluating model performance are the tall tower data, wind profiler data, and power output from 23 wind plants in the NSA and 34 wind plants in

Table 1. WFIP team partners and collaborating institutions.

Northern Study Area (NSA)	Southern Study Area (SSA)
WindLogics	AWS Truepower
NextEra Energy Resources	MESO, Inc.
Midcontinent Independent System Operator (MISO)	Electric Reliability Council of Texas (ERCOT)
South Dakota State University (SDSU)	Texas Tech University (TTU)
	University of Oklahoma's Center for Analysis and Prediction of Storms (OU CAPS)
	North Carolina State University (NCSU)
	ICF International
NOAA	**DOE Laboratories**
Earth System Research Laboratory	National Renewable Energy Laboratory
Air Resources Laboratory	Argonne National Laboratory
NWS/National Centers for Environmental Prediction	Lawrence Livermore National Laboratory
	Pacific Northwest National Laboratory

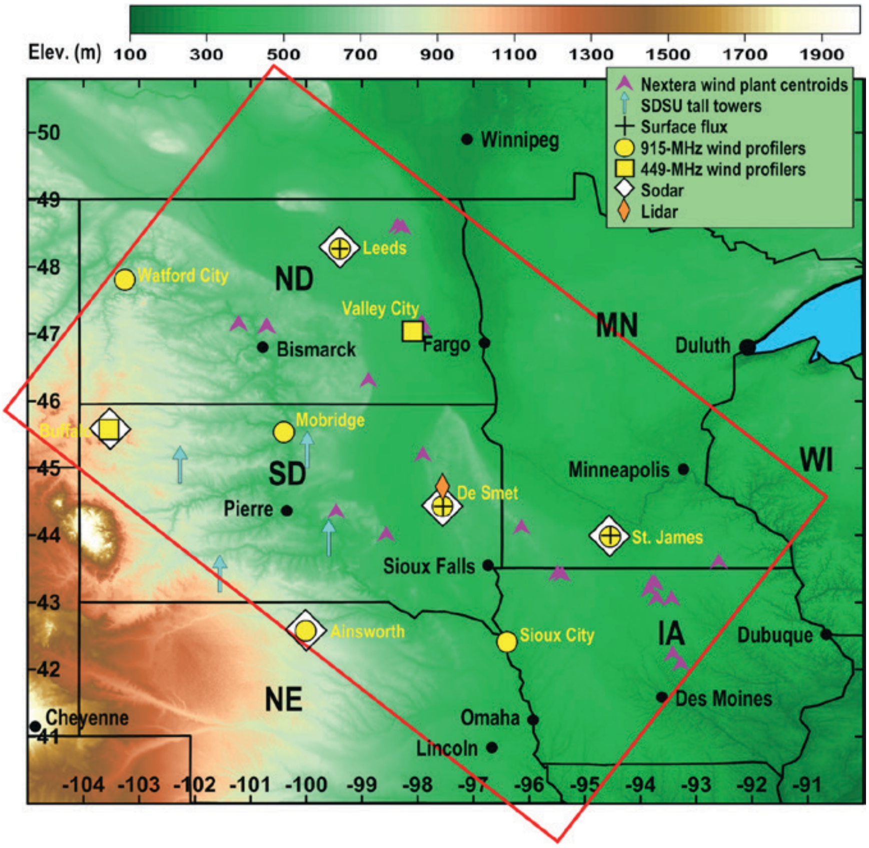

Fig. 1. Geographic domain of the NSA. Surface elevation is shown by color shading. Instrument types and locations are shown, as well as the locations of the NextEra wind plants. Four publicly available tall-tower sites from South Dakota State University (SDSU) are shown, but the industry-provided tall-tower locations are proprietary and are not displayed.

the SSA. The wind plant power data were independent of the data assimilation process.

To assist in maintaining the WFIP instrumentation throughout the year-long field campaign and in identifying potential model problems, the observations and model forecasts were displayed continuously on a real-time publicly accessible website, updated on a subhourly basis, with separate websites for the proprietary data. The public websites for the NSA and SSA can be found online (http://wfip.esrl.noaa.gov/psd/programs/wfip/).

A 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 an NWP model. New processing techniques were implemented to reduce spurious signals from birds and other contamination in wind profiling radar data (Bianco et al. 2013; Wilczak et al. 1995), and techniques were also developed to identify and correct direction offsets in the tall tower observations, as discussed in Wilczak et al. (2014). The wind profiler QC technique continues to be applied to boundary layer profiler wind observations with QC flags relayed through the Meteorological Assimilation Data Ingest System (MADIS; http://madis.noaa.gov), improving their assimilation into NOAA real-time model forecast systems.

NOAA WEATHER MODELS. Because of the focus on short-term forecasts (up to 15 h), the principal NOAA models used during WFIP were the hourly updated 13-km-resolution RUC, the 13-km-resolution RAP, and the 3-km-resolution HRRR (Fig. 3). RAP and HRRR both used version 3.4.1 of the Advanced Research core of the Weather Research and Forecasting Model (ARW; Skamarock et al. 2008). RUC and RAP provided forecasts out to 18 h, while HRRR provided forecasts out to 15 h.

RUC was the NOAA/NWS/National Centers for Environmental Prediction (NCEP) operational hourly updated forecast system through the first half of the WFIP field campaign, when it was replaced by RAP on 1 May 2012 (Table 3). Prior to this date, the RAP model was run at NCEP in a test mode, and we refer to both the operational and test versions as the NCEP_RAP. Research versions of the RAP (ESRL_RAP) and HRRR models were run in real time 24 h a day for 7 days a week by NOAA/ESRL through the entire WFIP campaign; these versions differed from NCEP_RAP as improvements to the model physics were incorporated over time. In particular, estimates of surface aerodynamic roughness lengths were improved, and the vertical resolution in the land surface model was increased, which improved the boundary layer diurnal cycle.

All of these models use three-dimensional (3D) variational data assimilation, with RAP using the

Gridpoint Statistical Interpolation (GSI) analysis system (Wu et al. 2002). The GSI is capable of assimilating a diverse set of observations, and new capabilities for assimilating energy-related observations (tall towers and nacelle anemometers) were developed for GSI as an outcome from WFIP. HRRR was in development during WFIP and at that time did not perform data assimilation on the 3-km grid, but used initial and boundary conditions obtained by direct interpolation from the 13-km ESRL_RAP. (In 2013, a 3-km data assimilation system was added to the HRRR.)

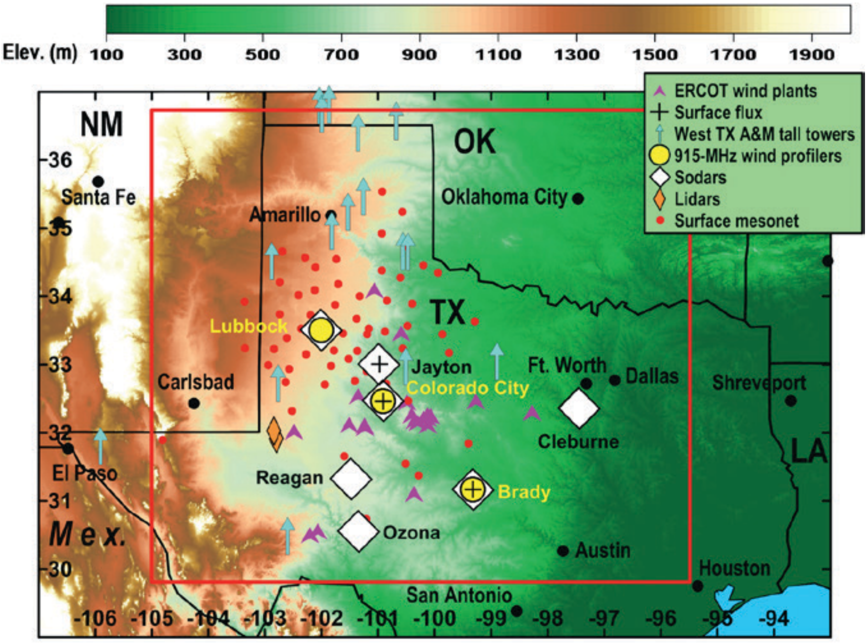

Fig. 2. Geographic domain of the SSA. Surface elevation is shown by color shading. Instrument types and locations are shown, as are the locations of WFIP-participating wind plants providing power to ERCOT. Publicly available tall-tower sites from West Texas A&M University are shown, but the industry-provided tall-tower locations are proprietary and are not displayed.

The intent during WFIP was to assimilate the special WFIP observations into the research ESRL_RAP and HRRR models, but not into the operational NCEP_RUC and NCEP_RAP, and then to compare the skill of these models. Inadvertently, the NCEP_RAP assimilated a small subset of the WFIP observations (one wind profiling radar and five sodars in the NSA; three sodars in the SSA; none of the tall towers, nacelle anemometers, or surface mesonet). Since these observations will have added some skill to NCEP_RAP, comparisons of the skill of ESRL_RAP and HRRR to the NCEP_RAP model, shown later in the "Northern study area" and "Southern study area" sections, will provide a conservative estimate of what the improvement would have been had none of the WFIP observations been assimilated into NCEP_RAP.

The latency of the NCEP_RUC, NCEP_RAP, and ESRL_RAP models was approximately 1 h during WFIP, while the latency of ESRL_HRRR was 1.5–2 h. When HRRR became operational at NCEP in 2014, the latency was reduced to approximately 1 h. All model forecast intercomparisons shown are independent of time latency. Details on RUC can be found in Benjamin et al. (2010), and for RAP and HRRR can be found online (http://rapidrefresh.noaa.gov and http://rapidrefresh.noaa.gov/hrrr, respectively).

Table 2. The types and numbers of meteorological observing instruments deployed in the two study domains. W-P designates wind profiling.

Instrument	NSA	SSA
915-MHz W-P radar	7	3
449-MHz W-P radar	2	
Doppler W-P sodar	5	7
W-P lidar	1	2 (short term)
Surface flux station	3	3
Surface meteorological station	8	63
Tall towers	133	51
Nacelle winds	405	—

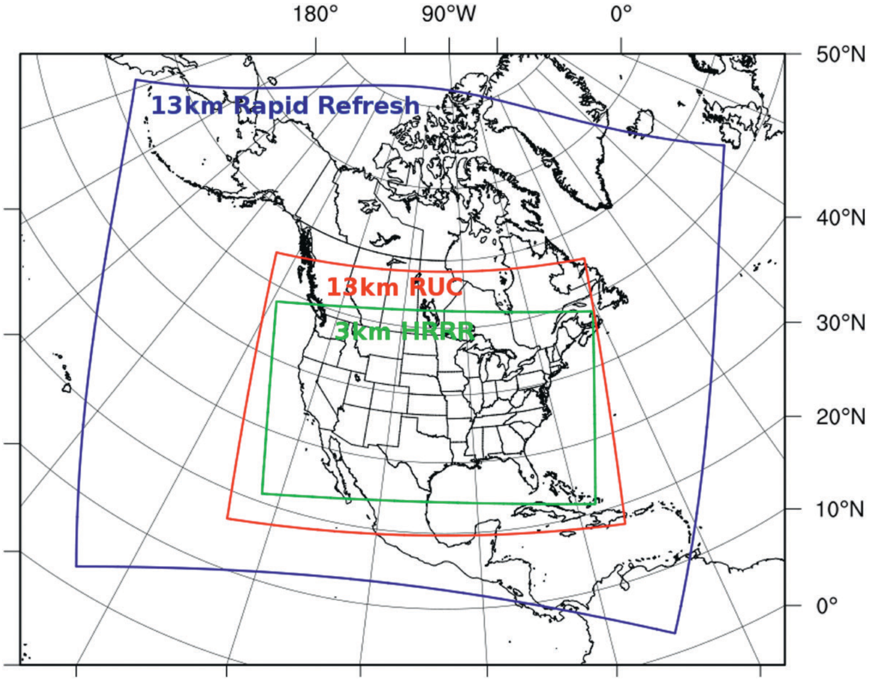

Fig. 3. NOAA model domains for the 13-km RAP (blue), the 13-km RUC (red), and the 3-km HRRR (green) simulations.

REAL-TIME FORECAST ERROR STATISTICS: RAP AND RUC MODELS.

NCEP_RUC was used as a baseline forecast against which to compare the upgraded WFIP forecasts until RUC ceased operations on 1 May 2012. The NCEP_RUC model did not assimilate any of the new WFIP observations while the research ESRL_RAP did, and so a comparison of these two models combines fundamental model improvements of RAP over RUC, as well as the impacts of assimilation of the WFIP data. The tall tower datasets are the primary source used for this evaluation. 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. To convert wind speed into power, we used a generic International Electrotechnical Commission class 2 (IEC 2005) wind turbine power curve, which is the most common type of wind turbine deployed by the NSA and SSA wind generator partners.

We have chosen to use a simple mean bias correction for the RUC–RAP comparison, and for the data-denial analysis that follows, after testing indicated that although more complex bias-correction techniques reduced the overall model error, they did not significantly alter the relative improvement between models or the improvement due to the assimilation of the new observations. The percentage root-mean-square error (RMSE) relative improvement of the ESRL_RAP model over NCEP_RUC is shown in Fig. 4 for the bias-corrected power evaluated using the 41 real-time tall towers in the NSA and the 27 real-time tall towers in the SSA. The improvement in hub-height power ranged from 12% to 5% for forecast hours 1–12.

For comparison purposes, an analysis of the operational NOAA/NWS North American Mesoscale Forecast System (NAM) and Global Forecast System (GFS) models 850-hPa vector-wind RMSE (using radiosondes for verification) encompassing the past 10 years (www.emc.ncep.noaa.gov/mmb/verif/vlcek/) over North America shows an annual improvement for the 12-h NAM forecasts (using the Eta Model before 2006) of 0.7% yr^{-1}, while for the GFS the value is approximately 0.4% yr^{-1}. Repeating the analysis shown in Fig. 4 for hub-height vector winds instead of power, similar improvements of 12%–5% for forecast hours 1–12 are found (not shown). Thus, the regional improvement from the combination of

Table 3. NOAA research and operational models and the dates that they provided forecasts for WFIP.

NOAA research NWP models	NOAA/NWS NWP models
	NCEP_RUC, Sep 2011–30 Apr 2012
ESRL_RAP, Sep 2011–Sep 2012	NCEP_RAP, Sep 2011–Sep 2012 (fully operational on 1 May 2012)
HRRR, Sep 2011–Sep 2012	NAM/NAM CONUS nest, Sep 2011–Sep 2012

RAP and assimilation of the WFIP observations in the NSA and SSA represents close to a *decade's* worth of improvement typically found in the operational models over North America, marking a significant advancement for the wind energy industry.

DATA-DENIAL NWP EXPERIMENT RESULTS. One of the primary goals of WFIP was to determine the impact of the special WFIP observations on the model forecast skill of turbine hub-height winds. Isolating the impact of the new observations required controlled data-denial simulations, where the identical NWP model was run twice: first as *a control run* that assimilated only the routinely available observations and second as *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.

Six separate data-denial episodes were chosen, ranging in length from 7 to 12 days, for a total of 55 days (Table 4). The intent in selecting these days was to get a distribution through all four seasons of the year. In addition, weeks were chosen when few observations were missing, a variety of meteorological phenomena were sampled (cold fronts, low-level jets, thunderstorms), and there were large-amplitude ramp events.

Figure 5 displays RMSEs of the tall tower–derived wind power for the control and experimental simulations, both using the RAP model. The RMSE (Fig. 5, top panels) is expressed as a percentage of the maximum wind power capable of being generated (the rated power). For all hours in both the NSA and SSA, the experimental simulations (that assimilate the WFIP observations) have smaller or equal RMSEs than the control. The improvement is slightly larger in the NSA, where there were more observations assimilated over a larger domain, than in the SSA. The bottom panels in Fig. 5 show the difference between the two curves of the top panels; this difference, which defines the improvement in the forecast, is approximately 1% of capacity at forecast hour 1 and is statistically significant through forecast hour 7 in the NSA, and through forecast hour 4 in the SSA, at the 95% confidence level. When expressed as a relative percentage improvement, the maximum RMSE improvement (at forecast hour 1) in the bottom two panels in Fig. 5 is equivalent to approximately 5%–6%. Similar magnitudes of improvement were found for r^2, the coefficient of determination, for the NSA and SSA (not shown). Interestingly, the SSA has higher values of RMSE than the NSA, perhaps as a result of more prevalent and more difficult to forecast low-level jets (e.g., Freedman et al. 2008), the presence of complex terrain (many of the wind plants are on mesa tops), and possibly more frequent convection. Additional statistical analyses can be found in the three DOE final reports from NOAA, WindLogics, and AWS Truepower (Wilczak et al. 2014; Finley et al. 2014; Freedman et al. 2014).

To demonstrate that the WFIP observations also have a positive impact on a deeper layer of the atmosphere than only at turbine heights, we show the

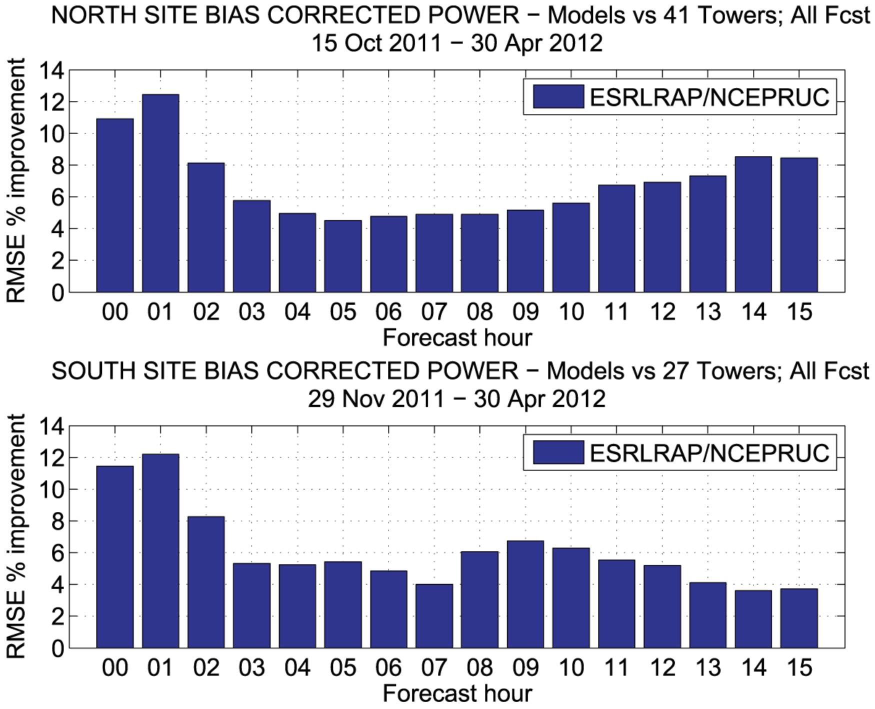

Fig. 4. Percentage RMSE relative improvement of the ESRL_RAP model over the NCEP_RUC model [defined as 100 × (NCEP_RUC$_{\text{RMSE}}$ − ESRL_RAP$_{\text{RMSE}}$)/NCEP_RUC$_{\text{RMSE}}$] for power as a function of forecast length, calculated using observations from (top) 41 real-time tall tower sites in the NSA and (bottom) 27 real-time tall-tower sites in the SSA during approximately the first half of the WFIP field campaign.

improvement in vector-wind 0–2-km layer-averaged RMSE, using data from the radar wind profilers as verification (Fig. 6). The RMSE improvement is large at the initialization time (hour 0), indicating a closer model fit to the newly assimilated observations, and this improvement diminishes with time but remains statistically significant for the first eight forecast hours. Previous profiler data-denial experiments (Benjamin et al. 2004, 2010) showed similar short-range forecast impacts from regional profiler networks.

In addition to the hourly updated RAP model, data-denial assimilation experiments were also run with the NOAA/NWS NAM 12-km parent and 4-km CONUS nest domains. These results are consistent with the RAP experiments, as discussed in Wilczak et al. (2014).

SPATIAL AVERAGING. The statistics shown in Fig. 5 are averages of power RMSE calculated at individual point locations. These statistics quantify forecast skill applicable to 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 observations with spatially averaged model forecasts. For example, a grid operator may be more interested in the aggregate wind power of the entire balancing area or the aggregate power in a geographic area that feeds into one transmission node. Spatially averaged forecast skill can differ from the average skill of individual point locations if there are compensating errors, where an overforecast at one point tends to balance an underforecast at another point. This difference is the same as that found between forecasting precipitation at a point location versus a catchment basin that spans many model grid points.

To evaluate the effects of spatial averaging, we used forecasts and observations from the NSA, since that domain had tower data covering a larger geographic area than the SSA. First, an 8 × 8 grid was overlain on the NSA domain (Fig. 7, left panel), with each grid box approximately 100 km (north–south) × 150 km (east–west). Within each of these grid boxes all of the tower observations and power forecasts for those towers were averaged at each hour. The RMSE was then computed for each of these 64 sets of aggregated observations and forecasts and averaged. The process was then repeated using a 4 × 4 grid, a 2 × 2 grid, and finally averaging the observations and forecasts for all of the tower sites together (a 1 × 1 grid) and, then, calculating the RMSE.

The forecast power RMSEs for the various degrees of spatial averaging are shown in Fig. 7 (right panel), with the solid curves for the average of the 55 days of the data-denial control simulations and the dashed curves for the experimental simulations assimilating the new WFIP observations, using all 133 tall towers for verification. The reduction in RMSE provided by spatial averaging is very large, with more than a factor of 2 difference between treating each tower individually to when all towers are aggregated together.

The difference between the dashed lines and solid lines shows the improvement from assimilating the new WFIP observations at the various degrees of spatial averaging. Interestingly, although the RMSE decreases continuously with more spatial averaging, the improvement from assimilating the WFIP observations remains fairly constant for all size averages until it finally decreases in the 1 × 1 box when all towers are combined into a single aggregate. This indicates that even for moderately large aggregation areas (the 2 × 2 boxes are 400 km × 600 km) forecasts can be improved significantly with assimilation of new observations. In this case the improvement averaged for forecast hours 1–6 for the 2 × 2 grid is 0.8% of rated capacity, while the relative improvement (0.8/13 × 100) is 6%.

The preceding analyses focused on the NOAA RAP and RUC forecast models and used data-denial studies to determine the impact of the new observations on forecast skill, using tall-tower or wind profiler observations for verification. The next two sections focus on specific WindLogics results from the northern study area and then AWS Truepower results from the southern study area, both using actual wind plant

Table 4. Dates for six data-denial studies.

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

power output for model evaluation (including ramp events), and both evaluating the impact of the HRRR model. The southern study area analysis also focuses on the use of ensemble forecast systems for wind energy and addresses the impact and forecasting of low-level jets.

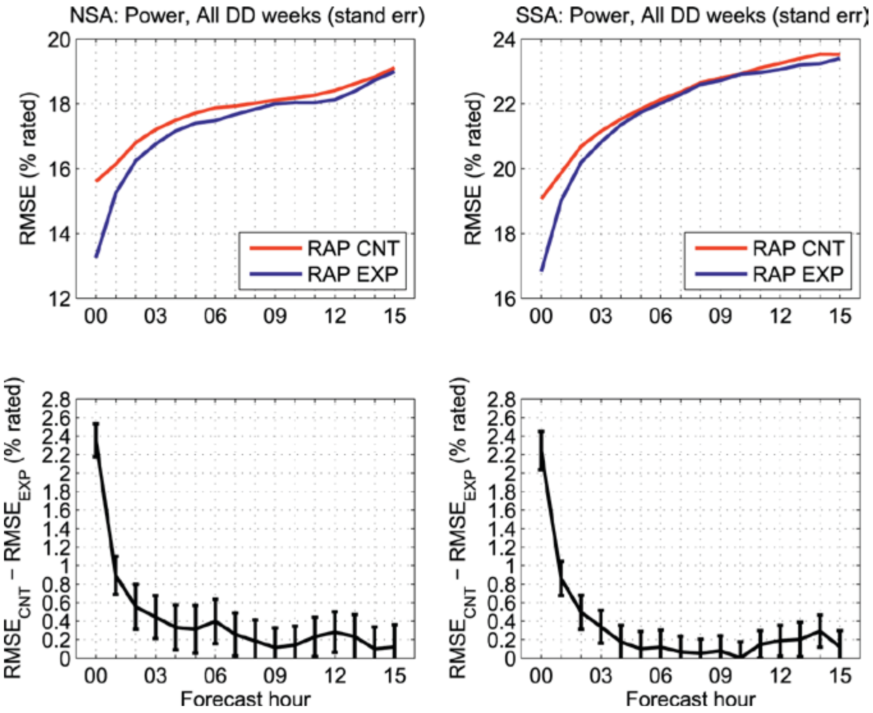

Fig. 5. RMSE for hub-height power, expressed as a percentage of maximum power (rated) capacity, for the (top left) NSA and (top right) SSA for all 55 data-denial episode days. The red curve is for the control simulations, and the blue curve is for the experimental simulations that also assimilated the WFIP observations. (bottom) The difference in RMSEs between the control and experimental simulations. Error bars represent the 95% confidence intervals defined as $(\pm 1.96\sigma/\sqrt{n})$, where n is reduced by the autocorrelation of the time series.

NORTHERN STUDY AREA. In the NSA, WindLogics made wind power forecasts for 23 NextEra Energy operational wind plants (as shown in Fig. 1). Aggregate forecasts were also created by summing forecasts for the individual plants. The analysis presented here will focus on the aggregate results. All forecast skill evaluations are based on observed wind plant power production.

The wind power forecasts generated from the various models applied several levels of postprocessing, a common practice among commercial forecast vendors. Although evaluation of postprocessing techniques commonly used in the wind energy forecasting sector was not the focus of WFIP, it is important to assess whether the fundamental wind speed forecast improvements achieved for the raw forecasts remain after typical postprocessing is applied to them. The levels of forecasts included 1) a "raw" forecast made using hub-height wind forecasts directly from the models and converted into power using a plant-specific power curve; 2) a "bias corrected" forecast made by calculating a rolling 2-week hub-height wind speed model bias relative to the turbine nacelle anemometer measurements over the previous 2 weeks (after a turbine-manufacturer-provided blade-wash correction was applied to them) for every forecast hour, and then bias correcting the wind speed before applying a plant-specific power curve; 3) a "trained" forecast that utilized sophisticated methods for creating nonlinear regression functions from a set of training data (Support Vector Machine, SVM; Cortes and Vapnik 1995; Chang and Lin 2011) using NWP model data as inputs and observed wind plant power data as the "target" variable to statistically correct the individual model-based wind power forecasts; and 4) a "trained ensemble," which combined trained forecasts from a short-range model with the NAM and the local wind plant persistence forecast to generate wind power forecasts. Note that to accurately predict power from an operating wind plant the forecast must consider turbine waking, which is accounted for explicitly in the bias-corrected forecasts and implicitly in the trained forecasts.

Since several months of data are required for the training process, the trained forecasts were generated starting in January 2012 and continued through the WFIP field campaign ending in August 2012. The forecast system was trained monthly using hourly data from the start of the field campaign through the end of a given month and, then, was used to produce trained forecasts for the following month. A comparison of the system-aggregate raw, bias-corrected, and trained power forecast RMSEs, expressed as a percentage of the rated capacity, for the first 12 forecast hours and for the 8-month period, is shown in Fig. 8. As can be

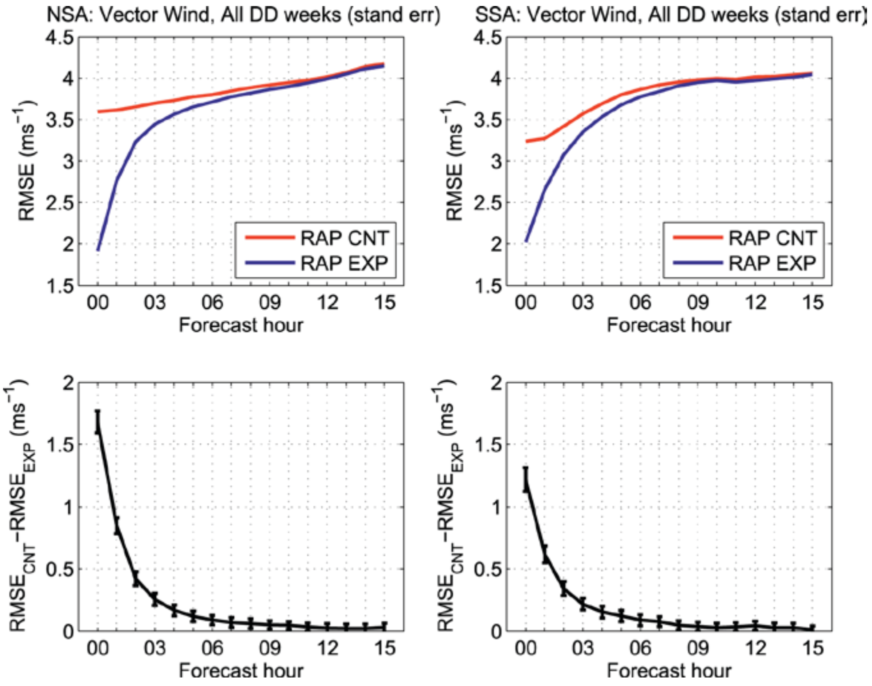

Fig. 6. As in Fig. 5, but for the 0–2-km layer-averaged RMS vector error from the RAP-based data-denial experiments, using all WFIP radar wind profiler observations for verification.

seen, bias correcting the model wind speeds prior to calculating the power improves the power forecasts for all three models (HRRR, ESRL_RAP, and NCEP_RAP), but particularly for HRRR. ESRL_RAP, which assimilates the WFIP observations, is more skillful than NCEP_RAP, which does not assimilate the WFIP observations. At the nonaggregated wind plant level (not shown), the ESRL_RAP-based bias-corrected power forecasts also had the lowest forecast errors at most wind plant locations, followed by the HRRR and NCEP_RAP bias-corrected forecasts, respectively.

For all models, the training process further improves the overall RMSE compared to the simpler bias-correction method, with absolute percentage improvements of 0.3%–0.6% of rated capacity averaged over the first 12 forecast hours (with improvements of 1%–2% of rated capacity compared to the raw forecasts). While the training process reduces the forecast error differences somewhat between the various model-based forecasts, the ESRL_RAP trained forecasts (which assimilate in the WFIP observations) are still the best of the individual models, with RMSEs lower by 0.62% of rated capacity compared with the NCEP_RAP trained forecasts.

Typically, the WindLogics operational forecasts are made from an ensemble of several trained models that also include persistence information for the first several forecast hours. An example of such an operational forecast is shown in Fig. 8 for the ESRL_RAP (plus NAM) two-member ensemble. All trained ensemble forecasts have lower RMSEs in the first 2 h when persistence information adds skill, with average RMSE improvements of 0.1% (trained ensemble ESRL_RAP plus NAM versus trained ESRL_RAP) and 0.7% (trained ensemble HRRR plus NAM versus trained HRRR) of rated capacity at later forecast hours resulting from the use of the ensemble.

While standard bulk statistical error metrics are useful for gauging forecast skill, they often are inadequate in capturing a complete sense of forecast impacts. This is particularly true for power system operations because reliability of the grid is of primary importance. Grid operators are concerned about very large forecast errors that develop quickly, even if they occur only rarely, because they can influence the reserves and operating practices that operators use to ensure reliability. These large forecast errors often occur as a result of actual or predicted "ramp events," when wind power production is changing rapidly. The ramp rate is of particular concern as there are limits to how quickly conventional generation can be ramped up (or down) to offset the changes in wind generation in order to keep the system balanced. Since grid operators must balance the system on a minute-by-minute basis, in this ramp analysis the observations were used at their highest temporal resolution (10 min) and the hourly model output was interpolated to these 10-min intervals.

Because the definition of a "wind energy ramp event" will vary from one operating area to another depending on the penetration of wind on the system and the other types of generation available, a suite of ramp definitions was used as follows:

- The power changes *X*% (of rated capacity) over a *Y*-h period (or less).
- The event can be longer than *Y* h as long as *d*(power)/*dt* ≥ (*X*% capacity)/(*Y* h) occurs at some point during the event.

- The beginning (end) of a ramp occurs when the 10-min ramp rate exceeds (falls below) 2.5% of the defined threshold rate.
- A correctly forecast ramp (or "hit") occurs when the midpoint of the predicted ramp is within ±2 h of the observed ramp midpoint and the magnitude of the predicted ramp is within ±50% of observed (and of correct sign).

Reasonable ranges of *X* and *Y* were chosen from scatterplots of observed ramp amplitude versus ramp duration for the entire system (or an individual site) for all ramps that equaled or exceeded 15% of the rated capacity [i.e., large enough to potentially create issues in Midcontinent Independent System Operator (MISO) grid system operations at the time of this study].

Several metrics were calculated to evaluate the accuracy of the wind power forecasts during wind energy ramp events. The metrics (calculated both at the wind plant level and the system aggregate level) included frequency bias, probability of detection, false-alarm rate, and critical-success index (or threat score), as well as errors in ramp-event timing, magnitude, duration, and ramp rate.

The frequency bias is defined as the ratio of the number of forecasted ramp events to the number of observed events and is shown in Fig. 9 for the aggregate of forecasts over a subset of ramp definitions. Results from a suite of ramp definitions are shown to illustrate the range of variability in the metrics for a given model forecast and the consistency in the results across ramp definitions. Only the bias-corrected forecasts are shown, as the trained forecasts were statistically optimized by minimizing the RMSE at the expense of losing the sharpness of ramp events, such that trained forecasts are not optimal for forecasting ramp events. As can be seen in Fig. 9, the bias-corrected ESRL_RAP-based power forecasts most accurately predict the total number of aggregate ramp events on average. (A frequency bias value of 1 indicates that the model predicts the same number of the events as was observed.) The HRRR-based forecasts tend to overpredict the number of aggregate ramp events by about 9%, and the NCEP_RAP-based forecasts tend to underpredict the number of events by about 10%. When a similar analysis is performed at the wind plant level, all model-based bias-corrected power forecasts tend to underpredict the number of events for most ramp definitions, but HRRR-based power forecasts do a significantly better job at forecasting the number of events for all ramp definitions as compared to the other forecasts (not shown), likely because of its higher resolution.

A comparison of the probability of detection values for all three bias-corrected model-based system-aggregate power forecasts for a subset of ramp-event definitions is also shown in Fig. 9. Probability of

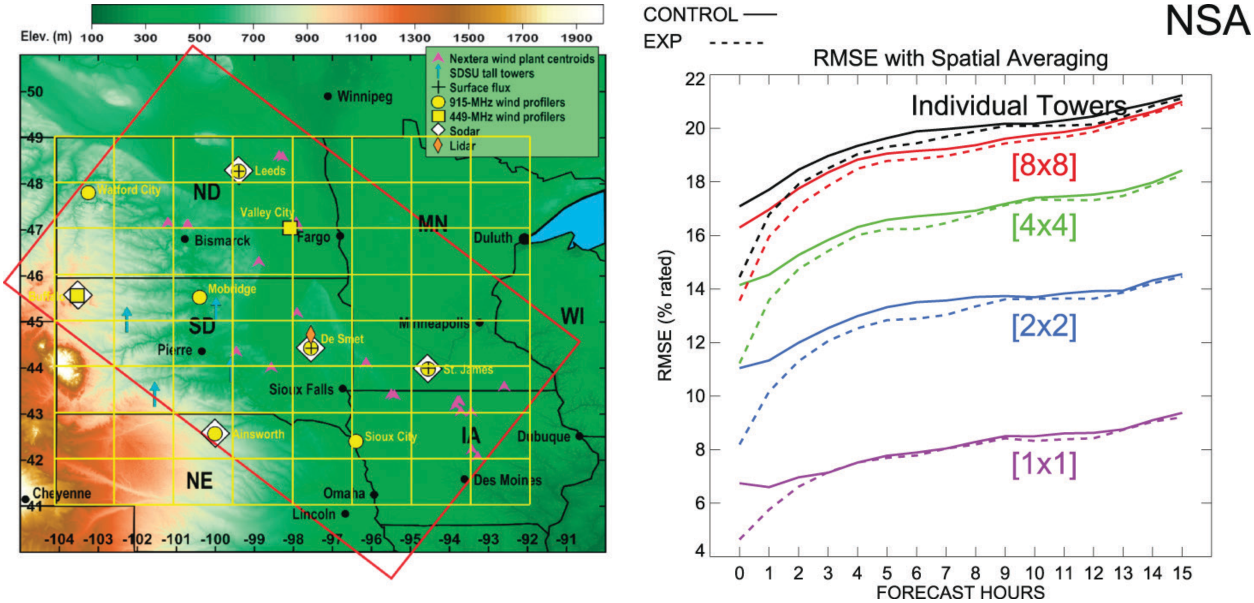

Fig. 7. (left) Grid used for spatial averaging in the NSA. (right) RMSE of power, expressed as a percentage of rated power, with different degrees of spatial averaging for all six data-denial episodes for the NSA. The solid lines are for the control simulations, and the dashed lines are for the experiments that assimilate the WFIP observations. The black lines are with no spatial averaging, and the purple lines are for the maximum spatial averaging with all tower location observations and forecasts aggregated into single time series.

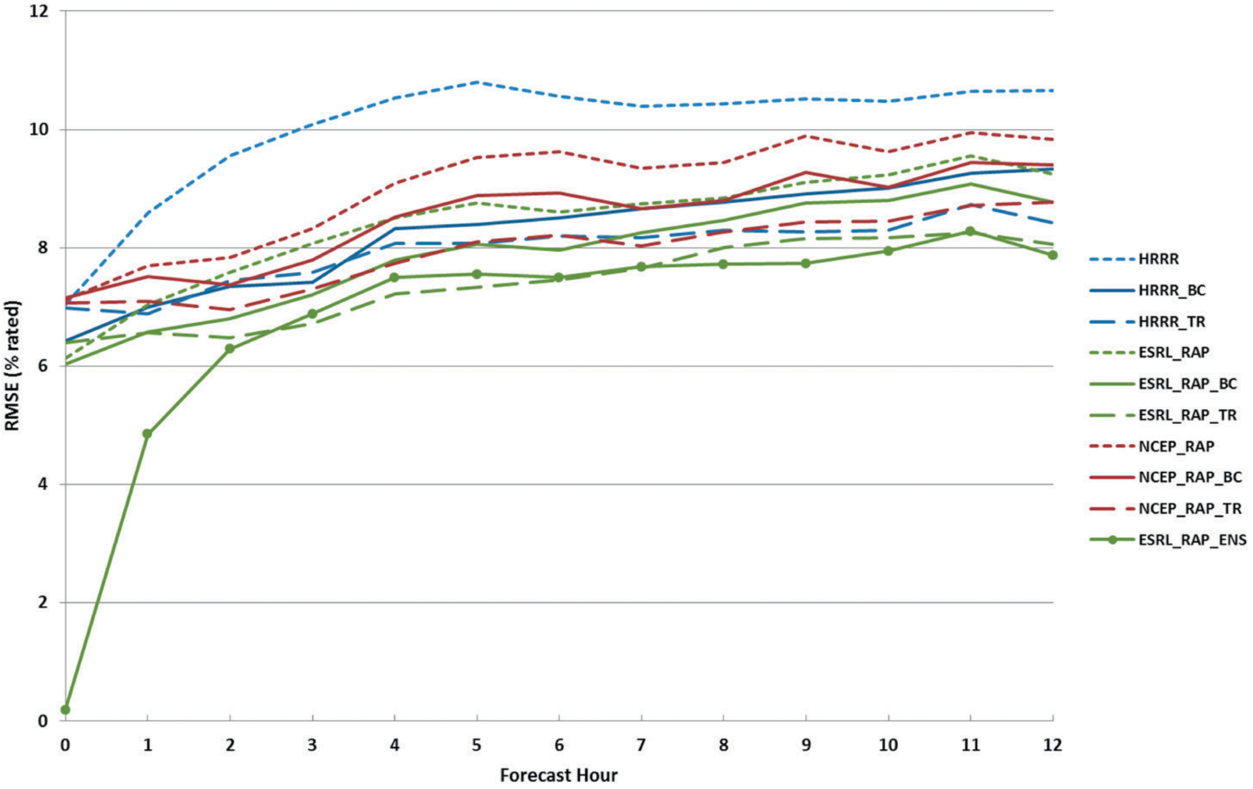

Fig. 8. RMSE (as a percentage of rated capacity) as a function of forecast hour for the system-wide aggregate raw, bias-corrected (BC), trained (TR), and trained ensemble (ENS, a combination of the ESRL_RAP-based and NAM-based power forecasts and the wind plant persistence forecast) over the time period for which trained forecasts were generated (Jan–Aug 2012). The forecast error is near zero at the start of the trained ensemble forecasts because the plant observations at forecast hour 0 are included in the ensemble forecasts.

detection is defined as the fraction of observed ramp events that is predicted correctly, and a value of 1 indicates a perfect forecast. The ESRL_RAP-based forecasts more accurately predict ramp events than do the HRRR and NCEP_RAP-based forecasts for most ramp definitions, by about 1% and 4%, respectively. A similar analysis done at the wind plant level showed that the HRRR-based forecasts were the most accurate, followed by those of ESRL_RAP and then NCEP_RAP (not shown).

Ramp-rate errors for the aggregate bias-corrected forecasts for all events classified as hits (i.e., correctly predicted ramps) are shown in Fig. 10. Negative values indicate that the forecast ramp rate is smaller than observed. As can be seen in Fig. 10, all forecasts underpredict the ramp rate for all ramp definitions, but the HRRR-based forecasts have significantly smaller ramp-rate errors (50%–60% less for most ramp definitions) than do the coarser-resolution model-based forecasts. This is largely because the HRRR-based forecasts significantly outperformed the coarser-resolution forecasts in accurately forecasting ramp duration. There is a clear demarcation in forecast ramp-rate errors as a function of ramp definition, with all forecasts more accurately predicting ramp rate for the smaller-magnitude (15% rated) ramp events. For these events, the HRRR-based forecasts are extremely accurate at correctly predicting the average ramp rate, with errors as much as 80%–90% smaller than the coarser-resolution forecasts. Similar results were obtained when the ramp-rate errors were calculated at the individual wind plants (not shown).

SOUTHERN STUDY AREA. For WFIP, a major component of AWS Truepower's contribution in the SSA region was the continued development and evaluation of ensemble forecast systems for wind energy. The WFIP SSA forecasting system (WFIPFS) is an enhanced and expanded version of AWST's operational eWind forecast system with five core components: 1) an ensemble of rapid-update short-term NWP forecasts, 2) a statistical adjustment procedure for each of the NWP forecasts, 3) a set of statistical time series prediction schemes, 4) an ensemble composite weighting algorithm, and 5) a wind plant output model. The WFIP enhanced observations were incorporated into most of the model system's data assimilation schemes. Actual wind plant power production from 34 plants was used for model postprocessing and validation.

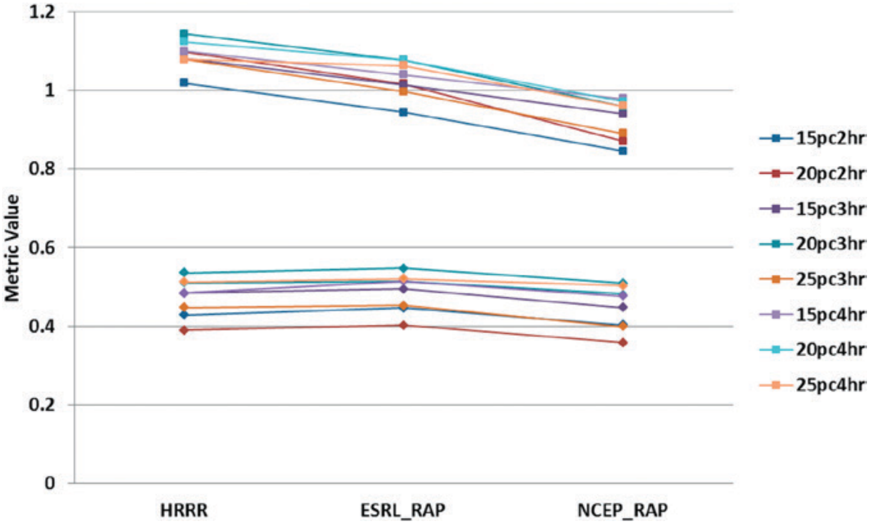

Fig. 9. Frequency bias (squares) and probability of detection (diamonds) for all wind power ramp events during the entire forecast period (Sep 2011–Aug 2012) for the aggregate bias-corrected HRRR, ESRL_RAP, and NCEP_RAP-based power forecasts. Results are shown for eight different ramp definitions as described in the text, where *XpcYhr* indicates an *X* change in rated capacity over a *Y*-h period.

Prior to WFIP, ERCOT was using two AWST forecast products: an optimized short-term wind power forecast (OPTENS_STWPF), which provided overall power production forecasts, and the ERCOT Large Ramp Alert System (ELRAS), which was used for ramp forecasts. These two were used as the baseline against which the WFIP forecasts were compared. ELRAS was primarily used to alert system operators that a major generation source of wind may become unavailable in a short amount of time (e.g., due to a forecast ramp event). If a large ramp is forecast, operators could subjectively decide to dispatch other available resources to meet whatever the load demand might be during their reliability update process.

The NWP component of the WFIPFS is composed of nine individual members based on three different modeling systems run by AWST, as well as HRRR. Most of the models have similar grid configurations. However, given the scale and nature of the phenomena affecting wind forecasts in the boundary layer (i.e., low-level jets, convective outflow boundaries, frontal systems), the following attributes were varied among the ensemble members:

- NWP models used to generate the simulations,
- source of lateral boundary conditions (BCs),
- boundary layer physics scheme,
- convective cloud scheme, and
- data assimilation scheme and incorporation of enhanced observations used to initialize the models.

The three models used by AWST were 1) WRF version 3.3.1 (Skamarock et al. 2005), 2) the Advanced Regional Prediction System (ARPS version 5.2.11) model (Xue et al. 2000, 2001), and 3) the Mesoscale Atmospheric Simulation System (MASS) model (Manobianco et al. 1996). All simulations had a horizontal grid spacing of 5 km and an update frequency of 2 h with most simulations using initial conditions (ICs) and BCs from ESRL_RAP. Three of the ensemble members were warm started (no spinup) using the previous forecast to initialize conditions for 11 of 12 runs per day, and a low-resolution (15 km) ensemble Kalman filter (EnKF) member was used to produce ICs and BCs for two of the ensemble members. The nine AWST high-resolution models produced 13-h power forecasts every 2 h for each wind plant for 1 yr, as well as a system-wide aggregate.

The baseline OPTENS_STWFP used two 8-km MASS runs employing NOAA GFS and NAM ICs and BCs, weighted each of three postprocessing methods [unadjusted, persistence adjusted, and model output statistics (MOS) adjusted] based on the relative performance over the previous month, and incorporated the latest 15-min power data for persistence corrections. An optimized WFIP ensemble forecast (OPTENS_WFIP) was similar except that it used forecasts from each of the 10 different NWP models (9 AWST plus HRRR) that used the ESRL_RAP ICs and BCs. The NWP forecasts were available for bias correction 2 h after initialization. A bias-corrected forecast was delivered every 15 min. The most recent wind plant tower and power generation data were used in the bias correction algorithm to include information for "persistence."

Comparisons of the WFIP ensemble forecasts (OPTENS_WFIP) with the baseline (OPTENS_STWPF)

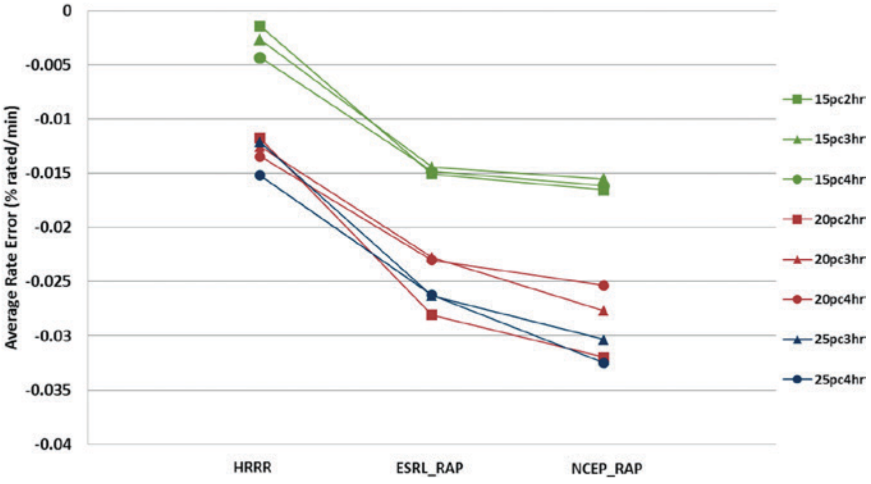

Fig. 10. Average ramp-rate errors for all ramp events classified as hits for the aggregate ramp events during the entire forecast period (Sep 2011–Aug 2012) for the bias-corrected HRRR-, ESRL_RAP-, and NCEP_RAP-based power forecasts. Results are shown for eight different ramp definitions as defined in the text.

are shown in Fig. 11. The greatest error reduction (30%–60%) occurs in the first 90 min of the forecast. Beyond 90 min, the forecast error reduction steadily decreases but is still apparent. The initial improvement before 90 min can be attributed to several factors, including more accurate (and a larger ensemble of) higher-resolution WFIP models, and the use of the ESRL_RAP ICs and BCs. Data-denial experiments (not shown) indicate lesser contributions to the magnitude of the overall forecast improvement at longer forecast time horizons from assimilated project observations. Additional sensitivity experiments are required to determine which component of the WFIPFS contributed most to forecast improvement. Although an ensemble forecast should outperform a deterministic forecast, and a large ensemble should outperform a small ensemble, the results here demonstrate the magnitude of these improvements when applied to wind energy forecasting.

One way to graphically summarize the improved ensemble performance (and compare the individual ensemble members) is through a Taylor diagram (Taylor 2001). The similarity between the ensemble performance and the observations is quantified in terms of their correlation (or their coefficient of determination), their centered root-mean-square difference, and the amplitude of their variations (represented by their standard deviations). Thus, Taylor diagrams can be especially useful in evaluating multiple aspects of model performance in a phase–amplitude space.

Figure 12 represents a Taylor diagram showing the individual ensemble members (various

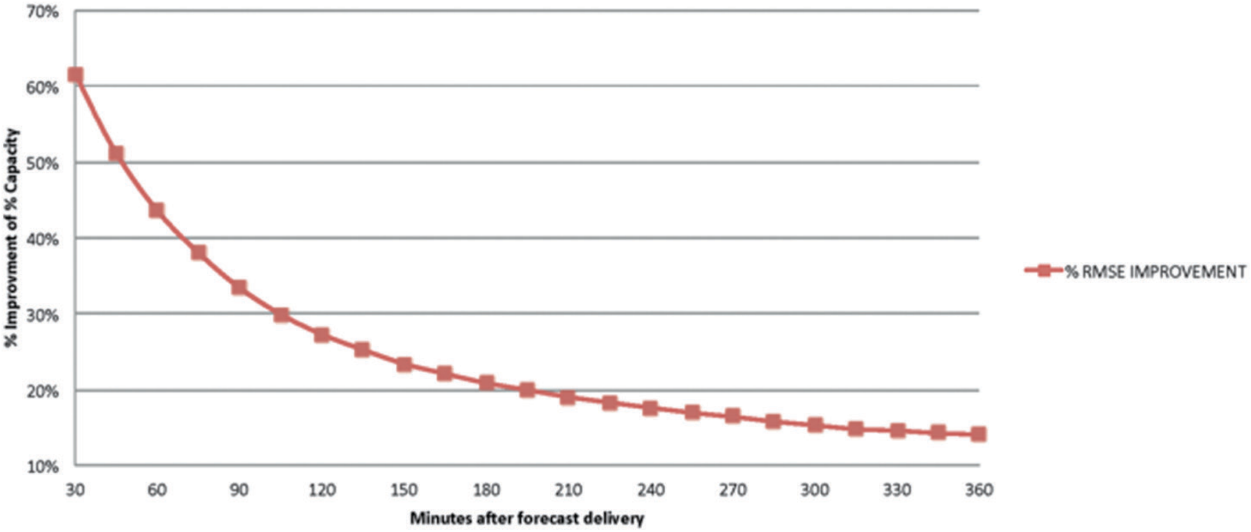

Fig. 11. Comparison of the percentage improvement in forecast performance as a function of forecast look-ahead time using the WFIP ensemble forecast system (OPTENS_WFIP) vs the baseline (OPTENS_STWPF) forecast, evaluated using wind-plant-observed aggregate power. The values represent the percentage reduction in RMSE of the OPTENS_WFIP over that of the OPTENS_STWPF forecast.

symbols), the OPTENS_STWPF (open upside-down black triangle), and OPTENS_WFIP (depicted by the open green triangle) 3-h forecast performance for the WFIP system-wide aggregate as compared with observations (black asterisk). Note that the individual model members (unoptimized, therefore requiring no statistical postprocessing) show considerable scatter in the phase–amplitude space, with the MASS and HRRR members performing best. There is also a significant increase in overall skill shown by the OPTENS_WFIP as compared with the baseline OPTENS_STWPF, with definitive movement toward minimizing RMSE, increasing r^2, and capturing the characteristic observational variability (solid blue arrow in Fig. 12). Note there is still significant room for improvement, as denoted by the dotted red arrow.

Next, the use of the WFIP ensemble for forecasting ramp events is investigated. Single forecasts from deterministic models cannot communicate the likelihood of occurrence or likelihood of different ramp event scenarios. Therefore, 6-h probabilistic ramp event forecasts were created every 15 min. For individual models these were derived using quantile regression of historical forecasts. These forecasts contain the probability of exceedance for several ramp-rate thresholds and a probability distribution of ramp rates. The WFIP probabilistic ramp forecasts were compared to ELRAS,

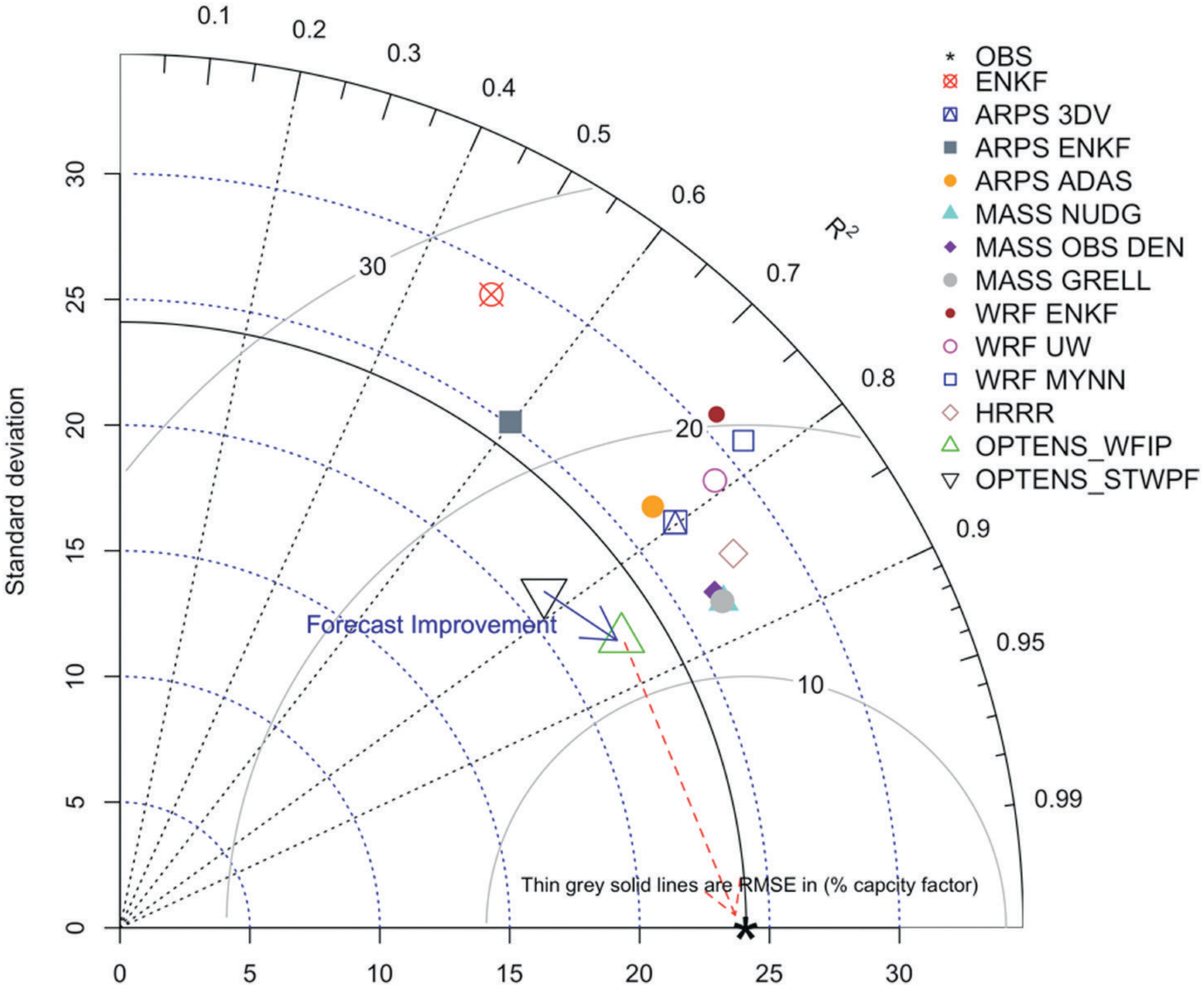

Fig. 12. Taylor diagram showing individual ensemble members (various colored symbols), OPTENS_STWPF (open upside-down black triangle), and OPTENS_WFIP optimized ensemble (open green triangle) 3-h forecast performance as compared with the observations (black asterisk). Thin gray solid lines represent the centered RMSE, black dotted lines depict the coefficient of determination, and blue dotted lines show the standard deviation. The solid blue arrow shows the forecast improvement from OPTENS_STWFP to the OPTENS_WFIP systems; dotted red arrow shows the trajectory to a perfect forecast.

LOW-LEVEL JETS IN THE SOUTHERN STUDY AREA

The low-level jet (LLJ) is a phenomenon that has been investigated by wind energy interests for nearly 40 years (Sisterson and Frenzen 1978; Kelly et al. 2004; Banta et al. 2008). LLJs occur regularly throughout the year in the southern Great Plains (e.g., Bonner 1968) and are especially prevalent over the WFIP SSA (Freedman et al. 2008). The height of the LLJ wind speed maximum varies between 50 and 400 m, but typically occurs at about 200 m (Banta et al. 2002). Thus, a special concern for wind energy interests and a forecasting challenge introduced by LLJs is the large vertical shear [upward of 8 m s^{-1} (100 m)$^{-1}$] that can occur across the turbine rotor plane.

Critical observational and forecasting issues concerning LLJs are 1) the strength of the vertical gradient of wind speed, 2) their formation and persistence, 3) spatial characteristics such as width and depth, and 4) intermittent turbulence leading to propagation of strong winds toward the surface. To ensure sufficient vertical resolution of the full profile of the LLJ, the field measurement campaign included the deployment of several integrated observation sites: that is, the collocation of a surface meteorological station, sodar, and wind profiling radar.

Qualitative analysis indicates that the LLJ is a regular, periodic (e.g., Fig. SB1), and dominant feature in the SSA that drives capacity factors to over 60% (and therefore a large fraction of power production) during the nocturnal hours. Given the large wind shears that can be generated by LLJs, model forecast errors that displace the wind profile by just a few tens of meters can lead to large errors in forecast power production, as can mistiming their onset or cessation.

OPTENS_WFIP produced a marked improvement in forecasting the amplitude and phase of the LLJ, as demonstrated in Fig. SB1 (showing the 3-h forecasts for a several-day sequence dominated by the formation and decay of the LLJ). In particular, the individual model-member raw forecasts were often significantly in error regarding the amplitude and phase of the diurnal wind speed cycle, while the OPTENS_WFIP forecasts (solid blue line in Fig. SB1) were more skillful. This is consistent with the findings of Deppe et al. (2013) and highlights a continuing issue that models have in capturing the temporal and spatial distributions of LLJs (Storm et al. 2009; Storm and Basu 2010). Accurate LLJ depiction is crucial for wind energy forecasts, and the results here illustrate the limitations of model parameterization schemes, especially for the PBL (Deppe et al. 2013; Werth et al. 2011). This further demonstrates the critical role played by the WFIP model system statistical postprocessing and bias correction schemes, which show much better alignment with the observations (solid black line in Fig. SB1). The large spread in individual model-member forecasts (Figs. 12 and SB1) is striking and suggests the value of probabilistic uncertainty forecasts for LLJ scenarios, a subject deserving of additional study but beyond the scope of this paper.

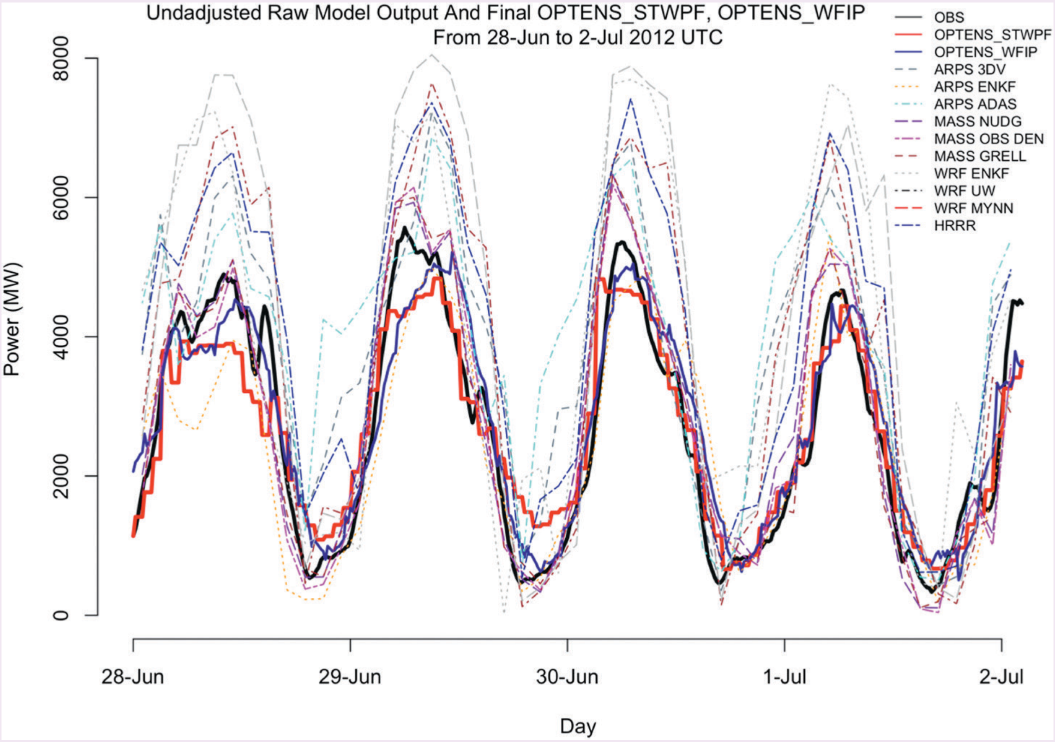

Fig. SB1. Observations (thick black line), OPTENS_STWFP baseline 3-h forecast (thick red line), OPTENS_WFIP ensemble 3-h forecast (thick blue line), and unadjusted individual model member 3-h forecasts (various line types and colors) for 28 Jun–2 Jul 2012.

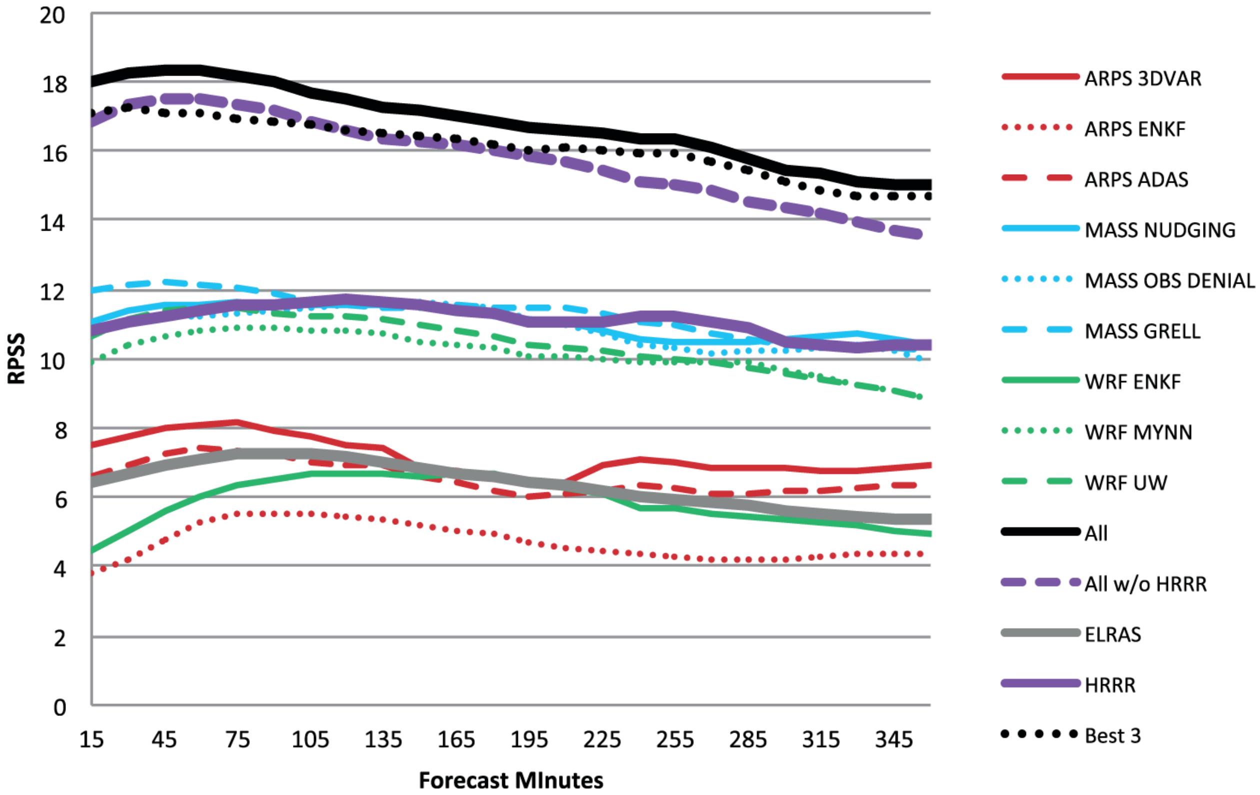

Fig. 13. RPSS for 60-min ramps for all individual WFIP member ramp forecasts, and ramp forecasts generated from a combination of WFIP forecast members (All, Best 3, All without HRRR).

which was based on a single ARPS model run and 3D variational data assimilation (3DVAR; Zack et al. 2011).

The probabilistic ramp forecasts were verified using the rank probability skill score (RPSS; Murphy 1969). The RPSS represents the improvement of the ramp probability forecast over the climatological ramp probabilities, and a value greater than zero indicates forecast skill greater than the reference (climatological) forecast. Several ensemble forecasts were generated using combinations of all ("All") and the best ARPS, WRF, and MASS members (i.e., "Best 3") from the MOS method. The results (Fig. 13) show that the ensembles produced a more accurate probabilistic forecast of 60-min ramps than any one of the single members. There is, on average, a 20% improvement in RPSS (forecast skill) of the ensemble forecasts (Best 3, All) over the best-performing single-member methods (i.e., HRRR, MASS). This result highlights the additional value from using an ensemble to generate a probabilistic ramp forecast. Of the single members, the HRRR probabilistic ramp forecast performed the best followed by the MASS and WRF [Mellor–Yamada–Nakanishi–Niino (MYNN, University of Wisconsin (UW)] forecast members. The ARPS [3DVAR, the ARPS Data Analysis System (ADAS)] and EnKF members (ARPS, WRF) performed poorly, mostly because of higher false-alarm rates. The baseline ELRAS also performed poorly when compared to the other WFIP members. The inclusion of all these members in the ensemble outperformed the Best 3 ensemble, highlighting the advantage of model diversity.

CONCLUSIONS. WFIP allowed for NOAA, DOE national laboratories, private-sector forecasting companies, universities, and electric grid operators to collaborate on improving wind energy forecasts. A significant number of new observing systems, including proprietary tall-tower and turbine nacelle anemometer measurements provided by the private sector, were assimilated into NOAA's regional and hourly updated forecast models, in both real-time and retrospective simulations. Data-denial experiments demonstrated that assimilation of these observations led to statistically significant improvements in turbine-height power forecasts. Improvements in the forecasts also occurred with the transition from the RUC to the RAP model midway through the WFIP field campaign. The improvements from the combination of assimilation of the additional observations and the upgrade to the RAP model ranged from 12%

to 5% for forecast hours 1–12 in the NSA and SSA, equivalent to approximately the previous decade's improvements achieved for NOAA/NWS operational forecasts for 850-hPa winds over North America.

The results from the NSA demonstrate that the research models (ESRL_RAP, HRRR) that included the additional WFIP data assimilation and improved model physics produced more accurate wind power forecasts than those created using the current operational model (NCEP_RAP), as illustrated in the general bulk error statistics and wind energy ramp forecast performance. A comparison of the various forecast metrics calculated with the two research models indicated that for overall bulk statistics, the 13-km-resolution ERSL_RAP-based forecasts are more accurate, while for ramp rates the 3-km HRRR-based forecasts performed best. These analyses collectively show that when it comes to power system operations, the complexity in identifying the best weather model for a particular forecast need is reason to have a diverse choice of models.

For the SSA, the AWST forecasts (OPTENS_WFIP) demonstrated impressive improvement in forecast power production compared with the baseline (OPTENS_STWPFS), with the largest improvement (60%) in aggregate capacity factor RMSE at hour 1, and consistently better performance (>20% decrease in RMSE) through hour 3. The probabilistic WFIP ensemble ramp predictions resulted in a large (20% or more) improvement in the RPSS as compared with the baseline (ELRAS) forecasts. Finally, the enhanced field observations facilitated identification and analysis of the principal phenomena (LLJs) responsible for the winds generating the larger capacity factors notable in the ERCOT domain and were also key to model system performance in capturing the phase and amplitude of the diurnal wind speed cycle.

Although the WFIP analyses have shown that significant improvements were made in wind energy forecasts, including wind ramp events, the remaining step of quantifying the economic benefits of these forecast improvements is still in process. One of the challenges of that analysis is in associating a monetary benefit to the improved grid reliability that is achieved through better short-term wind energy forecasts. Future efforts coordinated across DOE, NOAA, and the private sector are also needed to continue improvements to model forecast systems and data assimilation methods to optimally utilize additional observations, to investigate forecast uncertainty, and to develop techniques required in more extreme complex-terrain environments.

ACKNOWLEDGMENTS. Numerous individuals made significant contributions to the success of WFIP. These include Tim Martin, Clark King, Jesse Leach, Tom Ayers, Jim Jordan, Dan Gottas, David Welsh, Leon Benjamin, Matthew Filippelli, Kurt Elsholz, Vic Morris, and Dan Nelson (field deployment and data acquisition); John Schroeder, Tom Strong, Shane Beard, Dave Christensen, Dennis Finn, Roger Carter, Brad Reese, Robert Lipschutz, Jason Rich, and Charles Kovalsky (data communications); Philippe Beaucage, Katherine Rojowsky, Sukanta Basu, and Paul Svenson (data analysis); Jennifer Leise, Deborah Hanley, and Ken Pennock (project management); Carrie Gillespie, Ken DeRose, Dennis Todey, and Bob Conzemius (real-time meteorological tower and turbine nacelle data); Victor Yannuzzi and Francisco Guzman (forecasting support); Dennis Keyser, Steven Levine, Jeff Whiting, and Ming Hu (GSI and data assimilation); Eric Rogers and Curtis Alexander (model development); Richard Eckman and Elizabeth Weatherhead (statistical analysis); Stan Calvert, Will Shaw, Geoff DiMego, John Brown, and Allen White (program support); Brian Ancell, Keith Brewster, Kevin Thomas, and Steve Young (AWST modeling systems); Venkat Banunarayanan, Saleh Nasir, Kristen Orwig, Greg Brinkman, Greg Stark, Erik Ela, and Jie Zhang (ongoing economic analysis); Isabel Flores (ERCOT liaison); and Michael McMullen (MISO liaison). The authors gratefully thank three anonymous reviewers for many constructive comments and suggestions. This work was supported by the U.S. Department of Energy (DOE), Office of Energy Efficiency and Renewable Energy, and by the National Oceanic and Atmospheric Administration.

REFERENCES

Banta, R. M., R. K. Newsom, J. K. Lundquist, Y. L. Pichugina, R. L. Coulter, and L. Mahrt, 2002: Nocturnal low-level jet characteristics over Kansas during CASES-99. *Bound.-Layer Meteor.,* **105,** 221–252, doi:10.1023/A:1019992330866.

——, Y. L. Pichugina, N. D. Kelley, B. Jonkman, and W. A. Brewer, 2008: Doppler lidar measurements of the Great Plains low-level jet: Applications to wind energy. *IOP Conf. Ser.: Earth Environ. Sci.,* **1,** 012020, doi:10.1088/1755-1315/1/1/012020.

Benjamin, S. G., B. E. Schwartz, E. J. Szoke, and S. E. Koch, 2004: The value of wind profiler data in U.S. weather forecasting. *Bull. Amer. Meteor. Soc.,* **85,** 1871–1886, doi:10.1175/BAMS-85-12-1871.

——, B. D. Jamison, W. R. Moninger, S. R. Sahm, B. Schwartz, and 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, doi:10.1175/2009MWR3097.1.

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. Oceanic Technol.,* **30,** 2697–2703, doi:10.1175/JTECH-D-13-00089.1.

Bonner, W. D., 1968: Climatology of the low level jet. *Mon. Wea. Rev.,* **96,** 833–850, doi:10.1175/1520-0493(1968)096<0833:COTLLJ>2.0.CO;2.

Bossanyi, E. A., 1985: Short-term wind prediction using Kalman filters. *Wind Eng.,* **9,** 1–8.

Chang, C.-C., and C.-J. Lin, 2011: LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.,* **2,** Article 27, doi:10.1145/1961189.1961199.

Cortes, C., and V. Vapnik, 1995: Support-vector networks. *Mach. Learn.,* **20,** 273–297.

Deppe, A. J., W. A. Gallus Jr., and E. S. Takle, 2013: A WFR ensemble for improved wind speed forecast at turbine height. *Wea. Forecasting,* **28,** 212–228, doi:10.1175/WAF-D-11-00112.1.

EnerNex, 2011: Eastern wind integration and transmission study. Subcontract Rep. NREL/SR-550-47078, 242 pp. [Available online at www.nrel.gov/docs/fy11osti/47078.pdf.]

Finley, C., M. Ahlstrom, L. Sheridan, G. Brinkman, K. Orwig, G. Stark, D. Todey, and M. McMullen, 2014: The Wind Forecast Improvement Project (WFIP): A public/private partnership for improving short term wind energy forecasts and quantifying the benefits of utility operations—The northern study area. WindLogics Final Tech. Rep. to DOE, Award DE-EE0004421, 125 pp. [Available online at www.osti.gov/scitech/biblio/1129929.]

Freedman, J., M. Markus, and R. Penc, 2008: Analysis of West Texas wind plant ramp-up and ramp-down events. Analysis of wind generation impact on ERCOT ancillary services requirements, R. A Walling, AWS Truewind Rep., 250–278. [Available online at www.uwig.org/AttchB-ERCOT_A-S_Study_Final_Report.pdf.]

——, and Coauthors, 2014: The Wind Forecast Improvement Project (WFIP): A public/private partnership for improving short term wind energy forecasts and quantifying the benefits of utility operations—The southern study area. AWS Truepower Final Tech. Rep. to DOE, Award DE-EE0004420, 107 pp. [Available online at www.osti.gov/scitech/biblio/1129905.]

GE Energy, 2010: Western wind and solar integration study. Prepared for NREL Subcontract Rep. SR-550-47781, 536 pp. [Available online at www.nrel.gov/docs/fy10osti/47434.pdf.]

Giebel, G., and G. Kariniotakis, 2007: Best practice in short-term forecasting—A users guide. E*uropean Wind Energy Conf. and Exhibition,* Milan, Italy, European Wind Energy Association, BT2.2. [Available online at www.ewea.org/ewec2007/allfiles2/156_Ewec2007fullpaper.pdf.]

IEC, 2005: Wind turbines—Part 1: Design requirements. International Electrotechnical Commission IEC 61400-1, 179 pp. [Available online at http://webstore.iec.ch/preview/info_iec61400-1%7Bed3.0%7Den.pdf.]

Kelly, N. D., M. Shirazi, D. Jager, S. Wilde, J. Adams, M. Buhl, P. Sullivan, and E. Patton, 2004: Lamar low-level jet program interim report. NREL/TP-500-34593. [Available online at www.nrel.gov/docs/fy04osti/34593.pdf.]

Mahoney, W. P., and Coauthors, 2012: A wind power forecasting system to optimize grid integration. *IEEE Trans. Sustainable Energy,* **3,** 670–682, doi:10.1109/TSTE.2012.2201758.

Makarov, Y. V., Z. Huang, P. V. Etingov, J. Ma, R. T. Guttromson, K. Subbarao, and B. B. Chakrabarti, 2010: Incorporating wind generation and load forecast uncertainties into power grid operations. Pacific Northwest National Laboratory Rep. PNNL-19189, 169 pp. [Available online at www.pnl.gov/main/publications/external/technical_reports/PNNL-19189.pdf.]

Manobianco, J., J. W. Zack, and G. E. Taylor, 1996: Workstation-based real-time mesoscale modeling designed for weather support to operations at the Kennedy Space Center and Cape Canaveral Air Station. *Bull. Amer. Meteor. Soc.,* **77,** 653–672, doi:10.1175/1520-0477(1996)077<0653:WBRTMM>2.0.CO;2.

Marquis, M., J. Wilczak, M. Ahlstrom, J. Sharp, A. Stern, J. C. Smith, and S. Calvert, 2011: Forecasting the wind to reach significant penetration levels of wind energy. *Bull. Amer. Meteor. Soc.,* **92,** 1159–1171, doi:10.1175/2011BAMS3033.1.

Monteiro, C., R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, 2009: Wind power forecasting: State-of-the-art 2009. Argonne National Laboratory Rep. ANL/DIS-10-1, 198 pp. [Available online at www.dis.anl.gov/pubs/65613.pdf.]

Sisterson, D. L., and P. Frenzen, 1978: Nocturnal boundary-layer wind maxima and the problem of wind power assessment. *Environ. Sci. Technol.,* **12,** 218–221, doi:10.1021/es60138a014.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech NCAR/TN–468+STR, 88 pp.

[Available online at http://www2.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf.]

——, and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN–475+STR, 113 pp. [Available online at http://www2.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.]

Storm, B., and S. Basu, 2010: The WRF model forecast-derived low-level wind shear climatology over the United States Great Plains. *Energies,* **3,** 258–276, doi:10.3390/en3020258.

——, J. Dudhia, S. Basu, A. Swift, and I. Giammanco, 2009: Evaluation of the Weather Research and Forecasting Model on forecasting low-level jets: Implications for wind energy. *Wind Energy,* **12,** 81–90, doi:10.1002/we.288.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106,** 7183–7192, doi:10.1029/2000JD900719.

Wendell, L. L., H. L. Wegley, and M. G. Verholek, 1978: Report from a working group meeting on wind forecasts for WECS operation. Pacific National Laboratory Rep. PNL-2513, 14 pp. + appendixes, doi:10.2172/6548011.

Werth, D., R. Kurzeja, N. Dias, G. Zhang, H. Duarte, M. Fischer, M. Parker, and M. Leclerc, 2011: The simulation of the southern Great Plains nocturnal boundary layer and the low-level jet with a high-resolution mesoscale atmospheric model. *J. Appl. Meteor. Climatol.,* **50,** 1497–1513, doi:10.1175/2011JAMC2272.1.

Wilczak, J. M., and Coauthors, 1995: Contamination of wind profiler data by migrating birds: Characteristics of corrupted data and potential solutions. *J. Atmos. Oceanic Technol.,* **12,** 449–467, doi:10.1175/1520-0426(1995)012<0449:COWPDB>2.0.CO;2.

——, L. Bianco, J. Olson, I. Djalalova, J. Carley, S. Benjamin, and M. Marquis, 2014: The Wind Forecast Improvement Project (WFIP): A public/private partnership for improving short term wind energy forecasts and quantifying the benefits of utility operations. NOAA Final Tech. Rep. to DOE, Award DE-EE0003080, 162 pp. [Available online at http://energy.gov/sites/prod/files/2014/05/f15/wfipandnoaafinalreport.pdf.]

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, doi:10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2.

Xue, M., K. K. Droegemeier, and V. Wong, 2000: The Advanced Regional Prediction System (ARPS)—A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part I: Model dynamics and verification. *Meteor. Atmos. Phys.,* **75,** 161–193, doi:10.1007/s007030070003.

——, and Coauthors, 2001: The Advanced Regional Prediction System (ARPS)—A multi-scale nonhydrostatic atmospheric simulation and prediction model. Part II: Model physics and applications. *Meteor. Atmos. Phys.,* **76,** 143–165, doi:10.1007/s007030170027.

Zack, J. W., S. H. Young, and E. J. Natenberg, 2011: Evaluation of wind ramp forecasts from an initial version of a rapid update dynamical-statistical ramp prediction system. *Proc. Second Conf. on Weather, Climate, and the New Energy Economy,* Seattle WA, Amer. Meteor. Soc., 781. [Available online at https://ams.confex.com/ams/91Annual/webprogram/Paper186686.html.]