Wind Ramp Events Validation in NWP Forecast Models during the Second Wind Forecast Improvement Project (WFIP2) Using the Ramp Tool and Metric (RT&M)

IRINA V. DJALALOVA,[a,b] LAURA BIANCO,[a,b] ELENA AKISH,[a,b] JAMES M. WILCZAK,[b] JOSEPH B. OLSON,[a,b]
JAYMES S. KENYON,[a,b] LARRY K. BERG,[c] ADITYA CHOUKULKAR,[a,d] RICHARD COULTER,[e]
HARINDA J. S. FERNANDO,[f] ERIC GRIMIT,[g] RAGHAVENDRA KRISHNAMURTHY,[c,f] JULIE K. LUNDQUIST,[h,i]
PAYTSAR MURADYAN,[e] DAVID D. TURNER,[b] AND SONIA WHARTON[j]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *National Oceanic and Atmospheric Administration/Earth Systems Research Laboratories, Boulder, Colorado*
[c] *Pacific Northwest National Laboratory, Richland, Washington*
[d] *Vibrant Clean Energy LLC, Boulder, Colorado*
[e] *Argonne National Laboratory, Lemont, Illinois*
[f] *Civil and Environmental Engineering and Earth Sciences, University of Notre Dame, Notre Dame, Indiana*
[g] *Vaisala Inc., Seattle, Washington*
[h] *Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, Colorado*
[i] *National Renewable Energy Laboratory, Golden, Colorado*
[j] *Lawrence Livermore National Laboratory, Livermore, California*

(Manuscript received 7 May 2020, in final form 3 September 2020)

ABSTRACT: The second Wind Forecast Improvement Project (WFIP2) is a multiagency field campaign held in the Columbia Gorge area (October 2015–March 2017). The main goal of the project is to understand and improve the forecast skill of numerical weather prediction (NWP) models in complex terrain, particularly beneficial for the wind energy industry. This region is well known for its excellent wind resource. One of the biggest challenges for wind power production is the accurate forecasting of wind ramp events (large changes of generated power over short periods of time). Poor forecasting of the ramps requires large and sudden adjustments in conventional power generation, ultimately increasing the costs of power. A Ramp Tool and Metric (RT&M) was developed during the first WFIP experiment, held in the U.S. Great Plains (September 2011–August 2012). The RT&M was designed to explicitly measure the skill of NWP models at forecasting wind ramp events. Here we apply the RT&M to 80-m (turbine hub-height) wind speeds measured by 19 sodars and three lidars, and to forecasts from the High-Resolution Rapid Refresh (HRRR), 3-km, and from the High-Resolution Rapid Refresh Nest (HRRRNEST), 750-m horizontal grid spacing, models. The diurnal and seasonal distribution of ramp events are analyzed, finding a noticeable diurnal variability for spring and summer but less for fall and especially winter. Also, winter has fewer ramps compared to the other seasons. The model skill at forecasting ramp events, including the impact of the modification to the model physical parameterizations, was finally investigated.

KEYWORDS: Model comparison; Model evaluation/performance; Renewable energy; Wind effects

1. Introduction

The second Wind Forecast Improvement Project (WFIP2) included an 18-month field campaign (October 2015–March 2017) held in the Columbia River Gorge and basin, in Oregon and Washington states, a region well known for its excellent wind resources. WFIP2 was led by the U.S. Department of Energy (DOE) and by the National Oceanic and Atmospheric Administration (NOAA), and was supported by several other public and private institutions (Shaw et al. 2019). Its main goal was to increase the accuracy of numerical weather prediction (NWP) model forecasts of wind speed in complex terrain, through the improvement of NWP physical parameterization schemes. The models tested in this study are the High-Resolution Rapid Refresh (HRRR) model, with 3-km horizontal grid spacing, and the High-Resolution Rapid Refresh Nest (HRRRNEST) model, with 750-m horizontal grid spacing (Olson et al. 2019a). At the end of the campaign, four 6-week periods, one for each season, were identified to assess the model improvements. We will hereafter refer to these four 6-week intervals as ''reforecast periods.'' Over these four reforecast periods, the HRRR and HRRRNEST were run in control (CTL) and experimental (EXP) setups, where the experimental configuration included parameterization modifications developed, tested and applied to the models during the field campaign (Olson et al. 2019a,b) to improve the forecast of hub-height wind speeds. We will hereafter refer to these four 6-week testing experiments as ''reforecast runs.''

Model verification and validation were performed using a variety of instruments deployed during WFIP2 (Wilczak et al. 2019b; Pichugina et al. 2019; Bianco et al. 2019; Grimit 2020). Among those were 19 sonic detection and ranging (sodars) and 7 light detection and ranging (lidars), measuring wind speed and direction from a minimum of 10 m to a few hundred meters (or more) above ground level (AGL), depending on the system. These heights represent the layer of the atmosphere most relevant for wind energy production, as turbine rotor-disk

Corresponding author: Irina V. Djalalova, Irina.v.djalalova@noaa.gov

DOI: 10.1175/WAF-D-20-0072.1

height generally ranges from 40 to 150 m. Standard statistics [bias and mean absolute error (MAE)] of average hub-height 80-m wind speed used to evaluate the improvements in forecasts have been described in Bianco et al. (2019). In the present study we emphasize validation of the model improvements at forecasting wind ramp events.

Wind ramps are rapid changes of wind speed over short periods of time, which result in rapid changes in power production. The correct forecast of ramp events is important for wind energy operators since, if they are not forecast accurately in directionality, amplitude and timing, the operator might need to request other forms of power production in a short period of time to meet the power demand. These rapid changes in power source can be quite expensive to accommodate, and being able to rely on an accurate NWP model could result in significant cost savings. As wind power penetration increases, wind ramps can become larger and more critical for power operators.

There is an increasing interest in the evaluation of wind power ramp forecasting, which has resulted in several studies. Ferreira et al. (2011) presented a survey on wind power ramp forecasting, Gallego-Castillo et al. (2015) provided a review on the recent history of wind power ramp forecasting, and Zhang et al. (2014) analyzed the advantage of using a short-term wind power forecast to improve the accuracy of the wind power ramp forecasting during the first Wind Forecast Improvement Project (WFIP). These studies (and others such as Bossavy et al. 2010; Cutler et al. 2007; Greaves et al. 2009; Kamath 2010; Zack et al. 2010) used different methods for the identification of wind ramp events. The method used here (based upon Bianco et al. 2016) includes several aspects of the earlier approaches, and also enables us to consider different ramp definitions simultaneously.

Previous analyses of wind ramp events in the WFIP2 region provide insights into the daily distribution of ramp events and their relative severity levels (Kamath 2010), evaluating the Weather Research and Forecasting (WRF) Model at predicting ramp events using different boundary layer schemes (Yang et al. 2013), or introducing a probabilistic approach for forecasting ramp events (Worsnop et al. 2018). In the present study we use a tool that gives us the option to not only consider one or two ramp definitions, but a matrix of ramp definitions. Also, the skill of the model considers the difference in duration of observed and forecast ramps, as well as the difference in their amplitudes and timings.

The manuscript is organized as follows: section 2 describes the dataset utilized in this study, including both observations and NWP models and their correlation with the Bonneville Power Administration (BPA) power generation data; section 3 summarizes the Ramp Tool and Metric (RT&M) used to calculate model forecast skill of ramp events and the approach that we take to organize the model run outputs for this study; section 4 describes the observed diurnal and seasonal characteristics of 80-m wind speed and ramp events; section 5 contains the RT&M results of model skill score statistics, while section 6 presents the same RT&M results but for bias-corrected models. Finally, section 7 draws conclusions and outlines a future research plan.

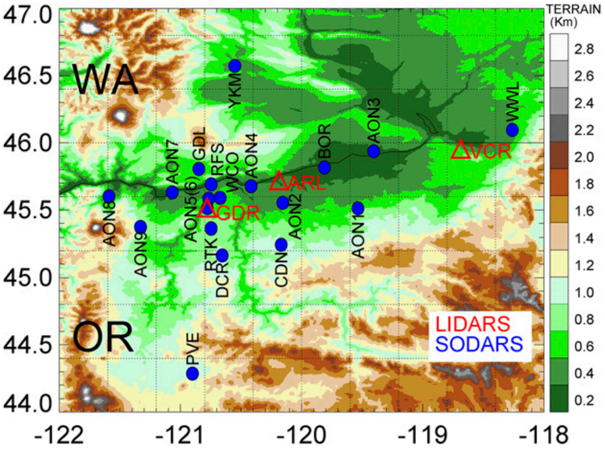

FIG. 1. Map of the 19 sodar (blue circles) and three lidar (red triangles) locations, selected for this study.

2. Dataset description

a. Observational dataset

Data from 19 sodars and seven lidars were collected during the WFIP2 field campaign. While most of the sites had either a sodar or a lidar, several sites had both instruments. At these locations, the instrument having the most complete dataset was selected to be included in this study. In addition, a lidar located at the Troutdale, Oregon, site has not been included in our analysis because this location is on the west side of the Cascade Mountains while our area of interest lies in the high wind energy production area east of the Cascades. This resulted in 22 sites, 19 with sodars and 3 with lidars used in this study. The map with the locations of the 22 sites is presented in Fig. 1 and instrument's location, manufacturer and institution in charge are described in Table 1 of Bianco et al. (2019).

Since not all of the 19 sodars and three lidars were continuously operational (hardware or software failures occurred during the experiment, as well as late deployments or early removals), measurements for our analysis were available from 20, 18, 16, and 19 instruments for the spring, summer, fall, and winter reforecast periods, respectively.

b. NWP models

As mentioned earlier, the models of interest in this study are the HRRR and the HRRRNEST, since their refined grid spacing compared to other NWP models better resolves the complex orography. For the four 6-week reforecast periods the HRRR and HRRRNEST were run out to 24 forecast hours, with 15-min output, using initial conditions initialized from the operational Rapid Refresh model (RAP; Benjamin et al. 2016); no additional data assimilation was used when initializing the HRRR for this study. Similarly, the HRRRNEST was initialized the same way, but getting its initial and boundary conditions from the 3 h forecasts provided by the HRRR. The four reforecast periods covered the four seasons: the spring period extended from 25 March to 7 May 2016; the summer period from 24 June to 7 August 2016; the fall period from 24 September to 7 November 2016; and the winter period from

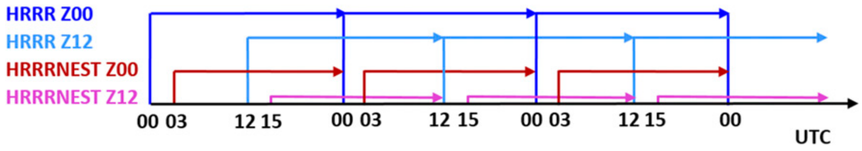

FIG. 2. Schematic of the HRRR and HRRRNEST model run outputs for the four 6-week reforecast runs (LST = UTC − 8).

25 December 2016 to 7 February 2017. The models were initialized every 12 h (at Z00—0000 UTC, and at Z12—1200 UTC; with local standard time, LST = UTC − 8). The HRRRNEST runs were delayed by 3 h to avoid spinup problems, so that a gap in the HRRRNEST model output exists from forecast horizon 00 to forecast horizon 02 (from 0000 to 0245 UTC for the Z00 initialized run, and from 1200 to 1445 UTC for the Z12 initialized run). The schematic of the availability of the HRRR and HRRRNEST model outputs for the reforecast periods is presented in Fig. 2.

For the reforecast runs, the models were run in both CTL and EXP setups. Major differences between the two include new parameterizations to the HRRR and HRRRNEST physics suite, improvements to existing parameterizations, and improvements to numerical methods. The reader is referred to Olson et al. (2019a,b) for details on these changes. Some developments, including mixing length revision, the addition of surface drag due to subgrid-scale orography, and modification of the horizontal diffusion, all improve the maintenance of cold pools. Since these developments show the biggest impacts on low-level wind compared to other model modifications, we may expect the biggest improvement of ramp skill in winter.

For our analysis, to compare to the observations, the 80-m wind field model output is horizontally interpolated bilinearly to the 22 site locations using the four closest grid points. In the vertical, model levels closest to the ground are approximately 10, 38, 88, 169, and 288 m AGL, from which wind speed is linearly interpolated to 80 m AGL using the two closest model levels. From here on, we reference the "model data" as the model output interpolated to the 22 instrument locations at 80 m AGL.

c. *Correlation with BPA power generation data*

The instruments used in our analysis are located inside the regional grid operator's (BPA) territory (Fig. 3), which includes Idaho, Oregon, Washington, western and small parts of eastern Montana, California, Nevada, Utah, and Wyoming (https://www.bpa.gov/news/AboutUs/). Wind power generation in the BPA territory reaches more than 6000 MW, with a peak value close to 4500 MW in the WFIP2 Columbia Basin area.

To evaluate NWP models, we convert 80-m wind speeds to rated pseudo-power (hereafter referred to simply as power), using a generic/normalized power curve provided in Wilczak et al. (2019a). The same power conversion curve is used for both observations and NWP model outputs. For the four reforecast periods, Fig. 4 compares the aggregated wind-power generated from the 22 WFIP2 observation locations to the actual wind power production on the BPA system (https://transmission.bpa.gov/Business/Operations/Wind/) and to the aggregated power forecasts from the HRRR CTL(Z12 initialized runs) model at the same 22 sites.

Although the area of the BPA system is much larger that the WFIP2 study area, the comparison shown in Fig. 4 indicates that the instruments deployed during WFIP2 provided a representative measure of the wind power production over the entire BPA system, due to the concentration of wind plants in the study area domain. This figure also shows how winter differs from all other seasons with a lot of days with zero, or close to zero, power production. These low-production periods are due to long-duration cold pool events in winter (McCaffrey et al. 2019). During these events, the river basin experiences stagnant cold air persisting over the area for several days, with associated weak wind speeds close to the surface.

Pearson correlation coefficients for observed versus BPA power, and for observed versus HRRR CTL power are presented in both Fig. 4 and Table 1. Table 1 also shows correlation coefficients for HRRR CTL versus BPA power for every reforecast period. Similar results are found for the Z00 initialized runs and for the other model runs. The purpose of this comparison is to prove that the observations used in this study represent the power production in the whole BPA area.

3. Ramp Tool and Metric

a. *Ramp analysis configuration*

A Ramp Tool and Metric (RT&M; Bianco et al. 2016; Akish et al. 2019) was developed during the first WFIP (Wilczak et al.

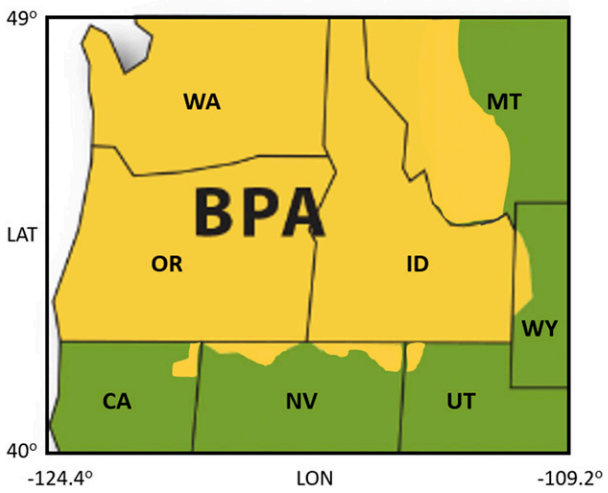

FIG. 3. Bonneville Power Administration area. The BPA renewable generation assets map is adapted from Western Area Power Administration (2018).

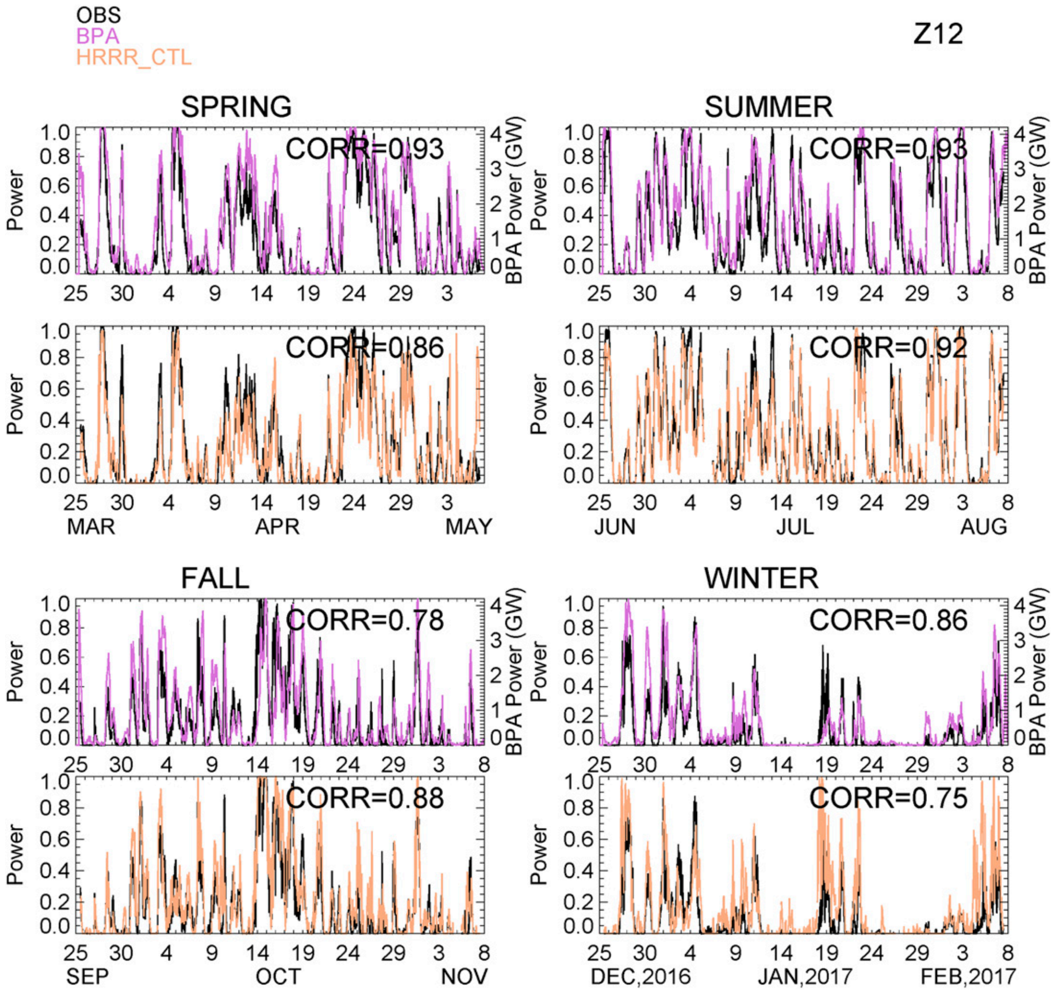

FIG. 4. Time series of aggregated rated-power for the four reforecast periods: (top left) spring, (top right) summer, (bottom left) fall, and (bottom right) winter. Power (left *y* axis) computed from observed wind speed data are shown in black on all panels. Actual power generation data (right *y* axis; GW) provided by the BPA are in purple in the top panels, and the HRRR CTL power data (Z12 initialized runs, left *y* axis) are in beige in the bottom panels.

2015) that was held in the U.S. Great Plains from September 2011 to August 2012. The RT&M consists of three components: in the first component, ramp events are identified (by the following parameters: central time of the ramp Ct; change in power Δp; and duration Δt) in the time series of observed and forecast power (80-m wind speeds are converted into power inside the RT&M); in the second component, each forecast ramp event is matched in time with all observed ramp events that overlap in duration; and in the third component, the skill score of the NWP model is computed on an hourly basis (taking into account errors in the center time, duration and change in power; i.e., the score is a function of the cube root of the product of the errors on Ct, Δp, and Δt; see Bianco et al. 2016, for complete formulas).

A ramp is detected when the change in power meets or exceeds a specified Δp threshold over a time interval equal to or shorter than a specified Δt threshold. Since no unique definition of a ramp event exists, the RT&M allows one to compute the ramp skill over a user-selected set of different ramp definitions and then averages over them. Each ramp definition has its own possible Δp and Δt. In the standard configuration, Δp is 30%, 40%, 50%, 60%, and 70% of the rated power capacity, and Δt is 30, 60, 120, and 180 min, for a total of 20 ramp definitions over which the skill is averaged. For each ramp definition the time interval Δt is used as a sliding window over which the power change is measured. If the power change exceeds the Δp specified for that ramp definition, a ramp is found. Continuous ramps for each ramp definition are concatenated in a single ramp. For this reason, ramp lengths could be longer than the time interval specified. The process is repeated for all ramp definitions. For simplicity, later in the analysis we will often show results from 30% Δp over 180-min Δt, because the lowest threshold of 30% Δp will also include larger magnitude ramps, and the longest temporal threshold of

TABLE 1. Pearson correlation coefficients for observed power vs BPA, for HRRR CTL vs observed power, and for HRRR CTL power vs BPA for each season.

Correlation	Spring	Summer	Fall	Winter
OBS/BPA	0.93	0.93	0.78	0.86
HRRR CTL/OBS	0.86	0.92	0.88	0.75
HRRR CTL/BPA	0.80	0.82	0.71	0.69

180-min Δt will also include shorter ramps, therefore a ramp definition of 30% Δp over 180-min Δt could potentially include many ramps of the 20 ramp definitions.

After the model and observation ramps are matched, the skill of the model at forecasting ramp events is finally computed, accounting for forecast ramps matched to observed ramps, forecast ramps not matched with observed ramps, and observed ramps not matched with forecast ramps. This results in eight possible scenarios (obs/model ramp: up/up, null/up, down/up, up/null, down/null, up/down, null/down, and down/down). For all possible ramp definitions, the score is calculated as the total score divided by the total number of matched and unmatched ramp events. Unmatched ramps have a null skill score. For matched ramps, the ramp skill score of the model has a value from -1 to $+1$ and is determined by three parameters: the error in power change ($|\text{model } \Delta p - \text{obs } \Delta p|$), the error in time duration ($|\text{model } \Delta t - \text{obs } \Delta t|$). and the difference in central time ($|\text{model Ct} - \text{obs Ct}|$) [see Eqs. (1)–(6) of Bianco et al. 2016]. A negative score is assigned when two ramps with opposite power changes are matched, while matched ramps with the same sign in a power change result in a positive skill.

b. *Artificial ramps: ''Stitching method'' versus ''forecast method''*

The RT&M offers two possible approaches for organizing the dataset, called the ''stitching method'' and the ''forecast method.'' The ''stitching method'' creates time series of model forecasts, ideally for each particular forecast horizon for hourly updated models (or for a certain number of forecast hours from each consecutive forecast runs, for nonhourly updated modes), while the ''forecast method'' simply searches for ramps through the duration of an individual forecast. Previous studies (Bianco et al. 2016; Akish et al. 2019) found that these two approaches give quite consistent results when using hourly updated forecasts. It is not clear a priori if the stitching method is still applicable when using simulations separated by greater lengths as in the WFIP2 reforecast runs. In this case, rather than concatenating by forecast horizon as the stitching method was originally designed, we could create a time series of the model outputs over the entire length of each of the four reforecast periods, by concatenating the reforecast runs consecutively.

Figure 5 provides a closer look at the RT&M on the WFIP2 dataset, highlighting differences between the stitching method and forecast method when applied to a 5-day period (29 June–3 July 2016). The time series of power shown in the three-top panels, are the observed values and model concatenated values from the Z00 and Z12 runs. The bottom panel shows the skill score for each pair of matching ramps.

Several cases of ''artificial ramps'' are apparent in both of the model time series (second and third panels), at the stitching times highlighted in black circles. The artificial ramps occur because of the large forecast lengths (24 h) between consecutive forecasts, which allows the forecast to drift considerably before it is reinitialized, and results in 40%–60% more ramps found in the stitching method than the forecast method at or near the time of concatenation. Although for the HRRR it would have been possible to concatenate forecasts every 12 h, this was not possible for the HRRRNEST because gaps in the concatenated model output would exist from forecast horizon 00 to forecast horizon 02 of each model run. Since we wish to compare the skill of the HRRR with the HRRRNEST, it was necessary to use the longer 24 h intervals between consecutive model runs for both models. Because of these difficulties with the ''stitching method,'' only the ''forecast method'' is used and discussed in this paper.

The ''forecast method'' experiences no degradation for infrequently reinitialized forecasts, with single forecast runs being compared independently to the corresponding observational time series. This method has its own limitations, however, since the initial and the ending forecast horizons can suffer from truncated ramps that could start soon before the beginning, or end soon after the end of the forecast cycle. Since each ramp skill score is assigned to the hour closest to the ramp central time, skill scores cannot be computed for some length of time near the beginning and end of each forecast, with the length dependent on the ramp time-length being examined. For the HRRR model, using a Δt ramp length specification of 3 h, ramp skill scores could be assigned to forecast horizons 3–21 h. However, due to the 3-h delay of the HRRRNEST model, ramp scores can only be assigned to forecast horizons 6–21 h.

Due to these considerations and because we want to assign a score to each hour of the daily cycle, the model skill score is only calculated from forecast horizons 6 to 17 of both the Z00 and Z12 initialization runs of both the HRRR and HRRRNEST models. Specifically, the model skill score will be averaged over hours 0600–1700 UTC for the Z00 initialization runs, and over hours 1800–0500 UTC of the next day for the Z12 initialization runs so that every hour of the day is covered by one of the two model runs and the HRRR and HRRRNEST models will be treated equally. The colored ramps in the model data are shown only if their central times fall within forecast horizons 6–17. These time periods are shaded in Fig. 5 (gray for the Z00 initialization runs and pink for the Z12 initialization runs). Ramps are only shown for the model over the shaded areas. This approach to only compute the ramp skill over forecast hours 6–17 for each model is used for the rest of the manuscript.

Model and observational ramps are matched when their duration times overlap. For this reason, one observational ramp could be matched to several model ramps and vice versa. The ramp skill score is given in the bottom panel of Fig. 5, where for the final model skill score we combine the information from the Z00 and Z12 initialization times for the WFIP2 model reforecast runs. The skill score of each of the seven model/observation pairs of matched ramps is almost always positive, and close to a +1 value, being negative in one case only (3 July 2016), when a down-ramp in the HRRR CTL model (Z00

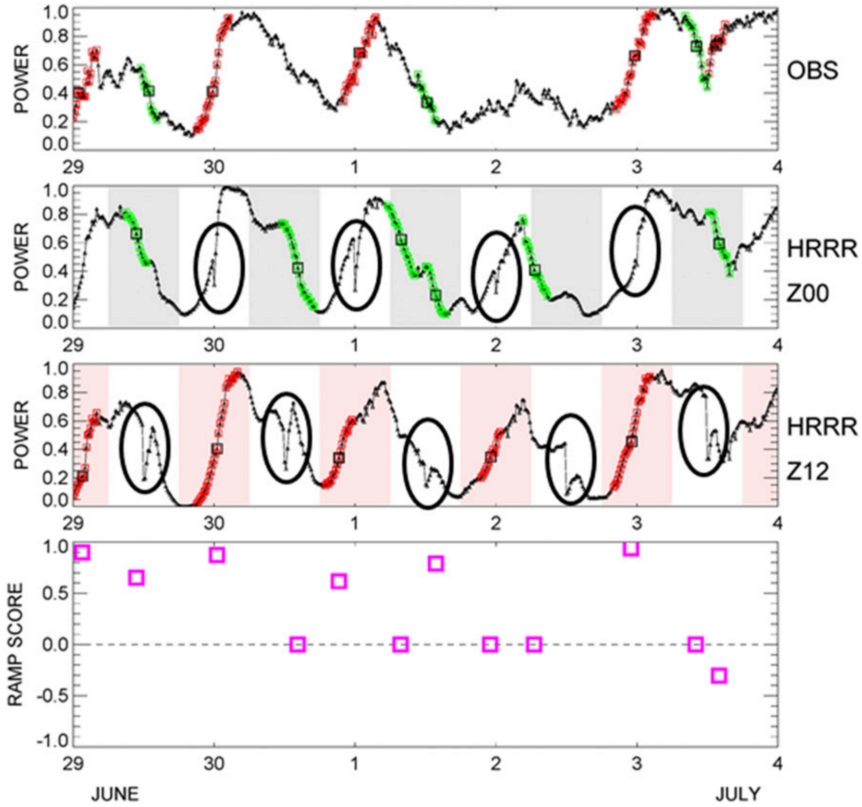

FIG. 5. Time series of normalized power from (from top to bottom) aggregated observations, HRRR CTL stitched outputs from Z00 and Z12 runs, and ramp skill score for 5 days in summer 2016. Ramps in the time series with at least 30% Δp over at least 180-min Δt are colored red for up- and green for down-ramps. Model ramps are shown only when their central time falls in the interval of the forecast horizons 06–17 (areas shaded in gray for Z00 runs and with pink for Z12 runs).

initialization time) is matched to an up-ramp in the observations, generating a negative score. All unmatched ramps, both in observational or modeled time series, are assigned with a null score (in this example there are five unmatched ramps).

As shown above, the stitching method results in spurious ramps when applied to infrequently reinitialized forecasts, and therefore is not recommended for the current dataset. Further, for the forecast method, windowing the number of hours in each forecast cycle allows for the same validation hours to be used and for the entire daily cycle to be covered, for the HRRR and HRRRNEST, so that fair comparisons can be made between the two models.

4. Diurnal and seasonal variability of 80-m wind speed and ramp events

Before showing the model ramp skill score results, we present a brief analysis of ramp statistics from the observed data. In particular, we analyze the dependency of the prevailing type of ramps on the diurnal variation of the change in mean wind speed in Fig. 6.

The gray shadow areas indicate the times during the diurnal cycle when the wind is decreasing (wind speed change is negative), while the pink areas indicate hours when there are more down ramps than up-ramps (up/down ratio $<$ 1). The agreement between the two is visible during all seasons, but is most distinct during summer. The correlation coefficients between these lines are shown in red with the highest correlation in summer (0.89), followed by spring (0.72), winter (0.6), and then fall (0.42). Thus, the diurnal variability of up- versus down-ramps is seen to be strongly dependent on the diurnal variation of the mean wind speed in summer, much less so in fall, with intermediate correlations in spring and winter. The strong dependence found in summer is due to thermally driven gap flows that result in large amplitude diurnal variations in wind speed (Pichugina et al. 2019; Wilczak et al. 2019b).

To examine the possibility of an elevation dependence of the diurnal and seasonal variability, the observed diurnal variability of 80-m wind speed is analyzed in Fig. 7 for different site elevations. While the black line is the aggregated 80-m wind speed over all sites, the blue line includes the sites with elevation between 0 and 300 m (AON3, AON7, BOR, RFS, ARL), the green line is for those with elevation between 300 and 700 m (AON2, AON4, AON5, GDL, WCO, WWL, YKM, VCR), and the red line includes the sites with elevation above 700 m (AON1, AON6, AON8, AON9, CDN, DCR, PVE, RTK, GDR).

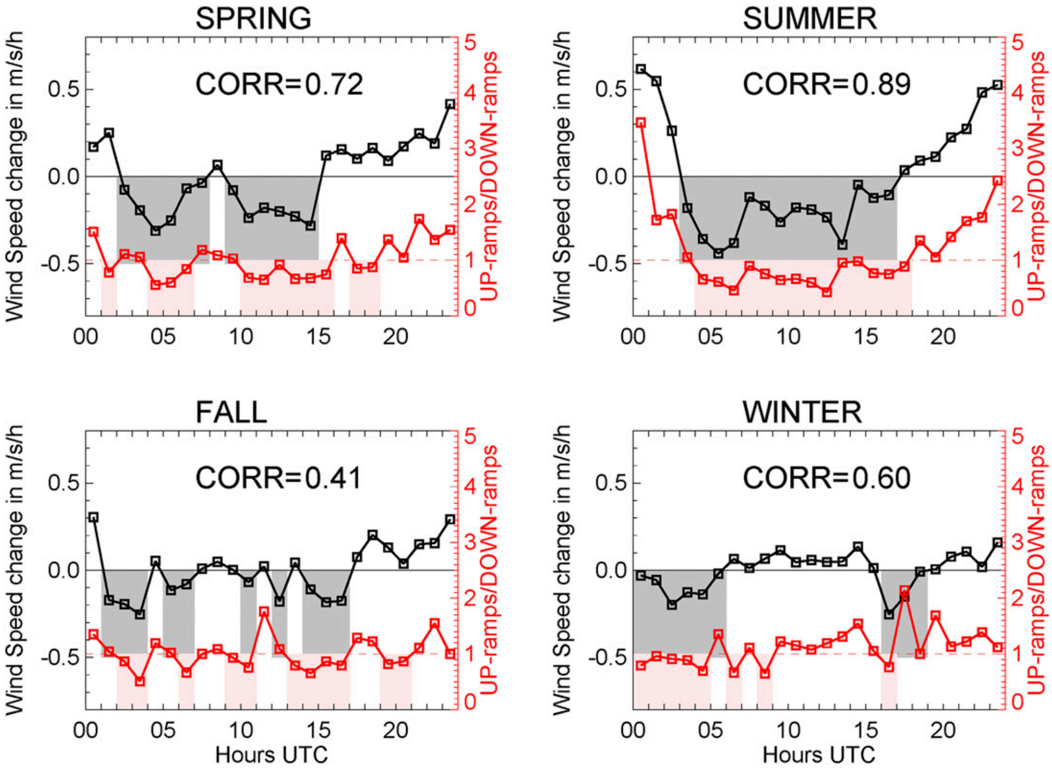

FIG. 6. Observed diurnal composite of aggregated 80-m hourly wind speed change (black lines) and corresponding ratio of the total number of up-ramps summed at individual sites to the total number of down-ramps (red lines) for a ramp definition of $\Delta p = 30\%$, and $\Delta t = 180$ min. Gray areas indicate the hours of decreasing wind speed ($du/dt < 0$), and pink areas indicate the hours with down-ramps prevailing (up-ramps/down-ramps < 1).

While wind speeds at all elevations show a similar diurnal pattern for each season, each elevation shows a different seasonal variability. Sites with lower elevation (blue and green lines) experience stronger 80-m winds in summer compared to those at higher elevation (red line), for fall and winter the opposite is true. This seasonal contrast is consistent with the seasonal variability of gap flow events. Gap flows occur preferentially in summer and spring (Sharp and Mass 2002), with accelerated winds at lower elevations. In contrast, cold pool events, with low wind speeds close to the surface, occur more often in winter and fall (Bianco et al. 2019; McCaffrey et al. 2019).

5. Models' skill at forecasting ramp events

To evaluate model ramp skill, it is informative to first examine time series of observed and forecast power and the detected ramps in each. Observations and HRRR CTL power data are shown in the three panels of Fig. 8 for all reforecast periods. The entire 24 h of the model simulations are concatenated, although ramps are only detected for forecast horizons 06–17. Identified up-ramps are colored in red while down-ramps are colored in green. The numbers of ramps are on the right side of each panel, with red numbers for up-ramps and green numbers for down-ramps.

In general, the number of ramps in the observations and the sum of the numbers of ramps in the Z00 and Z12 model runs (only taken from forecast horizons 06–17) are very similar, with values summarized in Table 2.

Spring, and especially summer, have large differences in the number of up and down ramps between the Z00 and Z12 runs. This contrast is due to the strong diurnal variation of the mean wind and ramp ratio shown in Fig. 6 and the hours of the day that each run spans. In contrast, the fall and winter periods have almost the same number of up and down ramps for both initialization times, as suggested by Fig. 6. Also, fewer ramps are present in the observed and model data in winter compared to other seasons, as expected from Fig. 4.

Winter and summer seasons have very different characteristics in 80-m wind speeds and ramp event distributions compared to each other and also to the other two seasons. Therefore, in Fig. 9 we introduce a new visual tool to help recognize possible patterns of diurnal distribution of ramp events and to compare diurnal ramp distributions of observed and model data. These diurnal distributions for the summer and winter seasons of up- and down-ramps with $\Delta p = 30\%$, and $\Delta t = 180$ min, are computed using the aggregated time series of power. The observed composite diurnal ramp distribution is in the upper panel, while all available models (HRRR CTL, HRRR EXP, HRRRNEST CTL, and HRRRNEST EXP) are from the second to the bottom panel of the figure. Similarl to Fig. 5, 06–17 forecast horizon hours are colored in gray for the data of the Z00 reforecast run and in pink for the same horizon

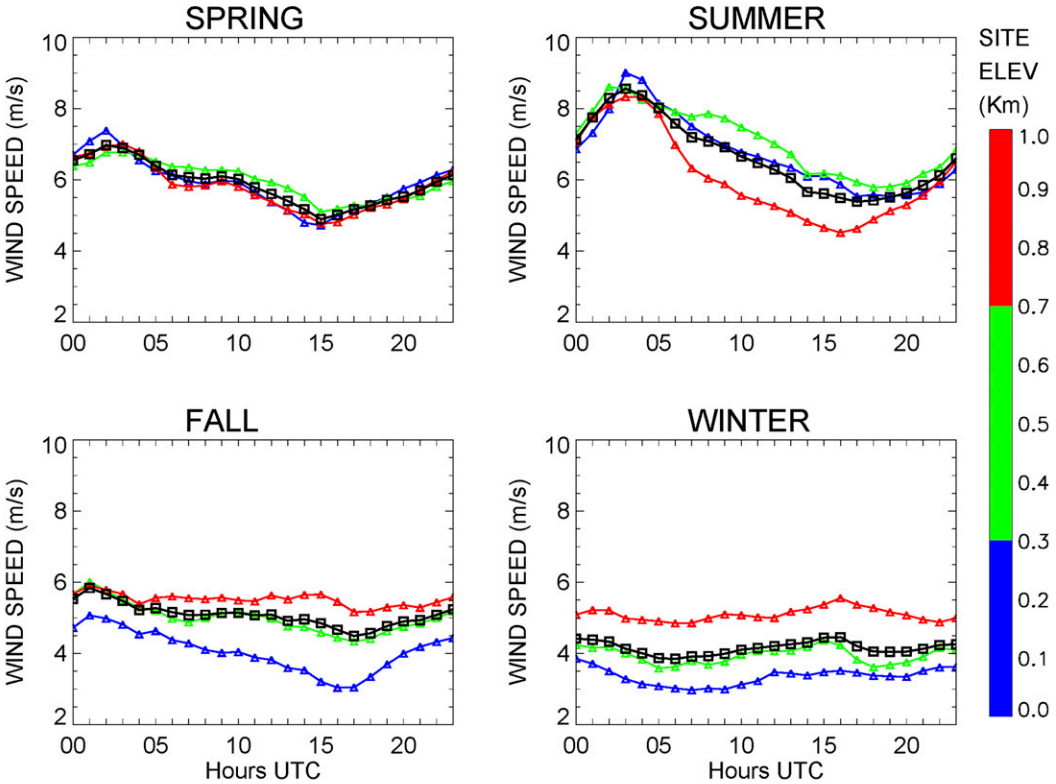

FIG. 7. Observed diurnal variability of 80-m hourly wind speed aggregated over all sites (black line) and over an average at three elevation ranges (0–300 m in blue, 300–700 m in green, and above 700 m in red) for the four reforecast periods [(top left) spring, (top right) summer, (bottom left) fall, and (bottom right) winter].

hours of the Z12 reforecast runs. This visual representation of the ramps helps in comparing the diurnal distribution of the ramps between the models and observation or between two models.

The diurnal cycle of ramp events in summer (left panels of Fig. 9) has a very apparent pattern with up-ramps happening mostly from 2000 to 0200 UTC (1200–1800 LST) while down ramps are mostly concentrated between 0400 and 1400 UTC (2000–0600 LST), due to the increase and decrease in wind speed at those times that was shown previously in Fig. 6 and possibly by changes in the diurnal cycle of vertical mixing in the boundary layer during the warm months compared to fall and especially winter seasons. This pattern is perfectly mimicked in all models except for the HRRR EXP, whose down ramps in summer occur on average almost 2 h earlier. This discrepancy results in a lower skill score for this model compared to the other models for the summer season.

In contrast to the summer period that has a distinct diurnal variation in the ramp climatology, due to more stable conditions winter ramps in the right panels of Fig. 9 are distributed evenly over the daily cycle without a specific pattern. It is apparent, however, that the HRRR EXP tends to find fewer ramp events in winter compared to its CTL version and to the observation, but HRRR EXP ramps are better matched to the observed ones compared to the HRRR CTL ramps, which results in higher skill for the HRRR EXP model, as will be highlighted later in the analysis. Three particular modifications of the HRRR EXP model, namely mixing length changes, a modified horizontal diffusion, and the inclusion of small-scale gravity wave drag, can reduce the near-surface wind speed in cold pool episodes, producing more skillful cold pool forecasts (Olson et al. 2019a,b).

The skill score of the four models analyzed in this study is presented in Fig. 10 for all reforecast periods and averaged annually (left panels), for all ramps together (top-left panel), for up-ramps (middle-left panel), and for down-ramps (bottom-left panel). Most of the models have a higher ramp skill score in spring and summer compared to fall and winter. This is related to the higher number of matched ramps found in spring and summer. Also, skill scores for up-ramps are noticeably higher compared to those of down-ramps for all seasons. Two interesting cases of significant skill score differences are highlighted in Fig. 10 (in spring for up-ramp events between the HRRR EXP in blue and the HRRRNEST EXP in light-blue, due to the model resolutions difference, and in winter for down-ramp events between the HRRR CTL and the HRRR EXP, due to the differences in model physics) and will be discussed in detail below.

The two right panels of Fig. 10 show the model improvements due to the different physical parameterizations of the EXP versus the CTL runs (top-right panel) and due to the different model resolution (bottom-right panel). The only

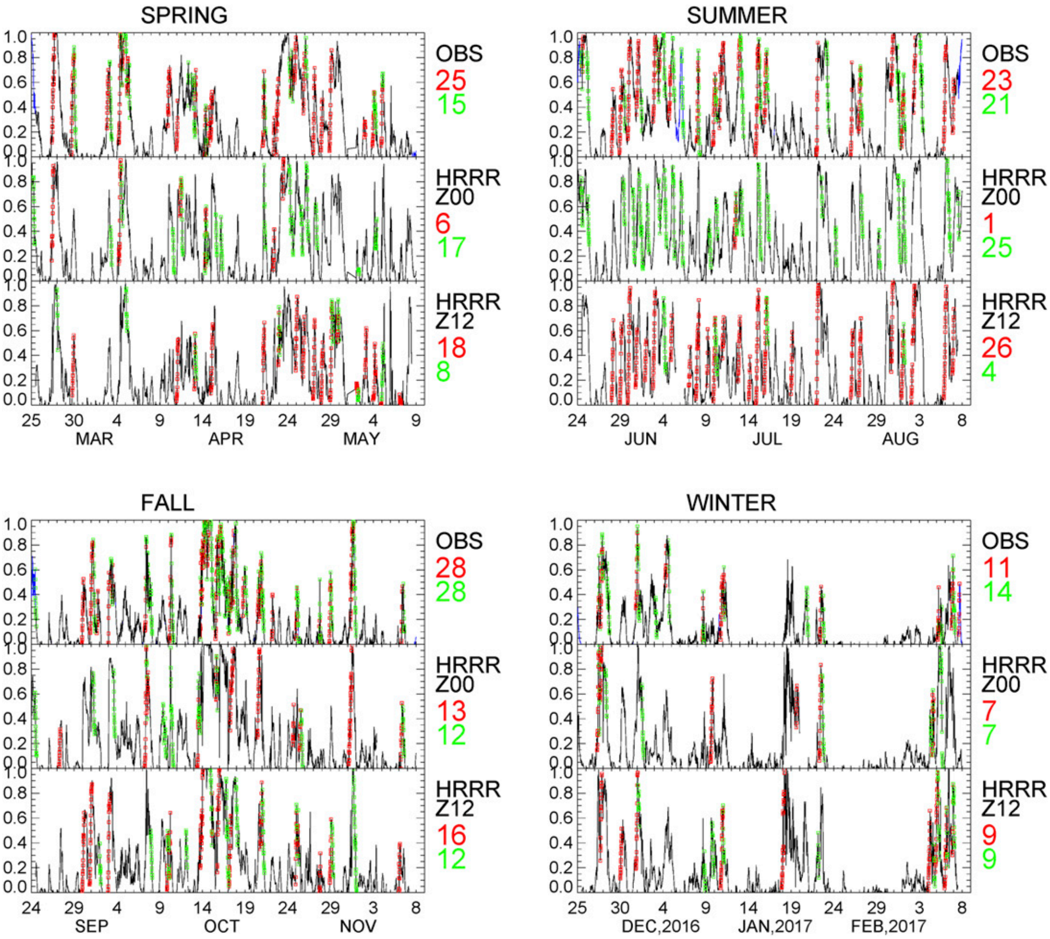

FIG. 8. Observed and modeled (HRRR CTL) time series of the aggregate normalized power for the four reforecast periods [(top left) spring, (top right) summer, (bottom left) fall, and (bottom right) winter]. Observed values are in the top panels of each of the reforecast periods and modeled ones are in the middle and bottom panels for Z00 and Z12 runs separately. Up-ramp events (in red) and down-ramp events (in green) ($\Delta p = 30\%$, and $\Delta t = 180$ min) are found only for forecast horizons 6–17 of each of the two model initialization times. The number of up- and down-ramps (in red and green, respectively) are presented by the numbers on the right side of each panel. Note that to get the number of up (or down) model forecast ramps, the number found in the Z00 initialization runs have to be added to those found in the Z12 initialization runs.

improvements due to the new physics in either model resolutions are for a much better performance of the HRRR EXP compared to the HRRR CTL in winter and a better performance of the HRRRNEST EXP compared to the HRRRNEST CTL in spring. We emphasize that the differences between the right upper panel of Fig. 10 and Fig. 14b of Olson et al. (2019a) are due to differences in the model ramp skill analysis approaches. While in Olson et al. (2019a) all available forecast hours are used for each model to compute the skill score, the approach used here only includes the skill scores from ramps centered at forecast hours 6–17, to treat the HRRR and the HRRRNEST equally. While larger improvements are expected for earlier forecast horizons, we nevertheless find a significant HRRR EXP improvement in winter for the later forecast horizons of 6–17, due to more accurate simulation of cold pool events. In stable, cold pool conditions the HRRR EXP calculates not only lower near-surface wind speeds, which reduce the wind speed bias, but also a reduced near-surface vertical wind shear (Olson et al. 2019a), both due to a drag parameterization improvement that accounts for subgrid-scale orography. Stable conditions were simulated more accurately in the HRRR EXP model (shown later) due to a

TABLE 2. Number of up- and down-ramps found in the observation and model for the four forecast periods.

		Spring	Summer	Fall	Winter
Up-ramps	Observation	25	23	28	11
	Model	24	27	29	16
Down-ramp	Observation	15	21	28	14
	Model	25	29	24	16

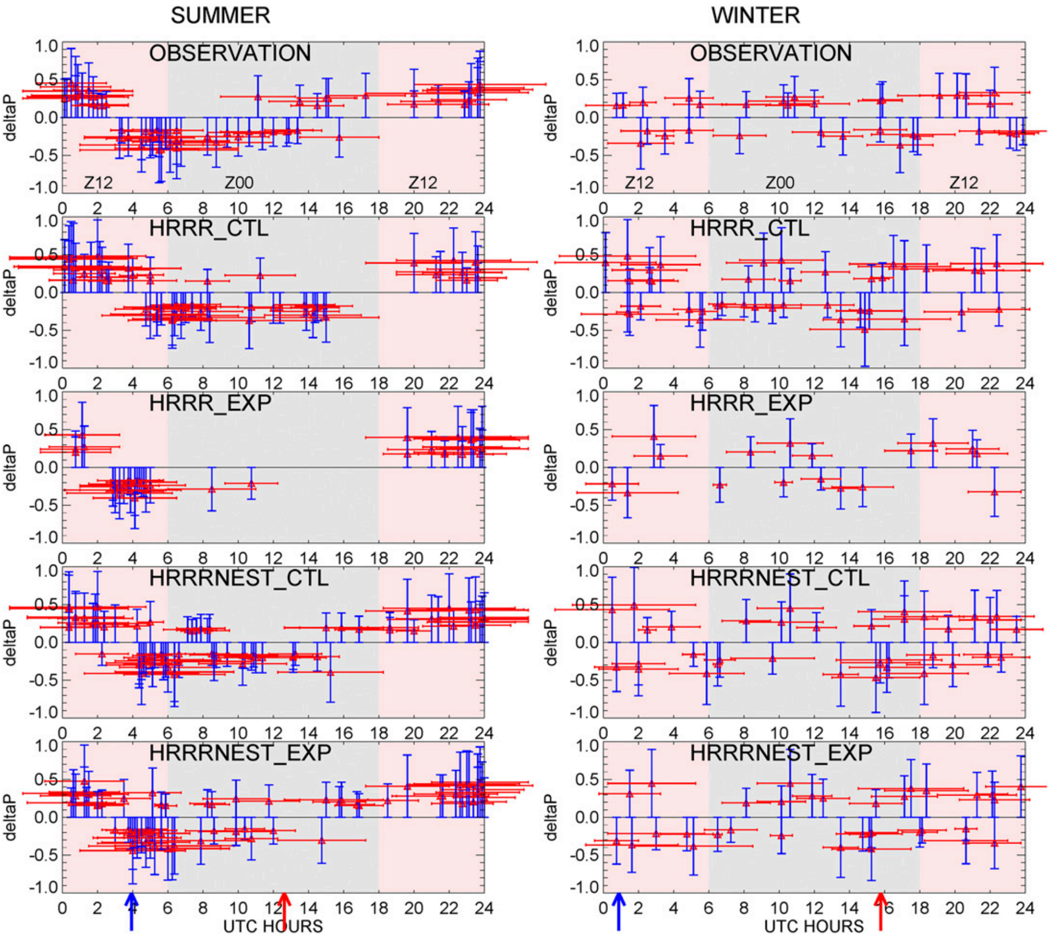

FIG. 9. Diurnal distribution of up- and down-ramps ($\Delta p = 30\%$, and $\Delta t = 180$ min) found in the (from top to bottom) aggregated observations, in the HRRR CTL run, in the HRRR EXP run, in the HRRRNEST CTL run, and in the HRRRNEST EXP run for the (left) summer and (right) winter reforecast periods. Each ramp is represented by two crossed lines, with the vertical blue line for the power change and the horizontal red line for the duration of each ramp. The central time of the ramp is the intersection of these two lines. Red and blue arrows mark sunrise and sunset, respectively.

reformulated mixing length in the revised MYNN scheme, but also could have been influenced by a better representation of the subgrid-scale clouds in the experimental model. A final improvement for cold pools came from a new numerical method that calculates horizontal diffusion on a Cartesian grid instead of using sigma coordinates. This change reduces undesired vertical mixing in regions of steep topography. All physical parameterization improvements were tested and evaluated for model resolutions > 1 km, and have not been established for model resolutions below 1 km. This explains why the HRRRNEST EXP does not show the same improvement for winter.

The skill improvement due to the resolution is more obvious: high-resolution models have higher scores compared to low-resolution models in almost all seasons. The higher skill seen in the higher-resolution simulations can come from better resolving atmospheric phenomena, by better resolving the surface characterization (terrain and land cover), or both. However, without a third set of simulations in which the coarser surface characterization is run with the higher-resolution atmospheric model, it is not possible to separate out these effects.

From the results presented in Fig. 10, one of the biggest differences in model ramp skill score is found in spring for up-ramp events, between the HRRR EXP (score equal to 0.25) and the HRRRNEST EXP (score equal to 0.34). To understand which of the three components of the skill score (errors on Δp, Δt, and Ct) contributes more to this large difference, in Fig. 11, we show the spring composite diurnal cycle of up-ramp events from the HRRR EXP (bottom-left panel), and HRRRNEST EXP (bottom-right panel) and ramps in the observations (both upper panels) for one ramp definition of $\Delta p = 30\%$ and $\Delta t = 180$ min.

Matched up-ramps in the upper/lower panels are colored in green, where the matching depends on the closeness in time of the observed and model ramp center times (Bianco et al. 2016). We note that two ramps that occur near the same time in the

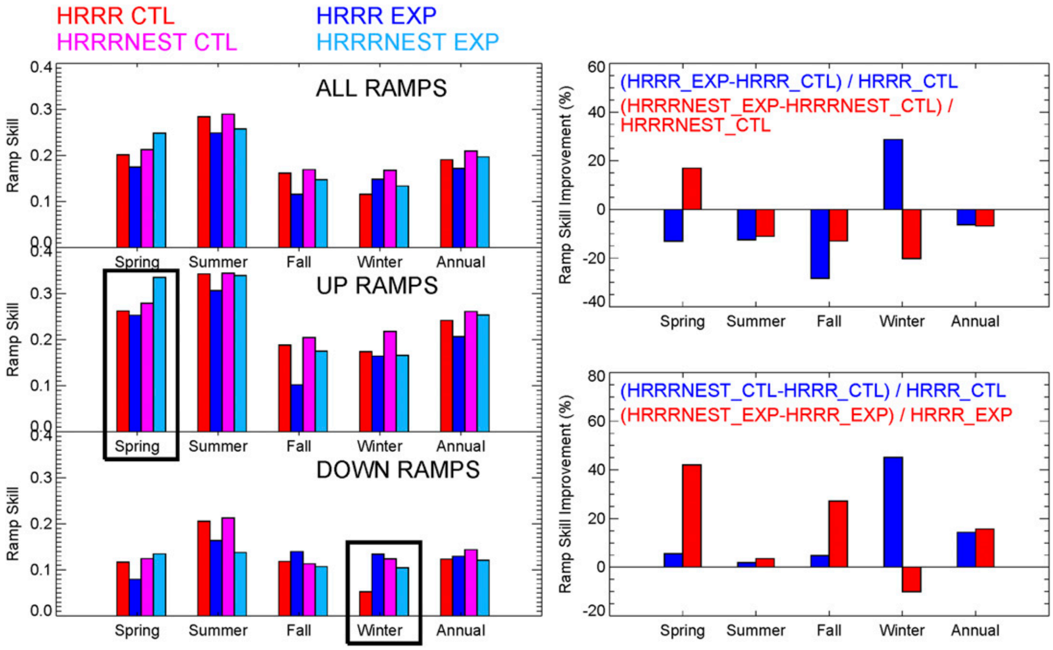

FIG. 10. (left) Skill score of the four models analyzed in this study, over the four reforecast periods, aggregating the power over all sites. (top right) Model improvements due to the physics difference between CTL and EXP runs (HRRR model is in blue, HRRRNEST in red). (bottom right) Model improvements due to the difference in resolution (the CTL runs are in blue, the EXP runs are in red).

diurnal cycle plot may not be matched because they occurred on different days. The number of all unmatched up-ramps between OBS and HRRR EXP ramps (14) is less than the number of unmatched up-ramps between OBS and HRRRNEST EXP (17), and yet the HRRRNEST EXP has a much better up-ramp skill score. This indicates that the higher skill score for the HRRRNEST EXP comes from better matched up-ramps. In fact, both models have almost the same up-ramps matched with the observation (17 for the HRRR EXP and 18 for the HRRRNEST EXP), but the average of the errors in power change (|model Δp − obs Δp|) for the matched ramps are almost twice larger for the HRRR EXP (0.18) compared to HRRRNEST EXP (0.1). This assessment is a result of our analyzing one ramp definition only, $\Delta p = 30\%$ and $\Delta t = 180$ min. When we consider all 20 ramp definitions, we found an even larger difference between the value of the averaged errors in power change of the HRRRNEST EXP (0.07) compared to the HRRR EXP (0.16), while the averaged difference in central time (|model Ct − obs Ct|) and averaged error in duration (|model Δt − obs Δt|) are very close (52 and 70 min for central time errors and 60 and 49 min for ramp duration errors for HRRR EXP and HRRRNEST EXP, respectively).

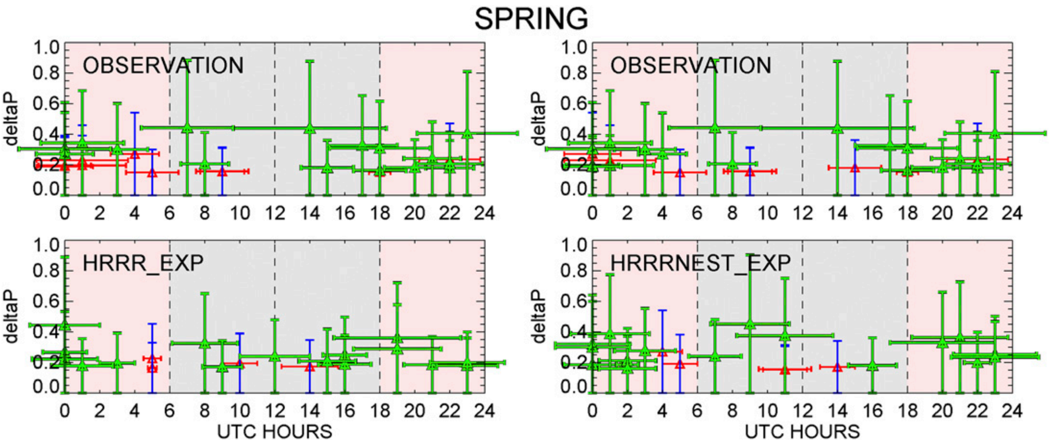

FIG. 11. Composite diurnal cycle of up-ramp events from (top) the observations, (bottom left) HRRR EXP, and (bottom right) HRRRNEST EXP for the spring reforecast period. Matched ramps are colored in green and unmatched ramps are colored in red–blue.

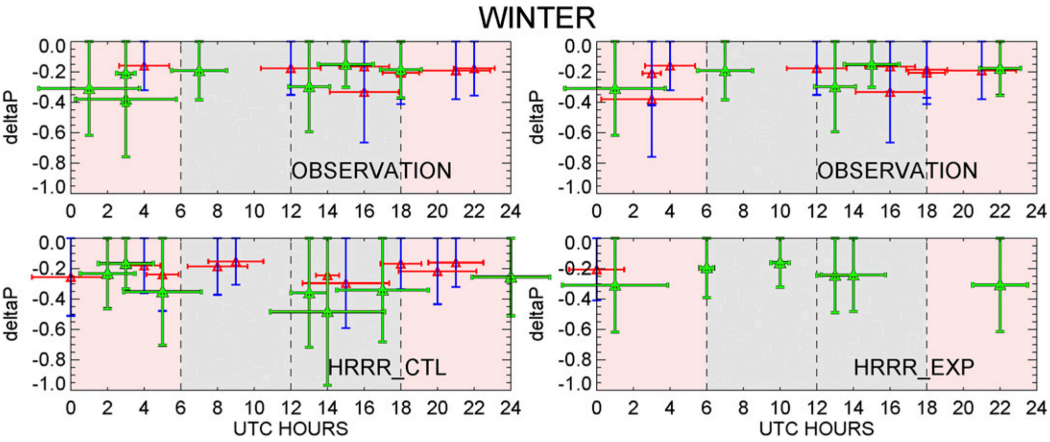

FIG. 12. Composite diurnal cycle of down-ramp events from (top) the observations, (bottom left) HRRR CTL, and (bottom right) HRRR EXP for the winter reforecast period. Matched ramps are shown in green.

Another interesting situation is the skill of the models at forecasting down-ramp events during the winter period, when the improvement in the model skill score due to the different physics (but same resolution) is evident between the HRRR CTL and HRRR EXP. First, both the observations and models for this time period have many fewer ramps compared to other seasons, as shown in Fig. 8, and second, the ramps are distributed sporadically throughout the diurnal cycle (Fig. 9). Figure 12 shows all down-ramps for the HRRR CTL on the left bottom panel and for the HRRR EXP on the right bottom panel, with the corresponding observed ramps in both top panels, for one ramp definition of $\Delta p = 30\%$ and $\Delta t = 180$ min. All matched ramps are colored in green. While the numbers of observational unmatched ramps in the top panels are almost the same (7 vs 9), the HRRR CTL has many more unmatched ramps compared to the HRRR EXP, 10 vs 1. When we analyze only the matched ramps, we find that the HRRR EXP model has smaller averaged errors in power change (0.09 vs 0.29 of the HRRR CTL) and averaged error in duration (69 min compared to 114 min of the HRRR CTL). Only the averaged difference in central time is slightly smaller for the HRRR CTL compared to the HRRR EXP, 57 min against 70 min, for a ramp definition of $\Delta p = 30\%$ and $\Delta t = 180$ min. When looking at all 20 ramp definitions, all ramp parameter estimates are better for the HRRR EXP compared to the HRRR CTL (averaged power error for HRRR CTL equals to 0.22 whereas the same power error for HRRR EXP equals to 0.08; averaged difference in central time and ramp duration errors equal to 53 and 78 min for the HRRR CTL and 48 and 50 min for the HRRR EXP), therefore providing altogether a higher score for the HRRR EXP for wintertime ramps.

Finally, to explain and demonstrate the better performance of the HRRR EXP model compared to other models in winter, in Fig. 13 we show the time series of the aggregated power from observations and from the four models, from both the Z00 and Z12 runs with the skill score of all ramps for two days in 4–5 February 2017.

These days are chosen because they present a rapid power change at the end of a long cold pool event (McCaffrey et al. 2019). Observed data are the same in all four columns. Both experimental models show better agreement with the observational time series compared to the control runs (second and fourth columns), but the HRRR EXP time series is much closer to reality. The HRRR EXP is the only model predicting ramps at the right time and with an almost perfect power change and duration (producing the highest skill score), while the other models forecast more unmatched ramps prior to the end of the cold pool event, due to their tendency to erode the cold pool earlier (Wilczak et al. 2019b).

6. Ramp statistics for bias-corrected data

As shown in Bianco et al. (2019), all models utilized in this study have wind speed biases that vary diurnally, with larger values at night and smaller values during daytime. After testing several ways to correct the models for their biases, seasonally, monthly, or diurnally, we found that a diurnal bias-correction results in the highest ramp skill scores. To apply this diurnal bias-correction, we calculated the power bias at each site, for each reforecast period and at each hour of the diurnal cycle, and removed it from the model output.

Figure 14 shows the statistical results for ramp skill score calculated on the diurnally bias-corrected data. This correction improves the ramp skill score for most of the models/seasons by increasing the skill score up to 17% with annual improvements of ~10%. The comparison of the original and bias-corrected ramp skill scores (hatched boxed versus solid color boxes in Fig. 14) shows a noticeable increase for the bias-corrected HRRR models ramp skill, both control or experimental, and much less of a change in ramp skill for the high-resolution models, especially for the HRRRNEST EXP, because the high-resolution models have much smaller biases compared to the low-resolution models. These ramp skill results are in an agreement with 80-m biases found in Bianco et al. (2019).

7. Summary and conclusions

The Ramp Tool and Metric was applied to forecasts of 80-m wind speed power data collected by 19 sodars and three lidars

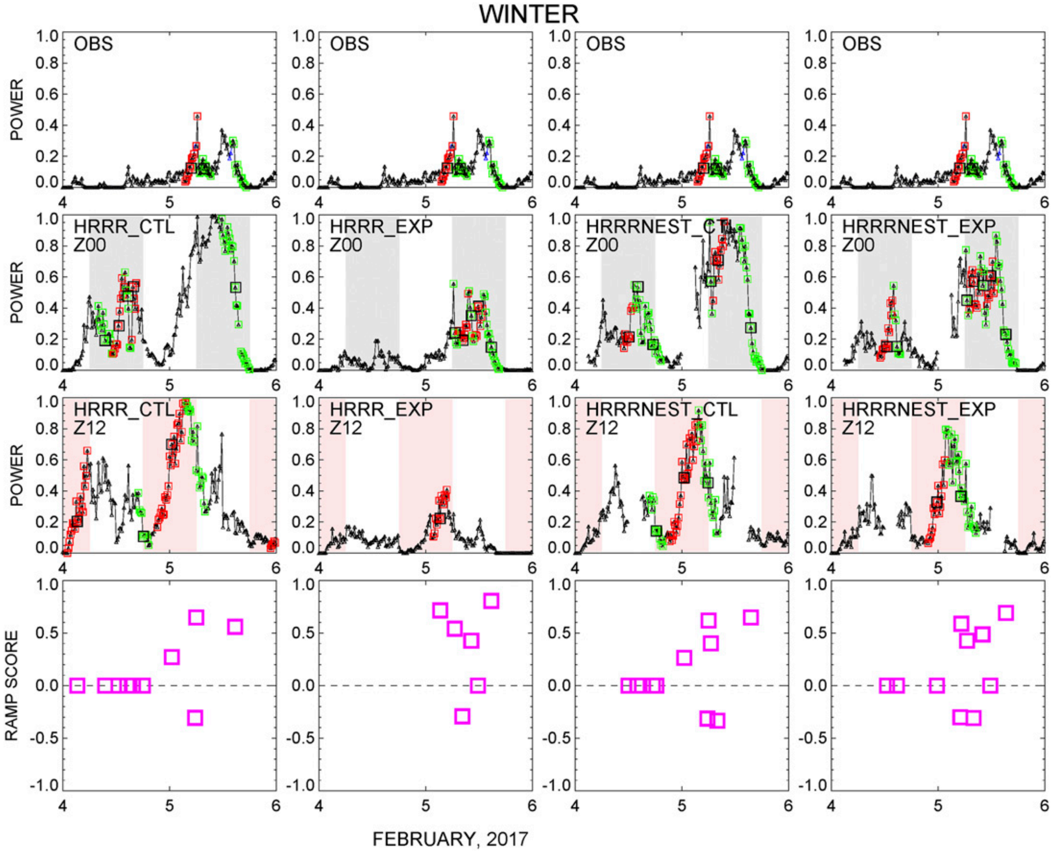

FIG. 13. As in Fig. 5, but for two days on 4–5 Feb 2017. Four columns of panels represent comparison of observed data and model data for forecast hours Z00 and Z12 from left to right for HRRR CTL, HRRR EXP, HRRRNEST CTL, and HRRRNEST EXP.

and the NWP HRRR and HRRRNEST models during the second Wind Forecast Improvement Project (WFIP2). The two models were chosen due to their small horizontal grid spacing, 3 km and 750 m, respectively, which is appropriate for an area of complex terrain such as that of the WFIP2 campaign. The two models were run in control and experimental configurations, with the experimental version including all the improvements introduced to the model physics during the campaign, summarized in Olson et al. (2019a,b). The CTL and EXP reforecast runs spanned four 6-week reforecast periods, one for each season, to test the dependence of the improvements on the season. The 80-m wind speeds measured by the 19 sodars, three lidars, and converted into power using a generic/normalized power curve, agreed well with the actual wind power generation data provided by the Bonneville Power Administration. The observational dataset was then used to quantify the seasonal and diurnal variability of wind speed and ramp events.

Due to the configuration of the reforecast runs, we calculated ramp skill over nonoverlapping 12-h blocks from consecutive simulations initialized at 0000 and 1200 UTC. The "forecast method" of the RT&M was adopted, which compares each reforecast model run independently with the corresponding observations, but was used to calculate ramp skill score only from forecast horizons 6–17 of each reforecast run to reduce the impact of truncated ramps at the beginnings and ends of the forecasts. Therefore, this result is only in qualitative but not quantitative agreement with previous ramp skill model estimation shown in Olson et al. (2019a) where all model forecast horizons were used for the ramp skill score calculation.

The main findings are:

- Spring and summer show a well-defined diurnal trend of wind speed, faster during the daytime and slower at night. This diurnal trend was not observed in fall and winter. As a consequence, in spring and summer, more up-ramp events are found in the afternoon and evening, while more down-ramps are found shortly before and after midnight. In contrast, the number of up- and down-ramps for fall and winter are equally distributed over the entire diurnal cycle.
- We introduced a new visualization for ramps, using two lines, a vertical one for power change and a horizontal one for time duration, crossed at the ramp central time. This visual tool helps to recognize patterns in the diurnal distribution of ramp events and enables comparison of distributions between observed and model data.
- Results from the RT&M over aggregated data showed higher scores for spring and summer compared to fall and

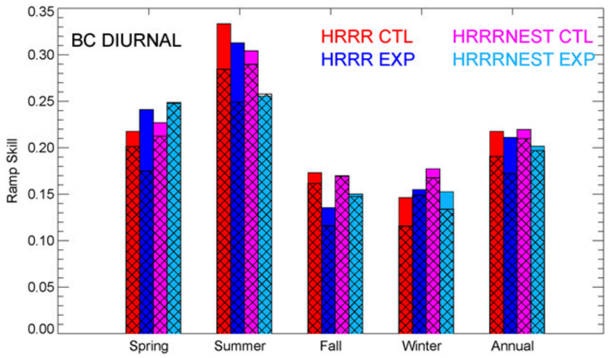

FIG. 14. Skill score of the four bias-corrected model runs for the four reforecast periods and annually averaged, aggregating the power over all sites. Hatched bars are the data without bias correction, the same as in the top-left panel of Fig. 10.

winter. The modified physics in the EXP runs produces better skill scores only for winter, while the higher resolution of the model improves the skill in almost all seasons except winter for the HRRRNEST EXP. For the winter period the HRRR EXP has fewer ramps compared to other models, but they are better matched to the observed ramps.

- The skill score improvements due to the resolution differences are almost always positive showing that the high-resolution model better resolves rapid changes in wind speed in the complex terrain.
- Our analysis also allowed us to analyze the contribution from the three terms of the ramp skill equation separately to find out why one model is better than another for certain reforecast periods.
- The diurnally bias-corrected models show higher ramp skill scores for almost all models/seasons compared to the uncorrected models, with up to 17% ramp score improvement for a single seasonal reforecast period and ~10% ramp skill score improvement averaged annually, indicating that simply improving the model's mean diurnal cycle will significantly improve ramp forecasts.
- Consistent with what was found during the first WFIP, more up-ramp events are found in the time series of both observed and forecast rated power than down-ramp events, (in this study this is especially true for spring and summer). Also, the skill score is always higher for up-ramps compared to down-ramps. The reason for this is an area for future research.

Before the end of the WFIP2 field campaign a model freeze was imposed to perform the four reforecast runs on which our analysis is based. Several other studies, already referenced throughout of the manuscript, were performed on these same reforecast runs to measure the impact of the improved parameterizations over different meteorological aspects, providing further insight on how the model parameterizations can be improved even further. On the basis of all these results, other improvements and additions to the parameterizations are currently under way, focusing on better simulating the diurnal cycle, which we believe will help to improve ramp prediction. These include adding momentum transport to help with the daytime nonlocal mixing, and seeking "missing" heating terms such as compensational environmental subsidence associated with nonprecipitating convective updrafts and heating due to the dissipation of TKE. If successful, these will be added to future versions of the models.

Acknowledgments. We thank all the engineers, scientists, and students involved in WFIP2 for site selection, leases, instrument deployment and maintenance, data collection, and data quality control. Funding for this work was provided by the DOE, Office of Energy Efficiency and Renewable Energy, Wind Energy Technologies Office, and by the NOAA/ESRL Atmospheric Science for Renewable Energy program. This work was authored (in part) by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract DE-AC36-08 GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. government. The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. government purposes.

Data availability statement. The operational HRRR model is not entirely open source (data assimilation, cycling scripts, etc.), but updates to the model parameterizations used in the HRRR are deposited periodically to the official repository for the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model, maintained by the National Center for Atmospheric Research (NCAR), which is open source (https://github.com/wrf-model/WRF, last access: 4 April 2020). A branch from this repository was created for WFIP2 testing, based on WRF-ARWv3.9. This branch is currently stored at zenodo (Olson and Kenyon 2019). This branch is no longer under development and all improvements have been transferred to NCAR's official repository. All datasets used in this study are freely available to the public from the DOE Data Archive and Portal (DAP; https://a2e.energy.gov/projects/wfip2, last accessed 4 April 2020). Please contact the corresponding author for additional details, if needed.

REFERENCES

Akish, E., L. Bianco, I. V. Djalalova, J. M. Wilczak, J. Olson, J. Freedman, C. Finley, and J. Cline, 2019: Measuring the impact of additional instrumentations on the skill of numerical weather prediction models at forecasting wind ramp events during the first Wind Forecast Improvement Project (WFIP). *Wind Energy*, **22**, 1165–1174, https://doi.org/https://doi.org/10.1002/we.2347.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Bianco, L., I. V. Djalalova, J. M. Wilczak, J. Cline, S. Calvert, E. Konopleva-Akish, C. Finley, and J. Freedman, 2016: A wind energy ramp tool and metric for measuring the skill of numerical weather prediction models. *Wea. Forecasting*, **31**, 1157–1156, https://doi.org/10.1175/WAF-D-15-0144.1.

——, and Coauthors, 2019: Impact of model improvements on 80 m wind speeds during the second Wind Forecast Improvement Project (WFIP2). *Geosci. Model Dev.*, **12**, 4803–4821, https://doi.org/10.5194/gmd-12-4803-2019.

Bossavy, A., R. Girard, and G. Kariniotakis, 2010: Forecasting uncertainty related to ramps of wind power production. *Proc. European Wind Energy Conf. and Exhibition*, Warsaw, Poland, European Wind Energy Association, 10 pp., https://hal-mines-paristech.archives-ouvertes.fr/hal-00765885/document.

Cutler, N., M. Kay, K. Jacka, and T. S. Nielsen, 2007: Detecting, categorizing and forecasting large ramps in wind farm power output using meteorological observations and WPPT. *Wind Energy*, **10**, 453–470, https://doi.org/10.1002/we.235.

Ferreira, C., J. Gamma, L. Matias, A. Botteud, and J. Wang, 2011: A survey on wind power ramp forecasting. Argonne National Laboratory Rep. ANL/DIS-10-13, 28 pp., http://ceeesa.es.anl.gov/pubs/69166.pdf.

Gallego-Castillo, C., A. Cuerva-Tejero, and O. Lopez-Garcia, 2015: A review on the recent history of wind power ramp forecasting. *Renewable Sustainable Energy Rev.*, **52**, 1148–1157, https://doi.org/10.1016/j.rser.2015.07.154.

Greaves, B., J. Collins, J. Parkes, and A. Tindal, 2009: Temporal forecast uncertainty for ramp events. *Wind Eng.*, **33**, 309–319, https://doi.org/10.1260/030952409789685681.

Grimit, E. P., 2020: The Second Wind Forecast Improvement Project (WFIP2) decision support tools. *11th Conf. on Weather, Climate, and the New Energy Economy*, Boston, MA, Amer. Meteor. Soc., 13.1, https://ams.confex.com/ams/2020Annual/webprogram/Paper367280.html.

Kamath, C., 2010: Understanding wind ramp events through analysis of historical data. *IEEE PES Transmission and Distribution Conf. and Expo 2010*, New Orleans, LA, IEEE, 1–6, https://doi.org/10.1109/TDC.2010.5484508.

McCaffrey, K., and Coauthors, 2019: Identification and characterization of persistent cold pool events from temperature and wind profilers in the Columbia River basin. *J. Appl. Meteor. Climatol.*, **58**, 2533–2551, https://doi.org/10.1175/JAMC-D-19-0046.1.

Olson, J. B., and J. S. Kenyon, 2019: joeolson42/WFIP2: WFIP2 Experimental HRRR version 1.0. Zenodo, accessed 4 April 2020, https://doi.org/10.5281/zenodo.3369984.

——, and Coauthors, 2019a: Improving wind energy forecasting through numerical weather prediction model development. *Bull. Amer. Meteor. Soc.*, **100**, 2201–2220, https://doi.org/10.1175/BAMS-D-18-0040.1.

——, J. S. Kenyon, W. M. Angevine, J. M. Brown, M. Pagowski, and K. Sušelj, 2019b: A description of the MYNN-EDMF scheme and coupling to other components in WRF-ARW. NOAA Tech. Memo. OAR GSD, 61, 37 pp., https://doi.org/10.25923/n9wm-be49, https://repository.library.noaa.gov/view/noaa/19837.

Pichugina, Y. L., and Coauthors, 2019: Spatial variability of winds and HRRR–NCEP model error statistics at three Doppler-lidar sites in the wind-energy generation region of the Columbia River basin. *J. Appl. Meteor. Climatol.*, **58**, 1633–1656, https://doi.org/10.1175/JAMC-D-18-0244.1.

Sharp, J., and C. F. Mass, 2002: Columbia Gorge gap flow—Insights from observational analysis and ultra-high-resolution simulation. *Bull. Amer. Meteor. Soc.*, **83**, 1757–1762, https://doi.org/10.1175/BAMS-83-12-1757.

Shaw, W., and Coauthors, 2019: The Second Wind Forecast Improvement Project (WFIP 2): General overview. *Bull. Amer. Meteor. Soc.*, **100**, 1687–1699, https://doi.org/10.1175/BAMS-D-18-0036.1.

Western Area Power Administration, 2018: Protecting industrial control systems. Western Area Power Administration, 13 pp., https://www.wapa.gov/About/the-source/Documents/2018-2-14_ARC_Industry_Forum.pdf.

Wilczak, J. M., and Coauthors, 2015: The Wind Forecast Improvement Project (WFIP): A public-private partnership addressing wind energy forecast needs. *Bull. Amer. Meteor. Soc.*, **96**, 1699–1718, https://doi.org/10.1175/BAMS-D-14-00107.1.

——, and Coauthors, 2019a: 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*, **22**, 932–944, https://doi.org/10.1002/we.2332.

——, and Coauthors, 2019b: The Second Wind Forecast Improvement Project (WFIP2): Observational field campaign. *Bull. Amer. Meteor. Soc.*, **100**, 1701–1723, https://doi.org/10.1175/BAMS-D-18-0035.1.

Worsnop, R. P., M. Scheuerer, T. M. Hamill, and J. K. Lundquist, 2018: Generating wind power scenarios for probabilistic ramp event prediction using multivariate statistical post-processing. *Wind Energy Sci.*, **3**, 371–393, https://doi.org/10.5194/wes-3-371-2018.

Yang, Q., L. K. Berg, M. Pekour, J. D. Fast, R. K. Newsom, M. Stoelinga, and C. Finley, 2013: Evaluation of WRF-predicted near-hub-height winds and ramp events over a Pacific Northwest site with complex terrain. *J. Appl. Meteor. Climatol.*, **52**, 1753–1763, https://doi.org/10.1175/JAMC-D-12-0267.1.

Zack, J. W., S. Young, J. Nocera, J. Aymami, and J. Vidal, 2010: Development and testing of an innovative short-term large wind ramp forecasting system. *Proc. European Wind Energy Conf. and Exhibition*, Warsaw, Poland, European Wind Energy Association.

Zhang, J., A. Florita, B.-M. Hodge, and J. Freedman, 2014: Ramp forecasting performance from improved short-term wind power forecasting. *Proc. ASME Int. Design Engineering Technical Conf. and Computers and Information in Engineering Conf.*, DETC2014-34775, Buffalo, NY, ASME, https://doi.org/10.1115/DETC2014-34775.