Received: 21 May 2018 | Revised: 15 March 2019 | Accepted: 24 March 2019

DOI: 10.1002/we.2347

RESEARCH ARTICLE

WILEY

Measuring the impact of additional instrumentation on the skill of numerical weather prediction models at forecasting wind ramp events during the first Wind Forecast Improvement Project (WFIP)

Elena Akish[1,2,3] | Laura Bianco[2,3] | Irina V. Djalalova[2,3] | James M. Wilczak[2] | Joseph B. Olson[2,3] | Jeff Freedman[4] | Catherine Finley[5] | Joel Cline[6]

[1] Science and Technology Corporation (STC), Boulder, CO

[2] NOAA/Earth Systems Research Laboratory, Boulder, CO

[3] University of Colorado/CIRES, Boulder, CO

[4] AWS Truepower, Albany, NY

[5] WindLogics Inc., St. Paul, MN

[6] DOE/Energy Efficiency and Renewable Energy, Washington, D.C.

Correspondence
Elena Akish, 325 Broadway, mail stop: PSD, Boulder, CO 80305, USA.
Email: elena.akish@noaa.gov

Present Address
Jeff Freedman, Atmospheric Sciences Research Center, University of Albany, State University of New York, Albany, NY.

Catherine Finley, Department of Earth and Atmospheric Sciences, Saint Louis University, St. Louis, MO.

Joel Cline, NOAA/National Weather Service, Washington D.C.

Funding information
National Oceanic and Atmospheric Administration; US Department of Energy (Office of Science, Office of Basic Energy Sciences and Energy Efficiency and Renewable Energy, Solar Energy Technology Program), Grant/Award Number: DE-EE0003080

Abstract

The first Wind Forecast Improvement Project (WFIP) was a DOE and NOAA-funded 2-year-long observational, data assimilation, and modeling study with a 1-year-long field campaign aimed at demonstrating improvements in the accuracy of wind forecasts generated by the assimilation of additional observations for wind energy applications. In this paper, we present the results of applying a Ramp Tool and Metric (RT&M), developed during WFIP, to measure the skill of the 13-km grid spacing National Oceanic and Atmospheric Administration/Earth System Research Laboratory (NOAA/ESRL) Rapid Refresh (RAP) model at forecasting wind ramp events. To measure the impact on model skill generated by the additional observations, controlled data-denial RAP simulations were run for six separate 7 to 12-day periods (for a total of 55 days) over different seasons.

The RT&M identifies ramp events in the time series of observed and forecast power, matches in time each forecast ramp event with the most appropriate observed ramp event, and computes the skill score of the forecast model penalizing both timing and amplitude errors. Because no unique definition of a ramp event exists (in terms of a single threshold of change in power over a single time duration), the RT&M computes integrated skill over a range of power change (Δp) and time period (Δt) values.

A statistically significant improvement of the ramp event forecast skill is found through the assimilation of the special WFIP data in two different study areas, and variations in model skill between up-ramp versus down-ramp events are found.

KEYWORDS

data assimilation, forecasting, ramp events

1 | INTRODUCTION

Grid operators responsible for making decisions on what kind of power generation to use to keep the grid in balance (conventional versus weather dependent, such as wind or solar) need a reliable numerical weather prediction (NWP) model to ensure grid stability. While conventional power is nearly always available, weather-dependent power can vary greatly over short periods of time. Wind speed variability is furthermore amplified through the wind turbine's power curve, which translates the wind speed into power production. Wind power production can feature large excursions, known as ramp events that can be very rapid as wind power is proportional to the cube of the wind speed in the middle portion of the turbine's power curve. Ramp events can be a challenge for grid operators as they must continuously keep power production in nearly exact balance with power demand. If large, sudden, and unanticipated changes in wind power production occur, the grid operator may be forced to bring online (or "spin up") other conventional energy generation units. Ramps that are not forecasted accurately (both in terms of amplitude and timing) will require large and sudden changes in output from conventional generation units, which can ultimately increase the costs of power generation and diminish the appeal of weather-dependent resources. However, evaluating the exact economic savings from improved short-term forecasts is challenging due to the complicated issue of applying a monetary value to grid reliability. In one of the few studies to estimate savings from short-term forecasts, Hodge et al[1] evaluated the cost savings from improving ultra-short-term forecasts (less than 40 minutes) on the California Independent System Operator (CAISO) system and found that for a future scenario with 25% wind energy penetration, a 50% forecast skill improvement resulted in an annual savings of $146 million.

To measure the skill of NWP models at forecasting ramp events, we developed a Ramp Tool and Metric (RT&M).[1] The RT&M is publicly available at https://www.esrl.noaa.gov/psd/products/ramp_tool/.

A description of the RT&M can be found in Bianco et al,[2] in which the tool was tested on data collected over a period of 9 days from a set of four 80-meter-tall towers and forecasts at the same locations from the 13-km grid spacing National Oceanic and Atmospheric Administration/Earth System Research Laboratory (NOAA/ESRL) Rapid Refresh Numerical (RAP) model, during the first WFIP. WFIP took place in the United States Great Plains from September 2011 to September 2012.[3-6] While the primary goal of WFIP was to test the impact of additional observations on the forecast skill of turbine-height winds, in this study, we focus on the skill of forecasting ramp events.

This paper is organized as follows: Section 2 presents the dataset used in this study; Section 3 gives basic details on the RT&M; Section 4 presents the results of our analysis, separately for the two main geographical locations of the WFIP campaign, and then assessing what types of observations are most useful to the data assimilation system to improve the skill of the models at forecasting ramp events; finally, Section 5 provides a summary and discussion.

2 | DATASET

The WFIP 1-year-long field campaign took place in two high wind energy resource areas of the United States, the upper Great Plains (hereafter referred to as North Study Area, NSA) and Texas (hereafter referred to as South Study Area, SSA). The additional observations in the NSA consisted of nine wind profiling radars (WPRs, seven of which were 915 MHz, and two of which were 449 MHz), five sodars, and 134 instrumented tall towers; while for the SSA, the additional observations consisted of three 915-MHz WPRs, seven sodars, and 49 instrumented tall towers.[6] Hourly averaged winds and RASS temperatures were provided by the WPRs, while the sodars provided 10-minute averaged data, and the towers provided 10-minute averaged wind speeds and directions at a height between 55 and 90 m. All three data sets were visually quality controlled to eliminate outliers before assimilation.[6]

The primary goal of WFIP was to improve short-term forecasts, so NOAA's RAP model was used in this study, which also supports the short-term forecasting of severe weather events for NOAA's National Weather Service. The RAP has 13-km grid spacing and was re-initialized every hour and run out to 15 hours, with model output available every 15 minutes, allowing us to analyze rapidly evolving ramp events. The version of the RAP model used in WFIP employs the Advanced Research version of the Weather Research and Forecast model (WRF-ARW)[7] version 3.4.1 as the forecast component, and a 3D variational data assimilation technique with a Gridpoint Statistical Interpolation (GSI) analysis system[8] each hour assimilating virtually every meteorological observation that is available, including satellite, aircraft, radiosondes, surface mesonet, and WPRs, among others. So, while the number of new instruments deployed for WFIP was relatively small, they were concentrated in the NSA and SSA which comprised a small part of the overall model domain.[6]

To measure the impact of the additional observations on forecast skill, six data denial periods were selected (16-25 September 2011; 13-20 October 2011; 30 November-6 December 2011; 7-15 January 2012; 14-25 April 2012; 9-17 June 2012), ranging in length from 7 to 12 days, for a total of 55 days. These periods were selected to represent all four seasons and to contain a large number of high-amplitude ramp events.

For the data denial periods, RAP simulations were run first assimilating only the conventional observations (Control Runs); subsequent runs assimilated both the conventional observations and the additional WFIP observations (Experimental Runs). A previous analysis of these model simulations found that the relative percent improvement in model forecast skill of the Experimental Run's RMSE and coefficient of determination of turbine hub-height winds (at forecast hour 1) was 6%, for both the NSA and SSA, decreasing at longer forecast horizons.[6] A later, second analysis investigated the impact of different components of the WFIP observing system, by separating the instrumentation into two groups, remote sensors

(WPRs and sodars) and in situ (tall tower vector winds and nacelle anemometer wind speeds), and assimilating them each independently for a subset of two of the six data denial periods (13-20 October 2011 and 7-15 January 2012).[9] That analysis demonstrated that the large numbers of in situ observations had a significant initial impact that diminished rapidly after only several hours, while the less numerous remote sensing instruments had a smaller initial impact that improved the forecasts for a longer time, due to their observing a deeper layer of the atmosphere.

Here, we first use all six data denial runs, following the all or only standard observation assimilation approach, to measure the percentage improvement of the model skill at forecasting ramp events. Second, we assess the data assimilation impact of the remote sensing data alone, the in situ data alone, and for the two combined, using the 13 to 20 October 2011 and 7 to 15 January 2012 data denial periods. The data set used for verification of this analysis includes the measurements collected by the tall meteorological towers (55-90 m above ground level), as we are interested in the ramps that happen at or near turbine hub-height.

This selection leaves us with 97 towers in the North Study Area and 46 towers in the South Study Area. Only towers with more than 50% data availability for a particular data denial period are considered in the rest of the analysis.

3 | THE RAMP TOOL AND METRICS

The first step of the RT&M is to generate equal-length time series of model forecast and observational wind speed data. The RT&M provides two possible approaches: the "stitching method" and the "independent forecast run method." For the "stitching method," a time series of model forecasts is created for each particular forecast horizon, and each of these is evaluated versus the corresponding observational time series. For example, with an hourly updated model such as the RAP, the 6-hour forecasts generated each hour are concatenated into a longer time series of hourly values of all 6-hour forecasts. For the "independent forecast run method," the RT&M is applied to each independent forecast run over its full length. The first method allows for forecast skill to be derived as a function of forecast hour, and it assumes that no artificial ramps are generated by concatenating different model runs. The second method avoids the possibility of generating artificial ramps through combining different model runs but has the disadvantage that for each model forecast the beginning and ending forecast hours will suffer from truncated ramps that potentially begin before the start or end of the forecast cycle.[2] This disadvantage is reduced for the "stitching method" as the concatenated time series will be much longer and the impact of having truncated ramps at the beginning and end will be negligible. For this reason, we use the stitching method to prepare the time series of model forecasts for each forecast horizon. Nevertheless, we tested the use of the "independent forecast run method" on this data set and found consistent results with the "stitching method."

Model data, output at 15-minute intervals, are linearly interpolated to the tall tower 10-minute time intervals. Both time series are then converted into power using a standard International Electrotechnical Commission (IEC) class 2 turbine power curve.[9,10] Model data that correspond to periods of missing data in observations are disregarded in the analysis as well as observational data corresponding to periods of missing data in the model.

Although ramps are often referred to as "a large change in power production over a short interval of time," no commonly accepted definition of a ramp event exists, nor are strict thresholds defining the "large change in power" and the "short interval of time" possible for all applications. Different threshold values might be more appropriate for different situations and different users. For this reason, the RT&M allows one to use a matrix of possible ramp definitions, each for different changes in power (Δp) and over different intervals of time (Δt). In the standard setting, threshold values for Δp are chosen to be 30%, 40%, 50%, 60%, and 70% of the rated normalized capacity, while threshold values for Δt are chosen to be 30, 60, 120, and 180 minutes. Using these thresholds for Δp and Δt provides the possibility to investigate the behavior of a NWP model for 20 different ramp definitions. This standard setting can be changed according to the user needs.

For each of these 20 ramp definitions, the RT&M follows three basic steps[2]:

- First, it identifies ramp events in the time series of observed and modeled power data. Three different identification methods are used, the "Fixed Time Interval Method," the "Min-Max Method," and the "Explicit Derivative Method." While the "Fixed Time Interval Method" only measures the difference in power over a determined time window, the "Min-Max Method" takes into account the maximum amplitude change in power within that window, and the "Explicit Derivative Method" analyzes the value of a smoothed time derivative of the power over that time window. A detailed description of these methods is presented in Bianco et al.[2]
- Second, the RT&M matches in time the observed ramp events with those predicted by the forecast model.
- Finally, it scores the ability of the model to forecast ramp events. The scoring metric accounts for both amplitude and timing (center point and duration) errors in the forecast, allowing one to evaluate up-ramp and down-ramp events separately. The particular scoring rules that we used are intended to reflect the perspective of a grid operator; however, the metric itself is flexible, and it could be easily modified by any user to reflect the needs of other participants in the energy generation system. Each ramp event is assigned a score in the range from −1 to 1; a score of 1 is assigned to an event when the forecasted ramp is identical to the observed, −1 when the forecasted ramp's characteristics match the observed one's except for the type of the ramp (eg, an up-ramp is forecasted when a down-ramp occurred), and 0 when the event is unmatched. The score can assume any value in the range from −1 to 1 according to the difference in change of power, duration, and time

of occurrence between the forecasted ramp and the observed one. An average score is calculated for all events found for each ramp definition, and the final score is the average across all ramp definitions. For more information on the scoring procedure, see Bianco et al.[2]

After this process is repeated for all ramp definitions, the NWP model is assigned a matrix of scores of the same dimension as the matrix of ramp definitions. Finally, this matrix of scores will be averaged to provide a final value for the NWP model skill at forecasting ramp events. This process can be repeated for all the forecast hours, to obtain the skill of the NWP model as a function of the forecast horizon.

4 | RESULTS

4.1 | North Study Area

Results averaged over the six data denial periods are presented in Figure 1 for the NSA. Skill was calculated using the three available ramp identification methods ("Fixed Time Interval Method," "Min-Max Method," and the "Explicit Derivative Method"), and because the results were consistent between the methods, an average over the three identification methods is presented here and for the rest of our analysis. The upper panel of Figure 1 shows the skill as a function of the forecast hour. The red line is used for the Experimental Runs (with additional observations), and the blue line is used for the Control Runs (without additional observations). The middle panel of Figure 1 shows the percentage improvement [100 × (skill_Experimental_Run − skill_Control_Run) / skill_Control_Run] as a function of the forecast hour. The lower panel in Figure 1 shows the difference between the two lines of the upper panel; this difference is statistically significant through forecast hour 12 in the NSA at the 95% confidence level (shaded contour).

A positive improvement is generated by the additional observations to the skill of the model at forecasting ramp events (middle panel of Figure 1), staying positive for all forecast hours. Averaging the NSA results over the first nine forecast hours, for all six data denial periods, and for the three ramp identification methods, we get a 7% improvement generated by assimilating the additional WFIP observations, which is even higher than the 6% found at only forecast hour 1 in the bulk RMSE of power forecast.[6] This means that the additional information is at least as beneficial to the model at forecasting rapid changes in wind speed as it is for improving the general forecast of hub-height winds.

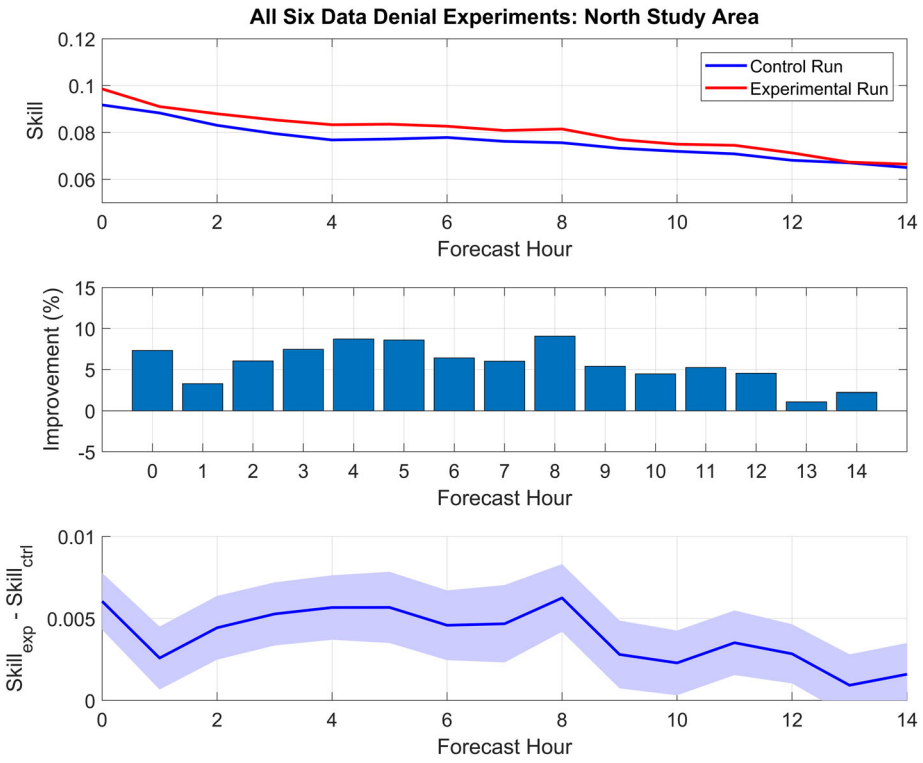

FIGURE 1 NSA results averaged over the six data denial periods and three ramp identification methods. Upper panel: Skill of the RAP model at forecasting ramp events as a function of the forecast hour (blue line for Control Runs and red line for Experimental Runs). Middle panel: Percentage improvement of the Experimental over the Control Run as a function of the forecast hour. Lower panel: difference in skill between the Control and Experimental Runs. Shaded area represents the 95% confidence interval defined as ($\pm 1.96\sigma/\sqrt{n}$) [Colour figure can be viewed at wileyonlinelibrary.com]

We also ran the RT&M looking separately at the skill of the RAP model at forecasting up-ramps and down-ramps. In both cases, the additional instrumentation improves the skill, as highlighted in Figure 2. Again, the skill is positive over all forecast hours, slightly decreasing with the longer forecast hours, but in both cases the improvement generated by the additional observations is evident in the lower panel of Figure 2. We also notice that both the skill of the Control and Experimental Runs are better at forecasting up-ramps (green lines in the upper panel of Figure 2, with larger values for the skill) versus down-ramps (purple lines in the upper panel of Figure 2, with smaller values for the skill). This behavior is consistent with our findings in the SSA (not shown).

4.2 | South Study Area

The SSA results are less compelling when compared with the NSA, as can be seen in Figure 3, in agreement with the results for regular statistical metrics.[6] The skill of both Control and Experimental Runs is again positive for all forecast hours (upper panel of Figure 3), slightly decreasing with the longer forecast hours, and its value is comparable with the skill in the NSA (upper panel of Figure 1). The improvement due to assimilation of the WFIP observations, although mostly positive through forecast hour 14 (middle panel of Figure 3), is not as good as in the NSA. When we average the SSA results over the first nine forecast hours, for all six data denial periods, and for the three ramp identification methods, we get a 2.2% improvement in the ramp skill score generated by assimilation of the additional WFIP observations. The difference in skill of Experimental and Control runs is only statistically significant for forecast hours 0, 2, 3, and 6 at the 95% confidence level.

One of the possible reasons for the smaller positive impact of the additional observations in the SSA is that, as we mentioned in Section 4, fewer WFIP instruments were deployed in the SSA, fewer tall towers were available, and both were less evenly distributed over the domain, compared with the NSA. It is also possible that the weather responsible for generating the ramp events in the SSA occurred on smaller spatial scales (eg, convective outflow boundaries), making it more challenging for their structure to be accurately assimilated into the model. In Figure 4, we show the position of the towers utilized in this study to assess the skill of the RAP model at forecasting ramp events, relative to the additional WFIP instrumentation. Because the tower data, as well as their position, are a proprietary information, Figure 4 shows only the outline of the tower locations.

We can see that the area where the 134 tall towers are located in the NSA (90% inside the cyan circle in Figure 4, left map) is better centered around the area where the additional instruments were deployed (nine WPR denoted by the black triangles, and five sodars denoted by the red

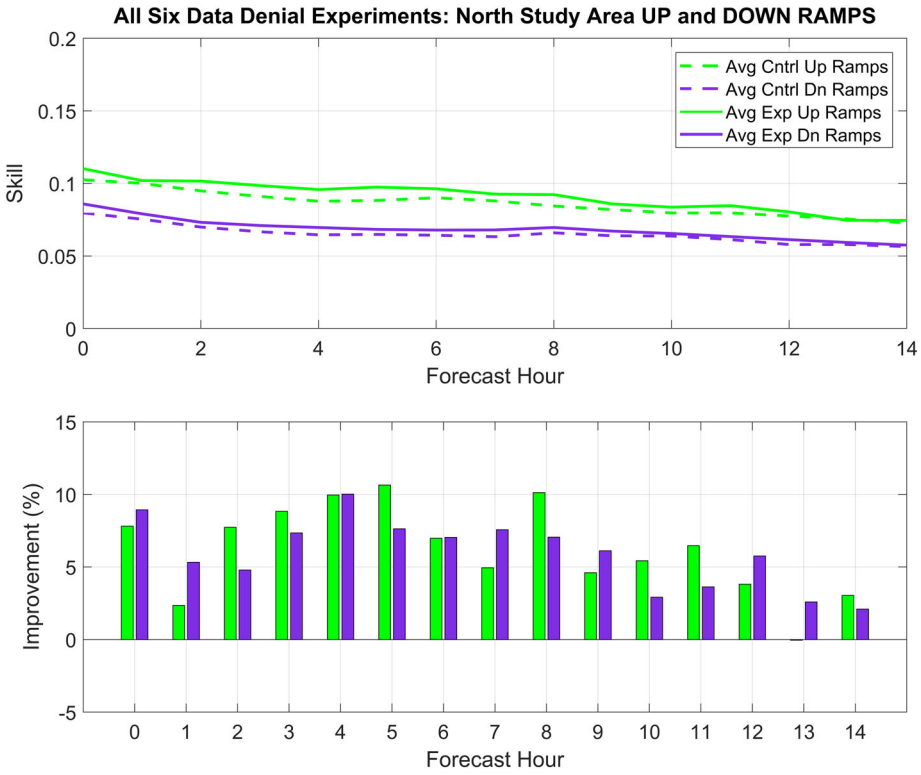

FIGURE 2 As in upper and middle panel of Figure 1, but for the up-ramp and down-ramp events separately. The skill of the RAP model at forecasting up-ramp events is shown in green, down-ramp events—in purple (dashed lines in the upper panel are relative to Control Runs and solid lines are relative to Experimental Runs) [Colour figure can be viewed at wileyonlinelibrary.com]

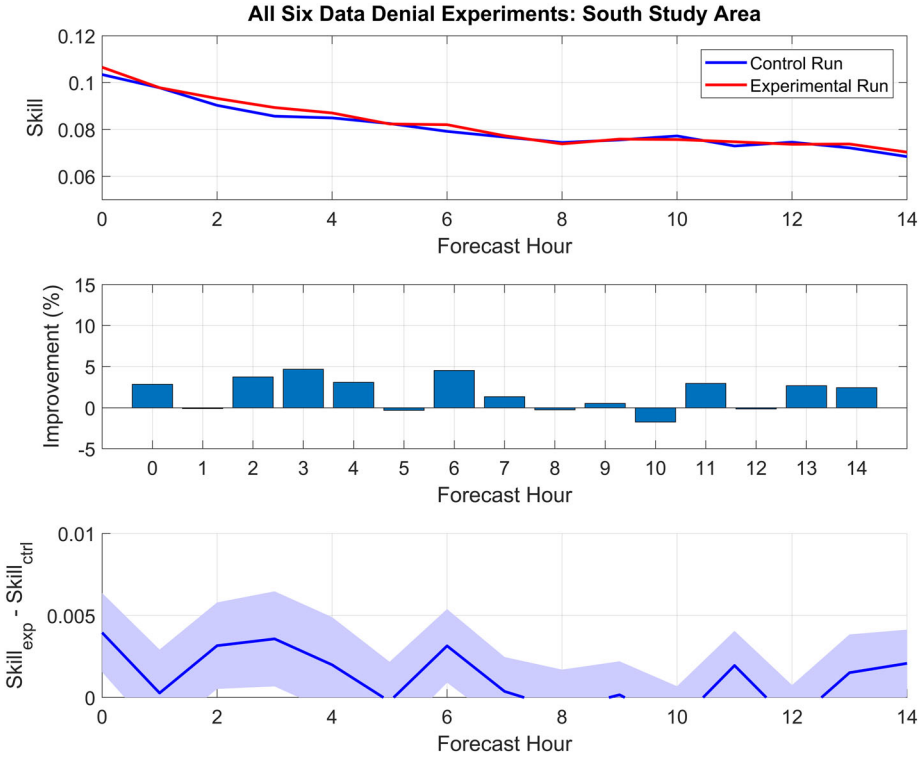

FIGURE 3 As in Figure 1, but for the SSA [Colour figure can be viewed at wileyonlinelibrary.com]

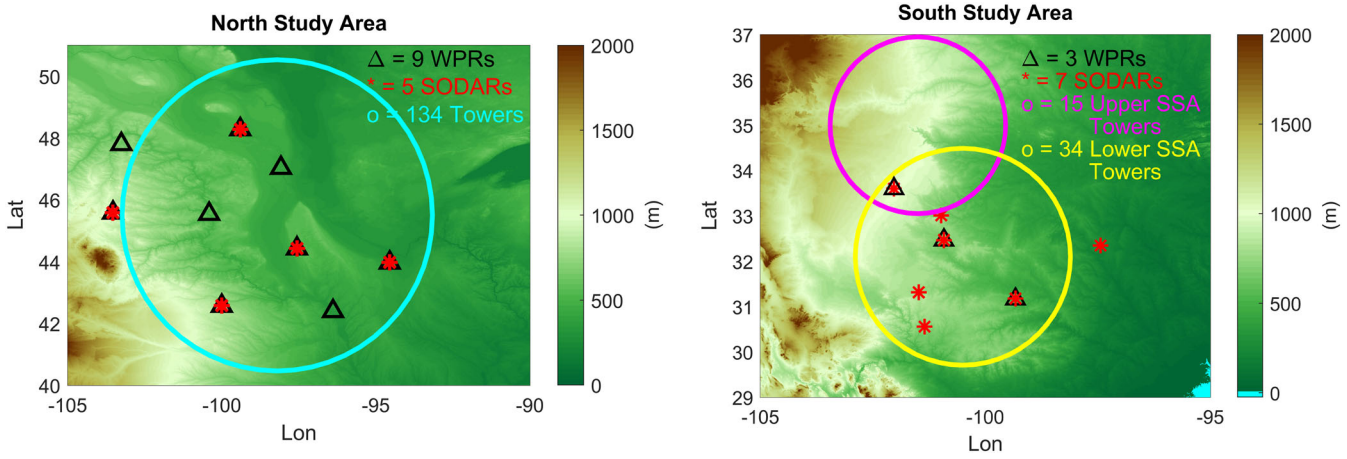

FIGURE 4 Position of the towers utilized in this study relative to the additional instrumentation deployed for WFIP. Left map: NSA with the cyan circle representing the area where 90% of towers are located. Right map: SSA, with the magenta circle representing the area where 80% of towers in Upper SSA are located and the yellow circle representing the area where all the towers in Lower SSA are located. Triangles show WPR locations, and asterisks show sodars locations [Colour figure can be viewed at wileyonlinelibrary.com]

asterisks). For the SSA (Figure 4, right map), we had two sets of tall towers, the Upper SSA Towers and the Lower SSA Towers, whose locations were approximately inside the magenta (80%) and yellow circles (100%), respectively. While the 34 towers in Lower SSA are better centered on the area where the additional instruments were deployed (three WPRs and seven sodars), the 15 towers in Upper SSA are largely to the north, making it possible that in some cases they might miss the benefit of the additional instrumentations.

Examining the six data denial periods separately, we found that the largest difference in the improvement due to the additional observations between the Upper SSA Towers and the Lower SSA Towers occurred for the 7 to 15 January 2012, data denial period. This is clear from Figure 5, where we present the improvement of the Experimental over the Control Run computed for both sets of towers (Upper and Lower SSA combined, upper left panel), improvement computed only over the Upper SSA dataset (central left panel), and improvement computed over the Lower SSA dataset (lower left panel). We see that the improvement is much more evident for the Lower SSA dataset compared with the Upper SSA dataset.

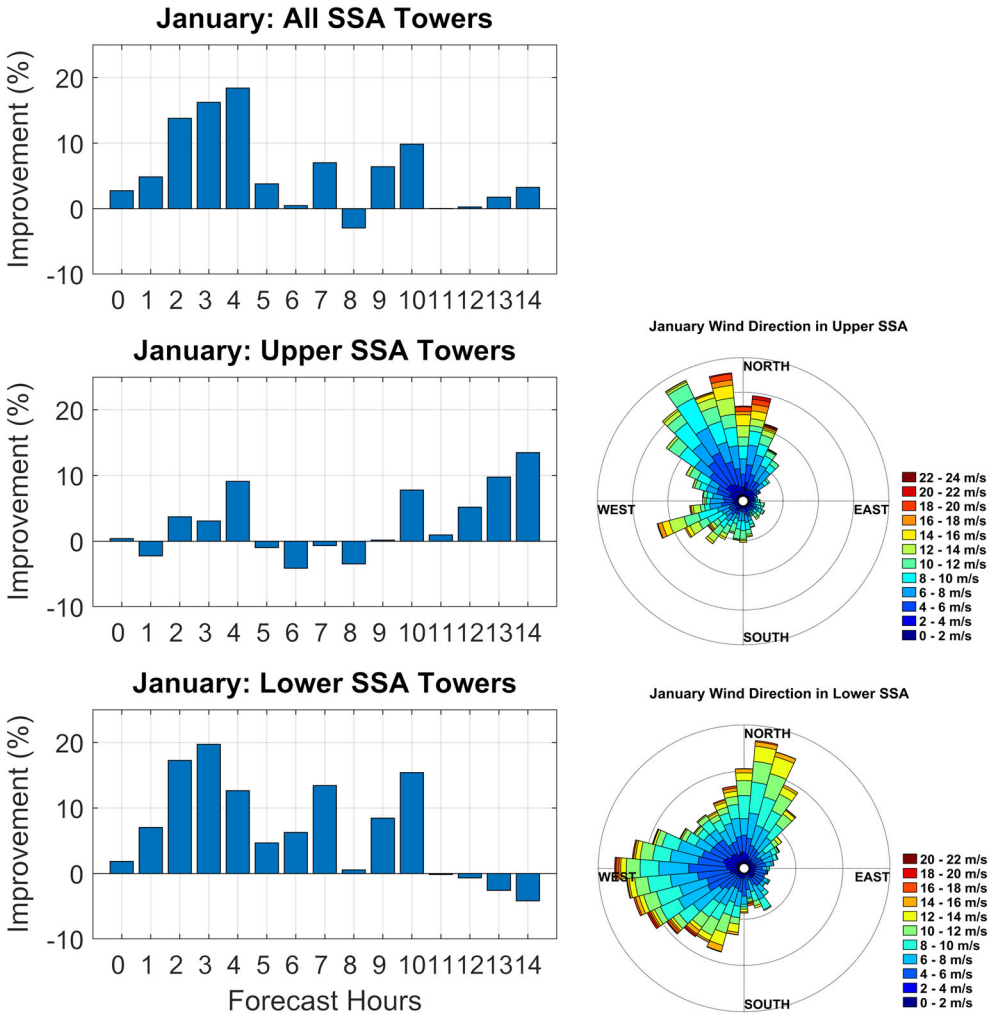

FIGURE 5 SSA: Left side: Percentage improvement of the Experimental over the Control Run for the average of three ramp identification methods as a function of the forecast hour for the 7 January to 15 January 2012 data denial period. Upper-left panel: Results for the entire SSA tower dataset. Central-left panel: Results for the Upper SSA dataset. Lower-left panel: Results for the Lower SSA dataset. Right side: Predominant wind direction at the Upper SSA and at the Lower SSA, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

For this data denial period, the improvement in the Upper SSA, averaged over the first nine forecast hours and the three ramp identification methods, is basically neutral (0.5%, from central panel of Figure 5), while it is largely positive in the Lower SSA (9.3%, from lower left panel of Figure 5).

The reason for this difference in the impact of the additional WFIP observations during the 7 to 15 January 2012 data denial period can be traced to the predominant wind directions in the Upper SSA and in the Lower SSA (right side of Figure 5). Here, we observe a predominantly northerly direction at the location of the Upper SSA Towers, and predominantly northerly and westerly wind directions at the location of the Lower SSA Towers. The towers in Lower SSA, which are located approximately in the yellow circle in the right map of Figure 4, are downwind to the additional instruments and can benefit more from the assimilation of these observations. On the other hand, the Upper SSA Towers, which are located approximately in the magenta circle in the right map of Figure 4, do not see any benefit as they are upwind of the additional WFIP instruments.

The next step of the analysis consisted in investigating the predominant wind direction during all the data denial experiments at the location of the two sets of SSA Towers. We selected the data denial periods with a clear predominant wind direction from which we could classify the Upper and Lower SSA as being upwind or downwind of the additional instruments. The selected data denial periods are: 7 to 15 January 2012, 14 to 25 April 2012, 9 to 18 June 2012, and 30 November to 6 December 2012, and their corresponding wind roses in the range between 55 and 90 m are presented in Figure 6, for the entire SSA. During the other two data denial periods, there was a less well-defined predominant wind direction making it difficult to classify the Upper and Lower SSA as being clearly upwind or downwind, and therefore we excluded them from this analysis.

We found that when the evaluation region has fewer additional upstream observations, the improvement is smaller. Thus, for the Upper SSA, a predominantly northerly wind direction (observed in Figures 6A-D for the 7 to 15 January 2012 and 30 November to 6 December 2011 data

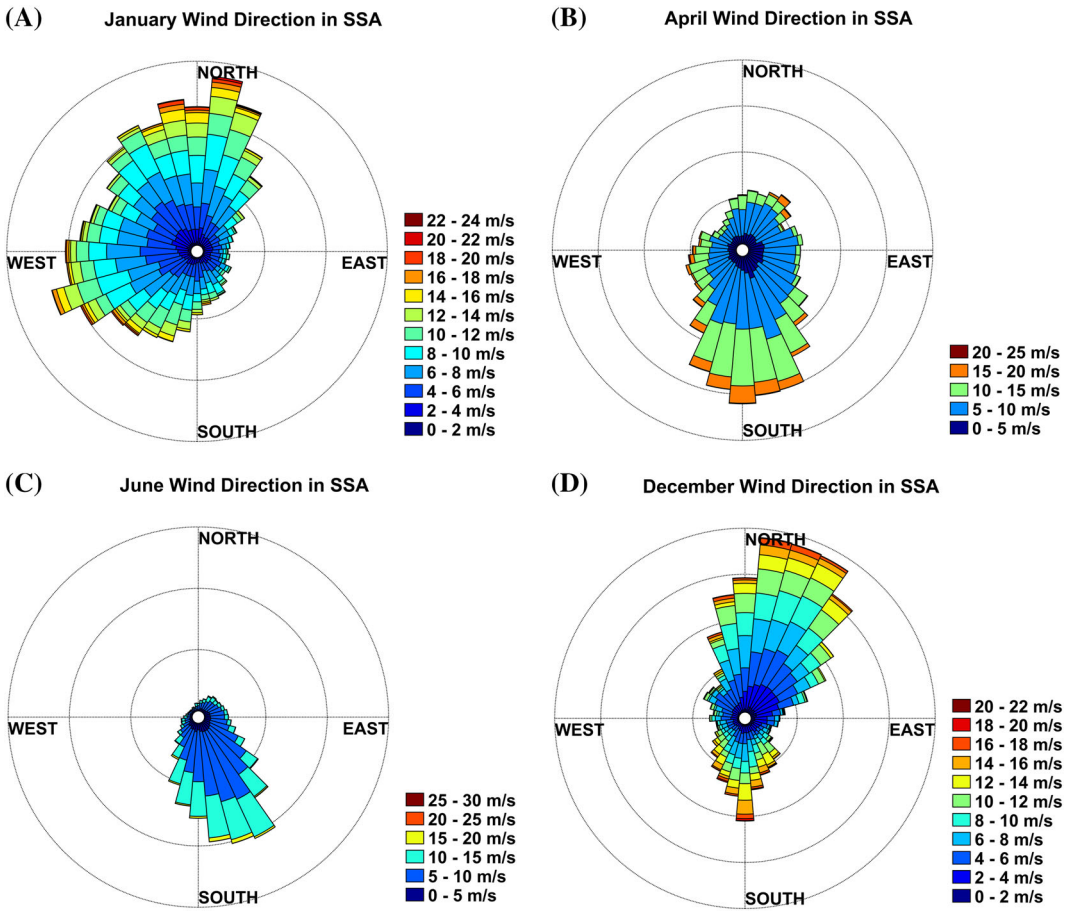

FIGURE 6 SSA: Predominant wind direction for the following data denial periods in the entire SSA: A, 7 to 15 January 2012, B, 14 to 25 April 2012, C, 9 to 18 June 2012, and D, 30 Nov to 6 December 2012 [Colour figure can be viewed at wileyonlinelibrary.com]

denial experiments), and for the Lower SSA, a predominantly southerly wind direction (observed in Figures 6B-C for the 14 to 25 April 2012 and 9 to 18 June 2012 data denial experiments) produce only a neutral impact (−0.7% improvement rate averaged over the first nine forecast hours) due to the additional WFIP observations. In contrast, for the Upper SSA, a predominant southerly wind direction, and for the Lower SSA, a northerly direction produced a larger positive impact from assimilation of the additional WFIP observations, with a 3.5% improvement over the first nine forecast hours. This highlights the importance of the strategic positioning of additional instrumentation when used for data assimilation relative to the area of desired improvement.

4.3 | Impact of different observations on the results

In this section, we assess the relative impact of the data assimilation for the remote sensing data alone, for the in situ data alone, and for the two combined, using the RT&M on the 13 to 20 October 2011 and the 7 to 15 January 2012 data denial periods. Although a heavily instrumented experiment such as WFIP provides a unique data set, the intent here is to provide information on the relative value of the different instrumentation types should either remote sensing or in situ observations become routinely available in the future. Results averaged over the 13 to 20 October 2011 and the 7 to 15 January 2012 data denial periods and over the NSA and the SSA are presented in Figure 7. The percent improvement of the skill of RAP simulations at forecasting ramp events is shown when assimilating the combined WFIP remote sensors and in situ instruments (black bars), when assimilating the WFIP remote sensor observations (green bars), and when assimilating the WFIP in situ observations (magenta bars). Similar to what was found by Wilczak et al,[9] the impact due to the assimilation of additional in situ observations has a significant initial impact that diminishes rapidly over several hours. In comparison, assimilation of the less numerous remote sensing observations has a smaller initial impact but remains positive for a longer forecast lead times. The longer duration of the positive impact is likely due to the remote sensors observing a deeper layer of the atmosphere. The sum of the improvements of the in situ and remote sensing data assimilated separately (magenta plus green bars) is generally close to, but different than, that when the data are assimilated together.

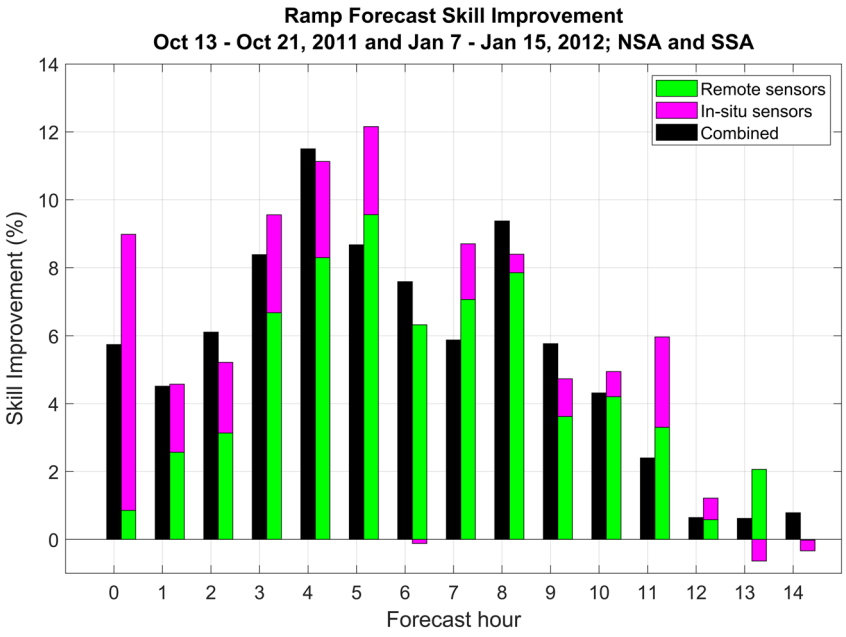

FIGURE 7 Skill percent improvement for RAP simulations that assimilate various combinations of the WFIP data, averaged for the 13 to 20 October 2011 and the 7 to 15 January 2012 data denial periods. The black bars are percent improvements due to the assimilation of the combined WFIP remote sensors and in situ instruments; the green bars show the impact of assimilating only the remote sensor observations; and the magenta bars show the impact of assimilating only the in situ observations. The green and magenta bars are incremental [Colour figure can be viewed at wileyonlinelibrary.com]

5 | SUMMARY AND CONCLUSIONS

In this study, we presented the results of the RT&M developed for the first WFIP at measuring the skill of the NOAA/ESRL RAP model at forecasting wind ramp events. WFIP took place in two different areas of the United States, with a different number of additional observations deployed for each area. The RT&M was applied to a set of six data denial experiments during which the RAP model was run first assimilating only the conventional observations (Control Run), and later assimilating both the conventional and the additional observations from the instruments deployed for WFIP (Experimental Run).

Our main results are as follows:

- The skill of RAP model at forecasting ramp events in both the Control and Experimental Runs is positive for all forecast hours, slightly decreasing with the longer forecast hours. This behavior is consistent among the three different ramp identification methods and is true for both the NSA and the SSA; however, the statistical significance of the improvement is more robust in the NSA.
- While we found a positive skill for the RAP model at forecasting ramp events for both of the Control and Experimental Runs, we found that the skill is higher at forecasting up-ramp events, compared to down-ramp events, for both the NSA and the SSA.
- The NSA and the SSA showed different results in terms of percentage improvement of the Experimental Run over the Control Run, with the impact of the additional WFIP instruments being more significant in the NSA (7%, averaged over the six data denial periods, first nine forecast hours) than in the SSA (2.2%).
- Factors that reduce the improvement in model skill in the SSA are the smaller number of additional observations compared to the NSA and the position of many of the towers used for verification relative to the WFIP remote sensing instruments. To this end, we found that the improvement from the additional instruments is positive when the verification towers are located downwind from the assimilated instruments, and it is neutral when the verification towers are upwind. This emphasizes the importance of the strategic positioning of instrumentation, particularly when the assimilation of the data is meant to benefit the wind energy industry.
- The RT&M was applied to a subset of two data denial experiments to measure the relative impact of assimilating separately remote sensing observations versus in situ observations. While assimilation of additional in situ observations had a significant initial impact that diminished rapidly after several hours, assimilation of the remote sensing observations had a smaller initial impact but remained positive for longer forecast horizon times.

ACKNOWLEDGEMENTS

This research has been funded by the US Department of Energy under the Wind Forecast Improvement Project (WFIP), award: DE-EE0003080 and by the NOAA Earth System Research Laboratory.

The authors wish to acknowledge Barb DeLuisi from the NOAA/ESRL/PSD group for maintaining the RT&M web page.

ORCID

Elena Akish https://orcid.org/0000-0001-8427-4188
Laura Bianco http://orcid.org/0000-0002-4022-7854
Irina V. Djalalova https://orcid.org/0000-0003-1299-5925
James M. Wilczak https://orcid.org/0000-0002-9912-6396

REFERENCES

1. Hodge B, Florita A, Sharp J, Margulis M, Mcreavy D. The Value of Improved Short-Term Wind Power Forecasting. National Renewable Energy Laboratory (NREL), Technical Report No. NREL/TP-5D00–63175. 2015.
2. Bianco L, Djalalova IV, Wilczak JM, et al. A wind energy ramp tool and metric for measuring the skill of numerical weather prediction models. *Weather Forecast*. 2016;31:1157-1177.
3. Finley C, Ahlstrom M, Sheridan L, et al. 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 Technical Report to DOE 2014, award number DE-EE0004421. 125 pp. http://www.osti.gov/scitech/biblio/1129929
4. Freedman J, Flores I, Zack J, et al. 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 Technical Report to DOE 2014, award number DE-EE0004420. 107 pp. http://www.osti.gov/scitech/biblio/1129905
5. Wilczak JM, Bianco L, Olson J, et al. 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 Technical Report to DOE 2014, award number DE-EE0003080, 162 pp. http://energy.gov/sites/prod/files/2014/05/f15/wfipandnoaafinalreport.pdf
6. Wilczak JM, Finley C, Freedman J, et al. The Wind Forecast Improvement Project (WFIP): a public-private partnership addressing wind energy forecast needs. *Bull Am Meteorol Soc*. 2015;19:1699-1718. https://doi.org/10.1175/BAMS-D-14-00107.1
7. Skamarock WC, Klemp JB, Dudhia J, et al. A description of the Advanced Research WRF version 3. NCAR Technical Note 475. 2008.
8. Wu W-S, Purser RJ, Parrish DF. Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon Weather Rev*. 2002;130:2905-2916. https://doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2
9. Wilczak JM, Olson J, Djalalova I, et al. Data assimilation impact of in situ and remote sensing meteorological observations on wind power forecasts during the first Wind Forecast Improvement Project (WFIP). *Wind Energy*. 2019;1-13. https://doi.org/10.1002/we.2332
10. International Electrotechnical Commission, 2007: Wind turbines—Part 12-1: Power performance measurements of electricity producing wind turbines. IEC 61400–12-1, 90 pp.

How to cite this article: Akish E, Bianco L, Djalalova IV, et al. Measuring the impact of additional instrumentation on the skill of numerical weather prediction models at forecasting wind ramp events during the first Wind Forecast Improvement Project (WFIP). *Wind Energy*. 2019;22:1165–1174. https://doi.org/10.1002/we.2347