A Wind Energy Ramp Tool and Metric for Measuring the Skill of Numerical Weather Prediction Models

LAURA BIANCO,[a] IRINA V. DJALALOVA,[a] JAMES M. WILCZAK,[b] JOEL CLINE,[c]
STAN CALVERT,[d] ELENA KONOPLEVA-AKISH,[e] CATHY FINLEY,[f,h]
AND JEFFREY FREEDMAN[g,i]

[a] *University of Colorado/Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado*
[b] *NOAA/Earth Systems Research Laboratory, Boulder, Colorado*
[c] *Office of Energy Efficiency and Renewable Energy, Department of Energy, Washington, D.C.*
[d] *SC Energy Consulting, Washington, D.C.*
[e] *Science and Technology Corporation, Boulder, Colorado*
[f] *WindLogics Inc., St. Paul, Minnesota*
[g] *AWS Truepower, Albany, New York*

(Manuscript received 22 October 2015, in final form 21 March 2016)

ABSTRACT

A wind energy Ramp Tool and Metric (RT&M) has been developed out of recognition that during significant ramp events (large changes in wind power Δp over short periods of time Δt) it is more difficult to balance the electric load with power production than during quiescent periods between ramp events. A ramp-specific metric is needed because standard metrics do not give special consideration to ramp events and hence may not provide an appropriate measure of model skill or skill improvement. This RT&M has three components. The first identifies ramp events in the power time series. The second matches in time forecast and observed ramps. The third determines a skill score of the forecast model. This is calculated from a utility operator's perspective, incorporates phase and duration errors in time as well as power amplitude errors, and recognizes that up and down ramps have different impacts on grid operation. The RT&M integrates skill over a matrix of ramp events of varying amplitudes and durations.

1. Introduction

One challenge in integrating weather-dependent renewable energy onto the electric grid is the temporal variability of the wind or solar resource. For wind, this variability is amplified by a wind turbine's nonlinear power curve that translates wind speed into power, creating large variations of wind energy production over short periods of time. Figure 1 (top panel) displays the time series of wind speed measured on a tall tower at 80 m above ground level (AGL), and the resulting wind power (bottom panel) produced by a turbine using a standard International Electrotechnical Commission (IEC) class 2 (International Electrotechnical Commission 2007) turbine power curve (center panel). These data were collected in South Dakota, a region featuring a large amount of wind energy production, where the class 2 turbine is the most common type of wind turbine deployed.

The wind power production in Fig. 1 has extended periods of time with either zero power production (for speeds below the turbine's cut-in speed, 3 m s^{-1}) or near 100% of its capacity for high speeds (between 13 and 25 m s^{-1}), with frequent jumps between small and large power values. These jumps, or ramp events, can be large because of the wind power increasing approximately as the cube of the wind speed in the middle portion of the turbine's power curve. Ramp

Denotes Open Access content.

[h] Current affiliation: Department of Earth and Atmospheric Sciences, Saint Louis University, St. Louis, Missouri.
[i] Current affiliation: Atmospheric Sciences Research Center, University of Albany, State University of New York, Albany, New York.

Corresponding author address: Dr. Laura Bianco, University of Colorado/CIRES, 325 Broadway, MS PSD, Boulder, CO 80305.
E-mail: laura.bianco@colorado.edu

DOI: 10.1175/WAF-D-15-0144.1

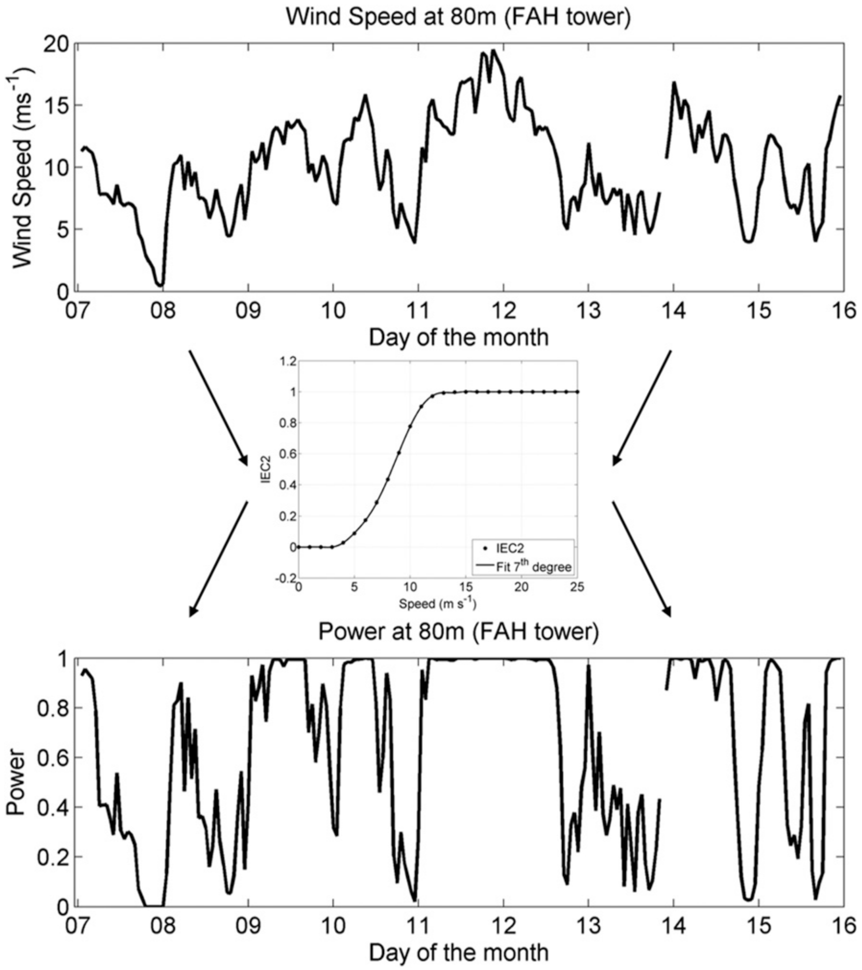

FIG. 1. (top) Time series of 80 m AGL wind speed (m s^{-1}) and (bottom) equivalent normalized capacity power using (center) an IEC2 wind power curve. Data from the SDSU FAH tall tower for the 9-day period, 7–15 Jan 2012.

events are important in real-time grid operations. If large changes in wind power production occur, grid operators must keep the grid in balance by making equally large and abrupt changes in conventional energy generation. Large and sudden changes in conventional generation can be costly and problematic (Bradford et al. 2010; Francis 2008), especially if the changes are not forecast accurately, both in terms of their amplitude and timing.

Standard metrics (e.g., mean absolute error and root-mean-square error) may not be well suited for wind energy forecast evaluation because wind power production can be near constant for considerable periods of time, especially near 100% or 0%; it is the periods of rapid transition between those two states that are most challenging for grid balancing. Therefore, a wind ramp metric can provide a useful statistical measure of the accuracy of the model at forecasting ramp events by weighting the model agreement for these events more than during periods of near-constant power. A ramp metric can be used to compare the skill of two models at forecasting ramp events, or for documenting progress in improving a given model. It could also be used to analyze the climatology of wind ramp events (i.e., seasonal means or interannual variations) from observations or models.

The identification and forecast evaluation of ramp events has similarities and differences with other meteorological phenomena, such as precipitation [rate of change; Hamill (2014)], aviation forecasting [timing; Isaac et al. (2014); Jacobs and Maat (2005); Wong et al. (2013)], air quality indices and severe convection (amplitude), and floods or droughts (direction of change). In addition, the response of a grid operator to a down ramp may be different than for an up ramp as curtailing output for up-ramp events may be easier than quickly bringing on additional generation for down ramps. The combination of rate of change, timing, and the directionality of ramp events makes for a unique forecasting problem. For an agency such as NOAA, the Ramp Tool and Metric (hereafter referred to as RT&M) can be useful for determining how potential changes to research or operational weather prediction models will impact users in the wind energy industry.

Despite the importance of ramp events for renewable energy, there is no commonly accepted definition of a ramp event, nor is a single strict threshold possible, because the threshold at which a ramp becomes important will vary from user to user, and from situation to situation (e.g., Zack et al. 2011). In this study we aim to develop a ramp tool that has the flexibility to be helpful for a variety of users and that can easily be modified or tuned to be valuable in a variety of situations. For models, the RT&M is most applicable as a diagnostic tool, for evaluating the skill of a long time series of forecasts, and was not designed or intended to be used for making real-time decisions.

The RT&M described here has three components. The first is the identification of ramp events in a time series of power data, for which several different methodologies are employed and compared. The second component of the RT&M matches observed ramp events with those predicted by a forecast model. If ramp events are defined such that they are rare events, matching is relatively simple. However, when the definition is relaxed so that ramp events become more frequent, matching events can be more difficult (Wandishin et al. 2014). The final component of the RT&M is a methodology for scoring the ability of a model to forecast ramp events. We develop a scoring metric that accounts for phase, duration, and amplitude errors in the forecast, and differentiates between the impacts of up- and down-ramp events. The particular scoring rules that we use are intended to reflect the perspective of a grid operator; however, the metric itself is flexible so that it could be easily modified to reflect the needs of other users. To test the RT&M, we used the 13-km horizontal resolution National Oceanic and Atmospheric Administration/Earth System Research Laboratory (NOAA/ESRL) Rapid Refresh (RAP) model and tall-tower anemometer observations that were collected during the Wind Forecast Improvement Project (WFIP), which took place in the U.S. Great Plains during 2011–12 (Wilczak et al. 2014, 2015). The RT&M results are applied to data collected over 9 days (7–15 January 2012) from a set of four 80-m-tall towers spanning an ~100 km × 100 km area, and to forecasts from the RAP model at the same locations. The RT&M is available online (http://www.esrl.noaa.gov/psd/products/ramp_tool/).[1]

This paper is organized as follows. Section 2 presents three different methods for defining ramp events and compares their results. Section 3 discusses the issues related to ramp matching. Forecast scoring and model evaluation procedures are presented for one ramp definition in section 4. The forecast skill score methodology is then extended for a range of ramp definitions in section 5. Section 6 provides a summary and discussion. In the appendix we discuss rules for applying a bonus to the model skill when excess energy production can be reduced through wind plant curtailment.

2. Ramp identification

As mentioned, there is no absolute definition of a ramp as the definition will depend on the particular application, and the application will change from user to user (Freedman et al. 2008). Also, users may require multiple definitions of ramps to be operative simultaneously. For example, if a ramp event has a 60% capacity change in generation over a 4-h period, it could inform a utility that a certain type of unit needs to be brought online. However, if within that 4-h period there is an embedded ramp with a 30% of capacity change in only 15 min, then a different type of unit may need to be brought online for that 15-min period within the 4-h duration ramp. A ramp metric must address a matrix of time and amplitude scales simultaneously to give a robust measure of model skill.

The type of power time series used in this tool may differ depending on the user (i.e., for a wind plant operator interested in forecast skill for a single plant or turbine, the time series considered would be the power production from that plant or turbine; for a grid operator concerned with the aggregate power generated by wind and solar over their entire balancing area, the time series to be considered would be the aggregate power). We assume that the time series in either case is the basis upon which operational decisions are made, so there is no reason to filter or modify the time series for the

[1] The RT&M is coded in Matlab, and users can download the main code, functions and instructions, and the data used for this study to test the RT&M, before running it on their own datasets. When the RT&M is run, a GUI opens and the user can choose several options, some of which will be introduced in the subsequent narrative of the paper, and the others are explained in a readme file, downloadable with the RT&M. We also created an executable version of the code for users that do not have Matlab. In this case they will not be able to modify the Matlab code, but can still use the GUI. When a user runs the RT&M, he or she can select whether the input data will be wind speed or power, according to what type of data are available.

analysis of ramp events. Data compression routines such as the swinging door algorithm (Bristol 1990; Florita et al. 2013; Zhang et al. 2014) reduce a time series to a shorter series of linear segments, and smaller changes in power are ignored as noise with the effect that the filtered time series will not contain the full range of power variations originally present. Also, although we do not apply any bias correction to the model output, it is possible for a user to apply postprocessing techniques before inputting their data into the RT&M.

Three ramp identification methods are presented in this section. Each is tested on 9 days of observations from four South Dakota State University (SDSU) 80-m-tall towers [Faith (FAH): latitude 45.0539°, longitude −102.2630°, altitude 797 m; Long Valley (LVL): latitude 43.4331°, longitude −101.5544°, altitude 944 m; Lowry (LWY): latitude 45.2772°, longitude −99.9861°, altitude 663 m; and Reliance (REL): latitude 43.9681°, longitude 99.5944°, altitude 628 m] and on forecasts for the same location.

Forecasts were generated by the hourly updated RAP model (http://rapidrefresh.noaa.gov). For comparison purposes the model forecast values at the tower location were determined through a horizontal parabolic interpolation of the 16 model grid points surrounding the tower location, followed by linear vertical interpolation of the model wind profiles to the tower instrument height. The RAP model provided output at 15-min intervals, while the SDSU tower data were available as 10-min averages. Time interpolation is discussed below.

The first step of the ramp identification process is to generate equal-length time series of model forecast and observational wind speed data. For this we take two different approaches that we will refer to as the stitching method and the independent forecast run method.

For the stitching method we create a time series of model forecasts for a particular forecast horizon. First, consider the simple case where the dataset consists of hourly model output, hourly observations, and model forecasts that are initialized on an hourly basis. A time series of forecasts for a fixed forecast horizon (say forecast hour 3) is created by concatenating all 3-h forecasts over a length of time t, where t is greater than the maximum length forecast (15 h for the RAP), and equal in length to the considered observed time series (here, t is 9 days, or 216 h).

The WFIP observed wind speeds have 10-min resolution while the model output has 15-min granularity, and the model initialization cycle is hourly. To make use of the high temporal resolution data for detecting ramps, the process described above is modified by extracting the four sequential 15-min forecasts that begin at a given forecast horizon hour, and then concatenating these groups of four forecasts. This time series of 15-min forecasts of similar forecast horizon values is then linearly interpolated to the 10-min intervals of the observations, and both time series are converted into power using the IEC2 turbine power curve.

For the independent forecast runs method we proceed using sets of individual forecast runs, not concatenated, in our case each with a length of 15 h, and to compare each individual forecast run against the corresponding observational time series. The advantage of this approach is that it does not have the potential to create artificial ramps through the stitching process, while it has the disadvantage of increasing the relative number of occurrences of ramps that are terminated at the start and end of each forecast run, as well as other disadvantages that will be discussed later.

We include both of these approaches in the RT&M, allowing the user [through the graphical user interface (GUI)] to choose the one that best addresses their analysis needs.

For both the stitching method and the independent forecast runs method, ramps are then identified within the two corresponding model and observational time series using one of the ramp detection methods described below.

a. Fixed-time interval method

Various methods for defining ramp events have been proposed (Cutler et al. 2007; Greaves et al. 2009; Kamath 2010; Zack et al. 2010; Bossavy et al. 2010; Ferreira et al. 2011), and the methods that we employ include aspects of these earlier studies. Ramps of different sign are recorded separately, so at the beginning two identical time series of logical "no ramp" values are initialized. The first of these time series will record only "up" events and the second only "down" events. The first ramp identification method, referred to as the fixed-time interval method, uses a sliding time window of length WL, over which we measure the change in power. This method tests if the difference in power $\Delta p = (p_s - p_e)$ between the starting and ending points in the time window WL equals or exceeds a threshold value Δp_{RD}, where Δp_{RD} is the ramp definition threshold, and p_s and p_e are the power values at the starting and ending points in the time window, respectively. If the threshold criterion is met, a ramp exists. If a ramp exists and $p_s < p_e$, the event is an up ramp; if $p_s > p_e$, it is a down ramp. If an up event is found in the window WL, then all points of the up time series within this window are changed from no ramp to up ramps, and if a down event is found in the window WL, then all points of the down time series within this window are changed to down ramps. Once a value is set as an up or down ramp, it cannot be changed during the rest of the process. The sliding window moves forward one time step, 10 min, and the process is repeated

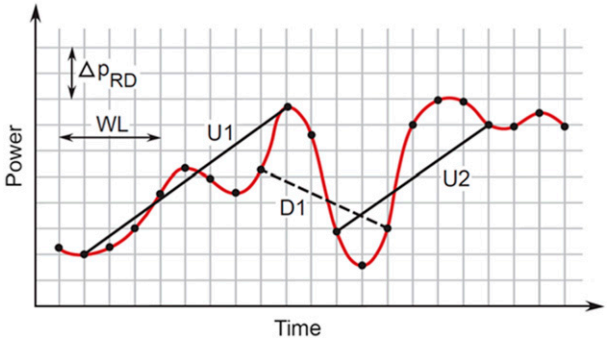

FIG. 2. Ramps identified by the fixed-time interval method (up ramps, solid; down ramps, dashed line) for a window length WL and a ramp threshold Δp_{RD}.

until the end of the time series is reached. At the end of the process each point in the up time series will be marked as a no ramp or an up ramp, and each point in the down time series will be marked as a no ramp or a down ramp. Any contiguous time steps marked as being part of up events are concatenated into a single up event, and the same is done for the time series of down events. The ramp event can be longer than the window length WL, but each point in the event belongs to a window of length WL that satisfies the criterion $|\Delta p| \geq \Delta p_{\mathrm{RD}}$.

Each concatenated ramp event is defined by its center time Ct, temporal duration Δt, and total power change $\Delta p = (p'_s - p'_e)$, where p'_s and p'_e are the power values at the start and end, respectively, of the concatenated ramp event. The schematic power time series in Fig. 2 shows a series of ramp events that would be detected by the fixed-time interval method.

Although appealing because of its simplicity, this method has the possible drawbacks that 1) the selected ramp events may not intuitively look like ramps, since larger values of Δp can occur within the ramp than those defined by its end points, and 2) two ramp events of opposite direction can be overlapping in time. These overlapping events could be truncated; however, since the fixed-time interval method is frequently used without truncation (Kamath 2010), these overlapping events are allowed to remain. The truncation of overlapping events will be implemented in method 3.

An example of ramp identification using the fixed-time interval method on observed and stitched modeled data (for the model initialization time, hour 0) is displayed in Fig. 3 for the 9 days of aggregate power data from the four SDSU towers.[2]

Using a ramp definition of a power change greater than 40% over a nominal 2-h period, seven up ramps (shown in red) and two down ramps (in green) are found through the time series of the observations. In contrast, due to the smoothness of the forecasted time series of power, the model finds only four up and two down ramps. Since Fig. 3 shows the time series of the aggregate power, the changes in power are smoother that those we would see if we considered the time series of the power for one tower only. For this reason, in this example there are no overlapping points of opposite-signed ramps.[3]

b. *Minimum–maximum method*

The next approach, referred to as the minimum–maximum (min–max) method, avoids the two problems previously noted for the fixed-time interval method. This technique finds the maximum amplitude change in power $\Delta p = (p_{\max} - p_{\min})$ within a sliding window of length WL, where $p_{\max}$ and $p_{\min}$ are the maximum and minimum power values within that window. If this change in power amplitude meets the criterion $\Delta p \geq \Delta p_{\mathrm{RD}}$, where Δp_{RD} is the ramp definition threshold, then a ramp event occurs. If more than one pair of points within the window meets the threshold criterion, only the shortest time Δt is used.

The initial ramp duration is determined by the times $t_{\min}$ and $t_{\max}$ that correspond to $p_{\min}$ and $p_{\max}$, so $\Delta t = |t_{\min} - t_{\max}| \leq \mathrm{WL}$. If $t_{\min} < t_{\max}$, the event is an up ramp; otherwise, it is a down ramp. All points within the interval Δt are marked as up or down, and separate time series of both all up events and all down events are recorded. The sliding window moves forward one time step, 10 min, and the process is repeated until the end of the time series is reached. Any contiguous time steps marked as being part of up events are concatenated into a single up event, and the same is done for the down events.

In these cases the ramps can have a duration Δt greater or smaller than WL. The magnitudes and start/end times for all up and all down ramps are then stored for the entire time series. Given the use of the min–max values, ramps of opposite signs cannot overlap, although the end point of one event may be the starting point of an opposite-signed event.

Figure 4 shows the same schematic power time series as in Fig. 2, but ramp events as detected by the min–max method are displayed. Now the identified ramp events look more intuitive, and more ramp events can be found

[2] When a user runs the RT&M using the downloadable GUI, he or she can choose whether to run the RT&M over individual sites and then average the statistics, or to run the RT&M on the aggregated sites.

[3] When a user runs the RT&M using the downloadable GUI, overlapping points will be shown with blue squares.

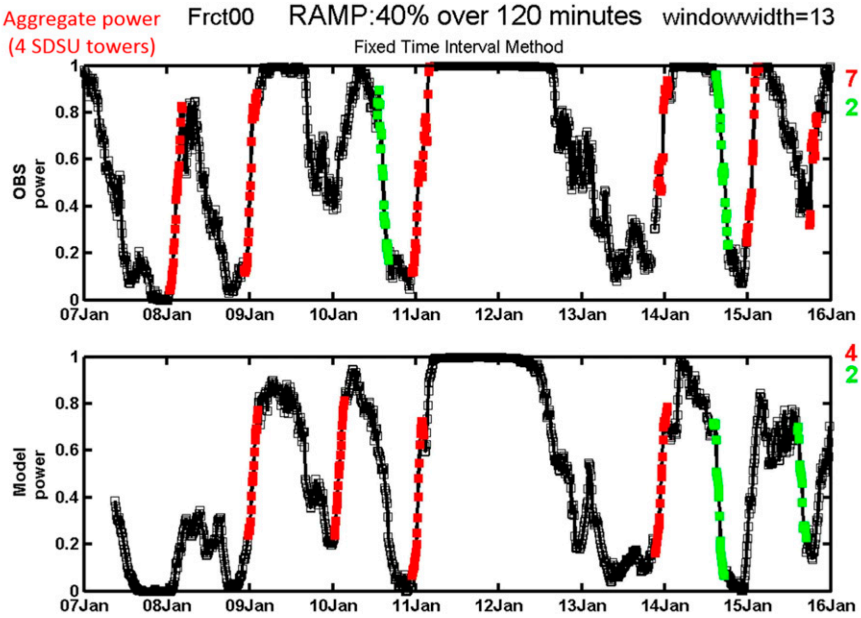

FIG. 3. (top) Time series of 10-min normalized aggregate power from anemometer measurements on the four SDSU towers and ramp events identified using the fixed-time interval method (using a 40% power change threshold over 2 h). (bottom) As in the top panel, but for the RAP model time series of power. Up ramps are in red and down ramps are in green. The numbers of ramps found in each time series are shown on the right.

in general as the selection of minimum and maximum power within WL shortens the duration of the event.

The min–max method on the same aggregate of observed and modeled data as in Fig. 3 (not shown) has generally similar results, finding seven up ramps and three down ramps in the time series of the observations and four up and two down ramps in the model simulation.

c. Explicit derivative method

The third method that we consider is referred to as the explicit derivative method. In this context, ''explicit'' means all points within the window WL are used to define a derivative and therefore the ramp. First, a smoothed time derivative of the power $\partial p/\partial t$ is defined as the slope of a linear least squares fit to the power over a time window WL. Next, if $|\partial p/\partial t| \geq \Delta p_{\mathrm{RD}}/\mathrm{WL}$, a ramp exists; if $\partial p/\partial t > 0$, it is an up ramp, and if $\partial p/\partial t < 0$, it is a down ramp. The beginning of an up-ramp event is found by searching for a minimum in power over the interval ½WL earlier in time than the first point where the derivative threshold is met, since those points were included in the derivative calculation. The end of an up ramp is found by searching for the maximum in power that occurs in the interval ½WL after the last point of the initial derivative ramp. Similar tests are done for the ends of a down ramp. As for the fixed-time interval method, with the explicit derivative method it is possible for two ramps of opposite signs to be partially overlapping in time. To accentuate differences between these two methods, we modify the explicit derivative results to truncate ramps that overlap. In the period of overlap of a down ramp followed by an up ramp, the minimum value of power is chosen as the end of the down ramp and the start of the new up ramp. If more than one occurrence of the same minimum value occurs, then the minimum closest to the down ramp is chosen as its end point, and the minimum closest to the up ramp is chosen as its beginning. If the period of overlap consists of an up ramp followed by a down ramp, the maximum

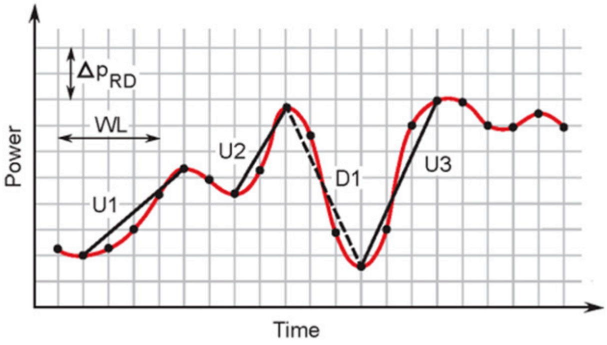

FIG. 4. Ramps identified by the min–max method for a window length WL and a ramp threshold Δp_{RD}.

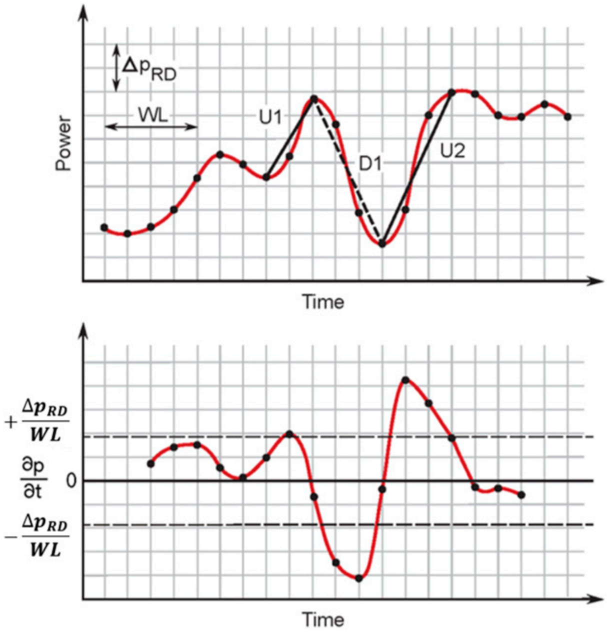

FIG. 5. (top) Ramps found by the explicit derivative method for a value of the smoothed derivative threshold given by Δp_{RD} and window length WL. (bottom) The smoothed power derivative corresponding to the power data in the top panel. The dashed lines indicate the smoothed derivative thresholds defining up and down ramps.

value of power is searched. As in the other two methods, any contiguous time steps marked as being part of ramp events of the same sign are concatenated into a single event, so that the ramp event Δt can be longer or shorter than the window length WL. Figure 5 displays a schematic time series of power (top panel) and the power derivative (bottom panel), and indicates ramps that are selected by this method.

Employing the explicit derivative method on the same aggregate of observed and modeled data in Fig. 3 (not shown) yields five up ramps and three down ramps in the time series of the observations and seven up ramps and two down ramps in the model simulation.

3. Matching of forecast and observed ramps

The next step is to develop a methodology for matching the observed and modeled events. The general philosophy we use is to match events that are closest in time; if multiple events have the same time separation, those with the closest ramp rate are matched. The inputs to the matching algorithm are the sequential list of ramp events over a time period t from the observations and forecasts, each defined by their duration of event (Δt_f and Δt_o), power change (Δp_f and Δp_o), and their center times (Ct_f and Ct_o), where the subscript f indicates the model forecast value and o the observed value. The number of events in the two time series in general will not be equal. Using these inputs, a matrix of dimension $N_f \times N_o$ is created, where N_f and N_o are the number of forecast and observed ramps, with the matrix populated by the differences in center times ($|\mathrm{Ct}_f - \mathrm{Ct}_o|$) for each combination of forecast and observed ramp events. A second matrix consists of the difference in the power ramp rate ($|\Delta p_f/\Delta t_f - \Delta p_o/\Delta t_o|$) between every pair of forecast and observed events, no matter what their time separation.

The matrix of differences in center times is then searched for the minimum value(s), corresponding to the model ramps that are closest in time to the observed. If this minimum value is larger than WL, then the event is unmatched. Frequently, multiple events with the same minimum timing error will be found, because of a limited number of time shifts possible for the discrete 10-min sample dataset. If more than one minimum exists, all of the minima are evaluated for instances when one model ramp is paired with two equally spaced (preceding and following) observed ramps. If the model value is paired with only one observation, these two events are matched, and then eliminated from any further searching. If the model ramp is paired with two observed events at the current time-shift minima, the choice of which one is matched with the model event is made based on the $\Delta p/\Delta t$ matrix, and the observed ramp with the smaller value in the ‘‘difference in power ramp rate’’ matrix is selected as the match. These matched modeled and observed ramp events are removed from both matrices, and the search for a minimum in the time shift is repeated. This process of searching through all model ramps is repeated until either all ramp events are matched or are determined to be unmatched (up/null or down/null events).

4. Forecast skill scoring methodology for a single ramp definition

The ramp identification and ramp matching procedures result in time series of matched pairs of forecast and observed ramps. Using this time series of events, a forecast score is determined by comparing the forecast and observed characteristics of each event. The ramp skill score accounts for forecast ramps matched to observed ramps, forecast ramps not matched with observed ramps, and observed ramps not matched with forecast ramps. The skill score incorporates phase, amplitude, and duration errors, and recognizes that up- and down-ramp events can have different impacts on grid operations. Also, the skill score is designed so that a set of random forecasts will have near-zero skill, as we

TABLE 1. Scenario definitions for matched and unmatched ramp events.

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

TABLE 2. Range of scores possible for all eight event scenarios for the simplified, symmetric case.

	Observed		
Model	Up	Null	Down
Up	From +1.0 to 0.0	0.0	From −1.0 to 0.0
Null	0.0	—	0.0
Down	From −1.0 to 0.0	0.0	From +1.0 to 0.0

verified by testing the RT&M on a randomly produced forecast. A negative score indicates the model is worse than random, and a positive score indicates the model has skill.

The first step is to classify the different types of ramp scenarios possible (Table 1). This is similar to a 3 × 3 contingency table (Wilks 2006) consisting of up, down, and null events, except that the null/null case is not considered and does not affect the skill score.

An equation is formulated to compute the scores for the nonnull scenarios (1, 3, 6, and 8) accounting for the timing, amplitude, and duration of the forecast and observed ramp difference. Nondimensional amplitude α, timing τ, and duration λ parameters are defined, with the score of an individual matched ramp event depending on a combination of the correctness of these parameters. For a perfect forecast the score reaches its maximum value, while for a forecast miss, the score will equal its minimum value. The score for these four scenarios is

$$\begin{aligned}\text{Score}_{\#} &= \text{MaxDistScore}_{\#}(\alpha_{\#}\tau_{\#}\lambda_{\#})^{1/3}\\ &\quad + \text{MinDistScore}_{\#}[1-(\alpha_{\#}\tau_{\#}\lambda_{\#})^{1/3}], \qquad (1)\end{aligned}$$

where the number sign (#) refers to the four nonnull ramp event scenarios, and MaxDistScore$_{\#}$ and MinDistScore$_{\#}$, respectively, represent limits to the score for perfect or missed forecasts, nominally 1, 0, or −1.

We develop values for α, τ, and λ, as well as for MaxDistScore$_{\#}$ and MinDistScore$_{\#}$ for two separate cases. The first is a simplified case (presented in this section) in which up/up ramps and down/down ramps are treated equally, and all unmatched events are given zero skill. For this case the value of MinDistScore$_{\#}$ will always be zero. The second case (presented in the appendix) is more complex and contains asymmetries, where up/up ramps and down/down ramps are not treated equally and unmatched events can have nonzero skill. For this case the value of MinDistScore$_{\#}$ can be different from zero. This case will be used when the user wants to take into account the possibility of curtailing wind production, which can result in forecast errors of different signs having different financial or grid reliability consequences.

The simplified scoring strategy uses the symmetric range of scores in Table 2, with up/up and down/down events having a score between +1 and 0, tending to 0 as the timing error approaches WL; down/up and up/down events having score ranges between 0 and −1; and all missed forecasts having zero skill.

The values of MaxDistScore$_{\#}$ in (1) are the scores with the maximum distance from zero listed in Table 2. Thus, for an up/up event (scenario 1) and a down/down event (scenario 8), MaxDistScore$_{1,8}$ is +1, while for an up/down event (scenario 3) and a down/up event (scenario 6), MaxDistScore$_{3,6}$ is −1. The forecast timing and amplitude skill parameters $\tau_{\#}$, $\alpha_{\#}$, and $\lambda_{\#}$ are defined as the linear equations:

$$\tau_{\#} = \left(1-\frac{|\text{Ct}_f-\text{Ct}_o|}{\text{WL}}\right), \qquad (2)$$

$$\alpha_{1,8} = (1-|\Delta p_f-\Delta p_o|), \qquad (3)$$

$$\alpha_{3,6} = \left(\frac{|\Delta p_f-\Delta p_o|}{2}\right), \qquad (4)$$

$$\lambda_{1,8} = \left(1-\left|\frac{\Delta t_f-\Delta t_o}{\Delta t_f+\Delta t_o}\right|\right), \quad \text{and} \qquad (5)$$

$$\lambda_{3,6} = \left(\frac{2\Delta t_{\min}}{\Delta t_f+\Delta t_o}\right), \qquad (6)$$

where $\Delta t_{\min} = \text{WL}$ for the fixed-time interval method because it defines the ramp using the points at the extremes of the window and $\Delta t_{\min} = 10$ min (the resolution of the data) for the min–max and the explicit derivative methods because for these methods the duration of the ramp can be smaller than the window length.

For all scenarios, the timing skill [(2)] falls in the range $0 \le \tau \le 1$, with a value of 1 when there is no timing error, decreasing linearly to zero when the timing error reaches WL.

For scenarios 1 and 8 the best skill is obtained when the forecast and observed ramps have identical power amplitudes ($\alpha_{1,8} = 1$), durations ($\lambda_{1,8} = 1$), and no phase error for their center times ($\tau_{1,8} = 1$). For this perfect forecast the score in (1) is equal to MaxDistScore$_{1,8}$ = 1. The values of $\alpha_{1,8}$, $\lambda_{1,8}$, and $\tau_{1,8}$ decrease toward zero as the forecast becomes less perfect and the score in (1) approaches MinDistScore$_{1,8}$ = 0.

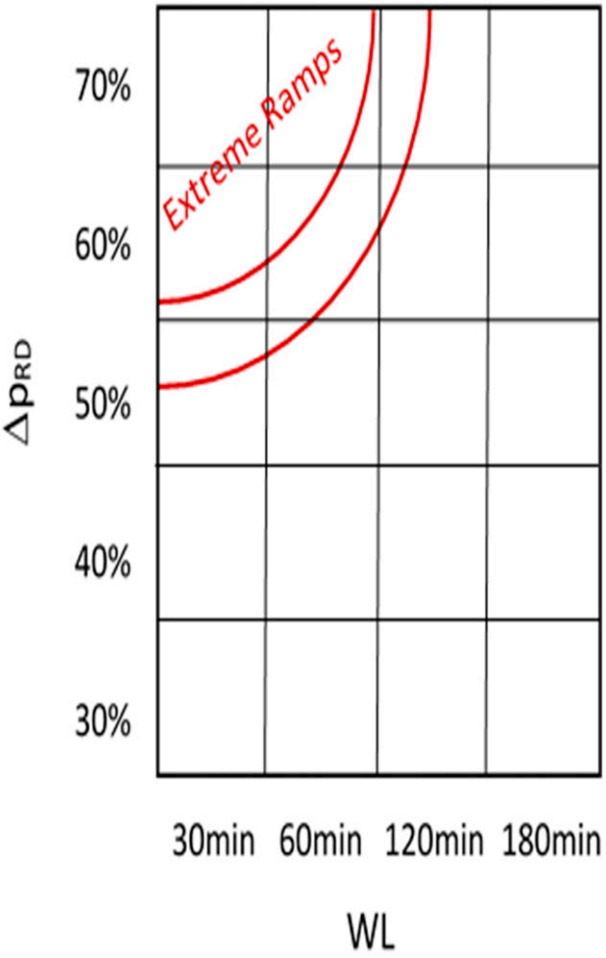

FIG. 6. Schematic diagram of a ramp matrix. Extreme ramps are in the top-left corner, and low-amplitude ramps of longer duration are in the bottom-right corner.

For scenarios 3 and 6 the worst case occurs when the forecast and observed ramps have identical (and opposite) maximum power amplitudes ($\Delta p_f = \pm 1$, $\Delta p_o = \mp 1$, so that $\alpha_{3,6} = 1$); the ramps are very sharp, approaching the resolution of the data (in which case $\lambda_{3,6} = 1$); and there is no phase error for their center times ($\tau_{3,6} = 1$). In this case the score in (1) is equal to $\mathrm{MaxDistScore}_{3,6} = -1$. As the amplitude of the ramps in these scenarios become smaller, $\alpha_{3,6}$ in (4) will decrease such that $\alpha_{3,6} = \Delta p_{\mathrm{RD}}$ at its minimum (with $\alpha_{3,6} \to 0$ as $\Delta p_{\mathrm{RD}} \to 0$). Also, $\lambda_{3,6}$ in (6) will decrease as Δt_f and Δt_o become longer, and the score in (1) approaches $\mathrm{MinDistScore}_{3,6}$ (in this case equal to 0) since the utility operator has more time to adjust for the wrong forecast.

5. Forecast skill scoring: Matrix of skill values

The ramp metric developed above applies to ramps defined by a single power amplitude threshold and window length. Ideally, one would like to know which model is best for a range of power thresholds and window lengths and then to average the model's skill over this range of values. For this reason, we consider a matrix of ramp skills schematically illustrated in Fig. 6.

For this study four different time windows (30, 60, 120, and 180 min) and five different power thresholds (30%, 40%, 50%, 60%, and 70%) have been chosen.[4] The ramp matrix has been designed so that each matrix element answers the question, does a ramp event of a particular duration exceed a Δp_{RD} threshold? Therefore, each matrix element contains the cumulative information for all ramps greater than that threshold. For example, a 70% Δp_{RD} over 2 h also fulfills the requirement of 60% Δp_{RD}, or lower, over the same WL; therefore, this ramp will be taken into account in more than one matrix bin.

Skill scores as defined in section 4 for each value of power threshold and window length are calculated and placed into each matrix element. Skill scores using extreme ramp definitions (largest power thresholds and shortest window lengths) are placed in the top-left corner of the matrix. Skill scores for more frequently occurring and weaker ramp events (lower power thresholds and longer window lengths) will be placed in the bottom-right corner of the matrix.

a. *Results from 9 days of observation from four SDSU tall towers*

To illustrate how the RT&M can be used to measure the skill of a model forecasting ramp events, we applied it to 9 days' worth of observations from the four SDSU tall towers introduced in section 2 and corresponding forecasts from the RAP model at the same locations. Recognizing that a single 9-day period is insufficient to definitively assess the skill of this model, the intent of this study is limited to describing the RT&M and how it can be used to measure model skill at forecasting ramp events on a real observational dataset. A future study is under way testing the RT&M on a larger observational dataset, comparing different models, and comparing similar models run at different spatial resolutions and with different model output frequencies.

For this exercise we examined the average of the statistical results for each tower site, instead of the aggregate of four towers, although as stated earlier the RT&M provides the option of aggregating the results

[4] When a user runs the RT&M using the downloadable GUI, he or she could also choose different time windows and power thresholds by modifying the Matlab code, but the minimum possible time window that can be chosen is equal to 2 times the resolution of the data, in our case the minimum time window would be equal to 2 × 10 min = 20 min.

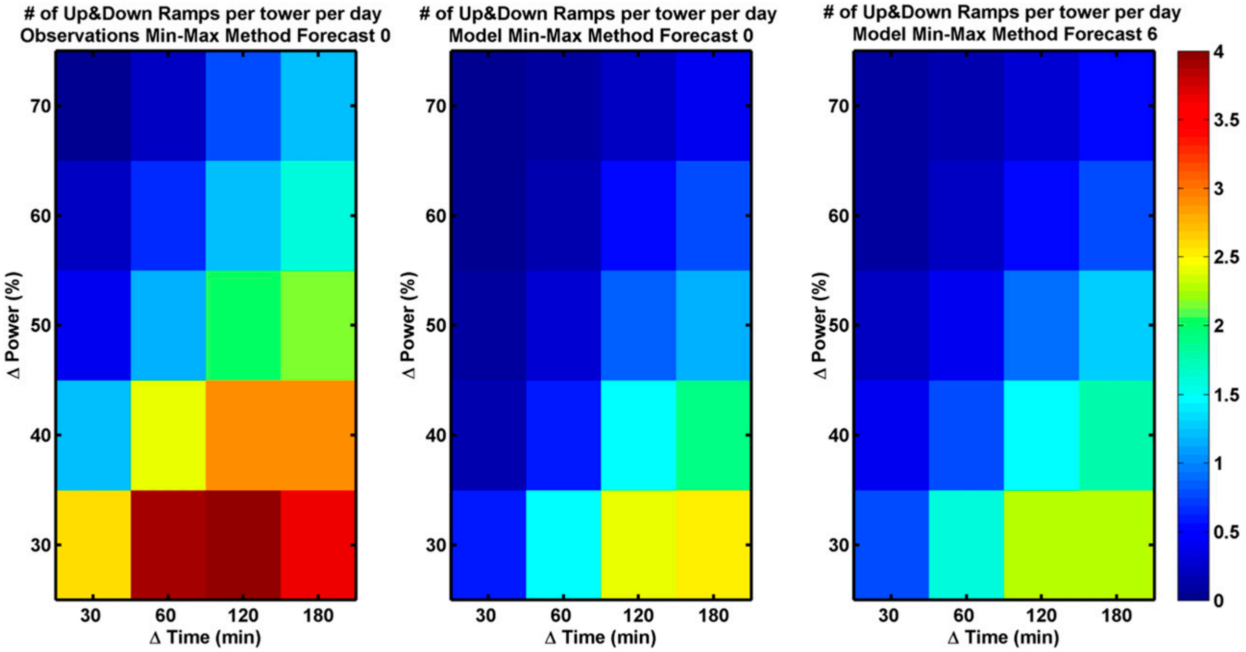

FIG. 7. Number of occurrences of ramp events per day per tower that fall into each matrix bin during the 9 days of analysis using the stitching method and using the min–max method, for (left) the tall-tower observations, and for (center) the forecast initialization time (hour 0) and (right) forecast hour 6 of the RAP model. The ramp definition power threshold ranges from 30% to 70%, and the window length ranges from 30 to 180 min.

according to the user needs. We also set the matrix of scores to be symmetric (as in Table 2). First, we present results obtained using the RT&M stitching method to compare the RAP model to observations.

Figure 7 displays the number of occurrences of ramp events found using the min–max identification method, as observed from the tall towers (left panel), and for the RAP model (forecast hour 0, initialization time, center panel; forecast hour 6, right panel), for the previously specified range of ramp power thresholds and window lengths. Relatively few extreme ramp events are found (top-left corner of each panel) while many small-amplitude and long-duration ramps are found (bottom-right corner of each panel). A similar number of events is found at forecast hour 6 compared to the initialization time. In contrast, the number of occurrences in the observations is larger. This is because at 13-km resolution the model has considerably smoother fields than the observed point location power time series.

Skill score matrices for the RAP model simulations using the fixed-time interval ramp definition method are shown in Fig. 8 for the initialization time and forecast hours 3, 6, and 14. The skill is larger for longer window lengths compared to the shorter windows. The skill is greatest at the initialization time, and slowly decays with forecast length. Skill scores for the other methods look qualitatively similar (not shown).

b. *Weighting matrix*

Using the methodology presented above, a perfect forecast of a 30% power capacity ramp over 2 h and one of 70% over 30 min may both have forecast skills of 1.0, yet forecasting the larger ramp will be more important and should have more value than the smaller ramp.

In place of averaging the ramp skill scores in all of the score matrix elements equally, a weighting function can be applied before averaging the score matrix that accounts for the fact that the skill scores for the more extreme events will likely have a greater impact on grid operations than the weaker ramps. The weighting matrix that we have used starts with a weight of 1.0 in the top-left corner of the matrix in Fig. 6 (most extreme ramps) and decreases the weight by 10% for each 10% change in the ramp power threshold and each increment in window length.[5]

The average score across the entire matrix is shown in Fig. 9 for the three ramp definition methods, using both

[5] When a user runs the RT&M using the downloadable GUI, he or she can choose to run the RT&M averaging the ramp skill scores in all of the score matrix elements equally, or apply the weighting function introduced here to weight the extreme events (or create their own weighting matrix, according to his or her needs, by modifying the Matlab code itself).

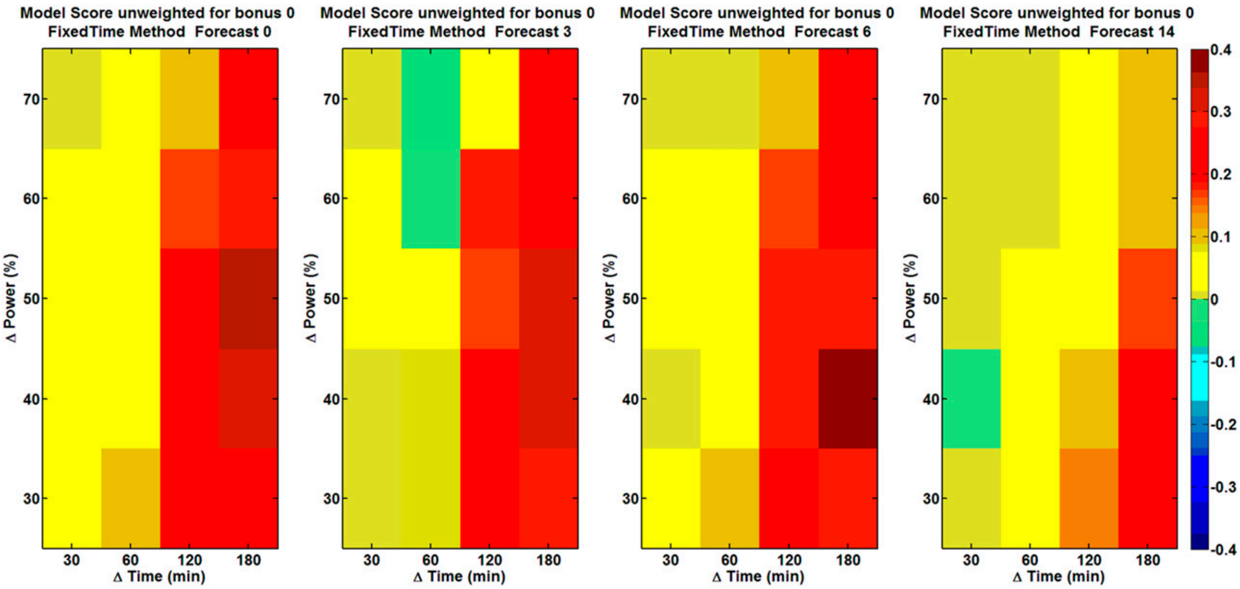

FIG. 8. Matrix of skill scores using the stitching method for the fixed-time interval method for the RAP model forecasts for the initialization time and forecast hours 3, 6, and 14.

an equal weighting of all the matrix elements (top panel), and when using the weighting matrix (bottom panel). The unweighted averaged skill score is greater than the weighted skill score, because the model has less skill at forecasting the most extreme ramps. Although the 9-day period used in this study is not long enough to claim definitive results, we notice that for this exercise the skill score of the RAP model is positive for all forecast hours and methods, decreasing with forecast length. We also note that model forecast skill tends to be greatest when using the explicit derivative method, which may be due to the fact that this method applies more temporal smoothing to the data then does, say, the min–max method. The choice of method will clearly depend on the user's specific forecast needs.

A breakdown of the ramp events into the eight different scenarios using the three ramp identification methods (not shown) shows the largest number of events by far occurs for the two model null event scenarios (4 and 5). This is most likely due to the fact that observational data have much more temporal variability compared to the model time series. For this reason, as noticed in section 2, fewer ramp events are found in the model time series and many observed ramps have no match in the model time series. The next most common events are scenario 1 (up/up), scenario 2 (up/null), scenario 8 (down/down), and scenario 7 (down/null), but with much lower frequencies than scenarios 4 and 5. The least common are scenarios 3 (up/down) and 6 (down/up), the worst scenarios possible.

The ramp skill scores can also be broken down into each scenario category (not shown). The positive contribution to skill score comes from scenarios 1 and 8, when the forecast accurately predicts up ramps and down ramps when they are observed. Scenarios 3 and 6 (ramp forecasts with the sign opposite than that observed) have a negative contribution smaller than the positive contribution of scenarios 1 and 8. Scenarios 2, 7, 4, and 5 (null events) have no net effect as the score for these scenarios is set equal to zero (see Table 2) for this exercise.

Also, the skill of a model at forecasting observed up-ramp events versus down-ramp events can be tested separately with this same RT&M, simply rerunning it first only using nonzero values for scenarios 1 and 6 in Table 2 (observed up ramps) and the second time only using nonzero values for scenarios 3 and 8 in Table 2 (observed down ramps). For this dataset this comparison is presented in Fig. 10, where the dashed lines are used for up events and the solid lines represent down ramps. The model has greater skill at forecasting observed up-ramp events compared to down-ramp events. Again the skill of the model at forecasting both up- and down-ramp events decreases with forecast length, but stays positive for all forecast hours.

Finally, in Fig. 11 we present the same results presented in Fig. 9 (bottom) and Fig. 10, but obtained by the RT&M independent forecast runs method to compare the model to the observations. In this case, the observed and model time series are shorter and equal to the length of the model forecast (equal to 15 h in the case of the RAP model used in this study). During a period of 9 days in January 2012, there are 216 (24 times 9) such

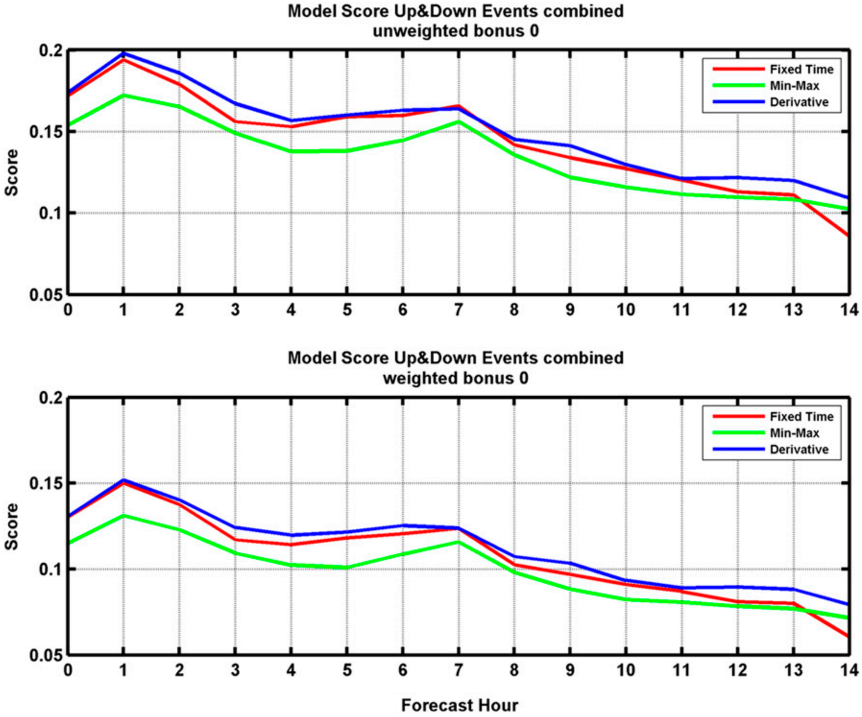

FIG. 9. Skill score results using the stitching method for the fixed-time ramp identification method (red), the min–max ramp identification method (green), and the explicit derivative ramp identification method (blue) with (top) equal weighting of all matrix elements and (bottom) when using the weighting matrix.

time series that allow comparison of the observed and model data for ramp identification. Ramps are determined and matched between each pair of time series in the same way they were determined and matched using the ‘‘stitching method’’ presented before, but when we measure the skill of the model for each particular ramp, we add that skill to the matrix of skills for the same forecast hour during which the central time of the model ramp occurs (in the appropriate matrix element, relative to that particular ramp definition). In this way we can preserve the information on how the statistics vary as a function of the forecast horizon. A disadvantage of this method is that the beginning and ending forecast hours will suffer from truncated ramps that potentially begin before the start of the forecast cycle or end after forecast hour 15. For instance, there

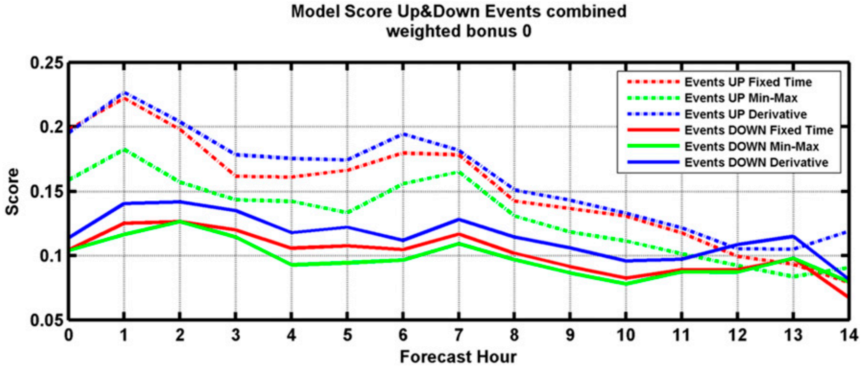

FIG. 10. Skill score results using the stitching method for the fixed-time interval method (red), the min–max method (green), and the explicit derivative method (blue) at forecasting observed up-ramp (dashed lines) vs down-ramp (solid lines) events. In all curves the skill score is computed using the weighting matrix.

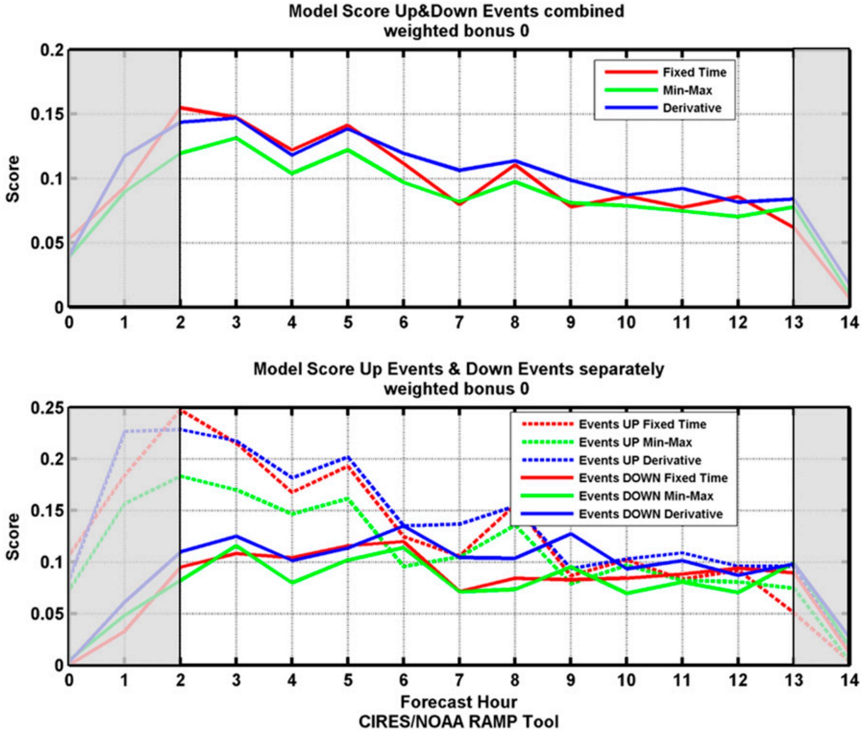

FIG. 11. (top) Skill score results using the independent forecast runs method for the fixed-time ramp identification method (red), the min–max ramp identification method (green), and the explicit derivative ramp identification method (blue). (bottom) Skill score results at identifying up-ramp (dashed lines) vs down-ramp (solid lines) events. In all curves the skill score is computed using the weighting matrix.

will be no ramps with central time at forecast hour 00:00, so forecast hour 0 skill will only have skill coming from ramps centered at 00:15, 00:30, and 00:45. Moreover, ramps centered at 00:15 will be limited in definition (no ramps with $\Delta t > 30$ min can be centered at 00:15, because the initial time of these ramps would happen before forecast hour 0000, which the model does not have). For this reason the skill of the model at forecast hour 0 will be less than what we were able to measure when the stitching method was used. Similar limitations will be true for forecast hour 1, as in this case there will be ramps with central time at forecast hour 01:00, but these will be limited in definition to those having a $\Delta t < 120$ min, because the initial time of the ramps with a $\Delta t > 120$ min would be before forecast hour 00:00, which again the model does not have. Similar considerations are true for forecast hour 14 (for the model we used in this study, or the last forecast hour of a model in general). The number of disadvantaged forecast hours will be a function of the ramp definitions chosen by the user.

When this second approach is run on the 9-day dataset, we found that the statistics for forecast hours 2–13 are very similar to those found with the stitching method (cf. Fig. 11 to Fig. 9, bottom panel, and to Fig. 10), proving that the RT&M is very robust. As expected from the considerations above, the skill of the RAP in Fig. 11 (both panels) is less that what we had found before in Fig. 9 (bottom) and Fig. 10 at forecast hours 0, 1, and 14. For this reason, Fig. 11 shows gray areas for the forecast hours that suffer from the deficiencies described above. Nevertheless, both of these approaches are available in the RT&M, and the user can choose the one that best addresses their analysis needs.

6. Summary and conclusions

A Ramp Tool and Metric was developed to test the ability of a model to forecast ramp events for wind energy. Power forecasts were evaluated by converting tall-tower observations and model forecast wind speeds to normalized capacity power using a standard IEC2 power curve. Two options are provided to the user to decide how to compare the model to observations.

The RT&M has three components: it identifies wind ramp events, matches forecast and observed ramps, and calculates a skill score for the forecasts.

The skill score incorporates phase, duration, and amplitude errors. Since no single pair of changes in power

TABLE A1. Scores for the four possible null scenarios when a bonus is applied to the situations that can be solved by curtailing wind energy generation.

Scenario	Model	Observed	Utility action	Score
2	Up	Null	Fuel spot market purchase	0
4	Null	Up	Wind curtailment or cancellation of planned increase of fossil fuel units	(0.1 × bonusweight)
5	Null	Down	Fuel spot market purchase	0
7	Down	Null	Wind curtailment or cancellation of planned increase of fossil fuel units	(0.1 × bonusweight)

and time thresholds may be representative, and some users may use different ramp definitions for different situations, the RT&M provides the option to integrate the skill over a range of changes in power and time windows. Although specific RT&M parameter values were used in this study, the tool is flexible and can be modified by users for their purposes. Also, a greater emphasis can be given to the more extreme events using a weighting matrix.

We tested the RT&M on 9 days' worth of observations from a set of four SDSU tall towers located in South Dakota, and used these data to illustrate how the RT&M can be employed to evaluate the skill of a model (RAP in this example) at forecasting ramp events. This hourly updated model runs operationally over North America at 13-km resolution, and was used during the 2011–12 WFIP campaign. For the RT&M analysis, RAP output at 15-min resolution was used, allowing us to consider short-duration ramps. The RAP model is found to have positive skill decreasing with forecast length, and greater skill at forecasting up-ramp events compared to down-ramp events. We developed and described three different methods for identifying ramp events but at this stage of the analysis it is not yet clear which one is best.

Since this RT&M is used to evaluate the skill of a model when forecasting ramp events in a time series of power data, in principle it could also be used on a time series of power data generated by solar plants. In this case the values of the time windows and power thresholds should be changed according to the expected behavior of the power data produced by the solar plants. This work is in our future plans. The RT&M is publically available online (http://www.esrl.noaa.gov/psd/products/ramp_tool/).

Acknowledgments. This research has been funded by the U.S. Department of Energy under the Wind Forecast Improvement Project (WFIP), Award DE-EE0003080, and by NOAA/Earth System Research Laboratory. The authors wish to acknowledge Joseph Olson from the NOAA/ESRL/GSD group for providing the RAP model outputs, Barb DeLuisi from the NOAA/ESRL/PSD group for maintaining the RT&M website, and three anonymous reviewers for the helpful comments.

APPENDIX

Curtailment Bonus

The simplified scoring strategy presented in the main body of the manuscript uses the symmetric range of scores shown in Table 2. However, symmetric scoring does not account for the fact that forecast errors of different signs (underprediction versus overprediction of power) may have different financial consequences. In particular, if the observed wind power P_o exceeds the amount of power $\hat{P}$ that the grid operator expects to occur during the ramp event power ($P_o > \hat{P}$), the grid operator has the opportunity to simply curtail the wind production. On the other hand, when the actual wind power produced is less than that expected ($P_o < \hat{P}$), the grid operator could be forced to make a power purchase on the spot market. In some markets curtailing wind may be much cheaper than dealing with the opposite case of making a power purchase on the spot market, so that an underprediction does not carry the same cost as an overprediction of the same magnitude. To account for this asymmetry, we describe below the concept of a curtailment bonus that can be incorporated into the calculation of the ramp score for a subset of the scoring

TABLE A2. Range of scores possible for all eight event scenarios for the case with bonuses.

	Observed		
Model	Up	Null	Down
Up	From +1.0 to (0.1 × bonusweight)	0.0	From −1.0 to 0.0
Null	(0.1 × bonusweight)	—	0.0
Down	From −1.0 to (0.2 × bonusweight)	(0.1 × bonusweight)	From +1.0 to (0.1 × bonusweight)

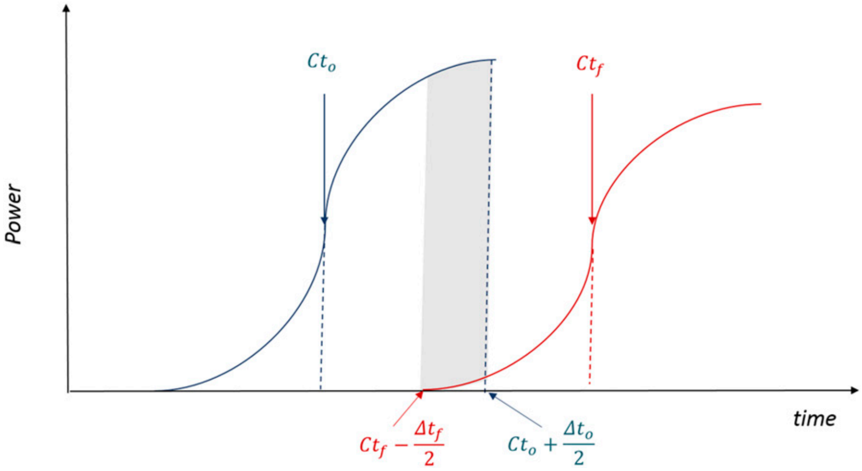

FIG. A1. Schematic representation of the first situation when the bonus will be applied to the model. The shaded area represents the amount of wind power production during an observed ramp event that could be curtailed and that overlaps with a matched forecast ramp.

scenarios. By ‘‘bonus’’ we do not imply that curtailment is necessarily good, only that in some situations curtailing wind can be less expensive than bringing up a fossil fuel generator.

In general terms, the bonus should be proportional to the integrated amount of wind that can be curtailed:

$$\text{Bonus} \propto \int_{t_s}^{t_e} (P_o - \hat{P})\,dt \quad \text{when} \quad P_o > \hat{P},$$

where t_s and t_e are appropriately defined start and end times of the integration period. Rather than use the full integral, we wish to define a simplified ramp bonus based on the few parameters used to define a ramp: Ct_f and Ct_o, Δp_f and Δp_o, and Δt_f and Δt_o.

Implicit in our curtailment analysis is the assumption that the grid is always balanced at the start of a ramp event, even if the model forecast is already in error at that point. Therefore, for example, a forecast down ramp that has been matched with an observed up ramp will always provide the opportunity for curtailment, even if the actual time series of forecast power remains greater than the observed power during the ramp event. That is, the grid operator will have accounted for the initial error in the forecast, so that the expected amount of power to be produced will be offset from the raw model power forecast. Since the tool only analyzes ramp events, a second assumption that we make to simplify the analysis is that a curtailment bonus will only be applied for those time periods when either an observed or forecast ramp is present.

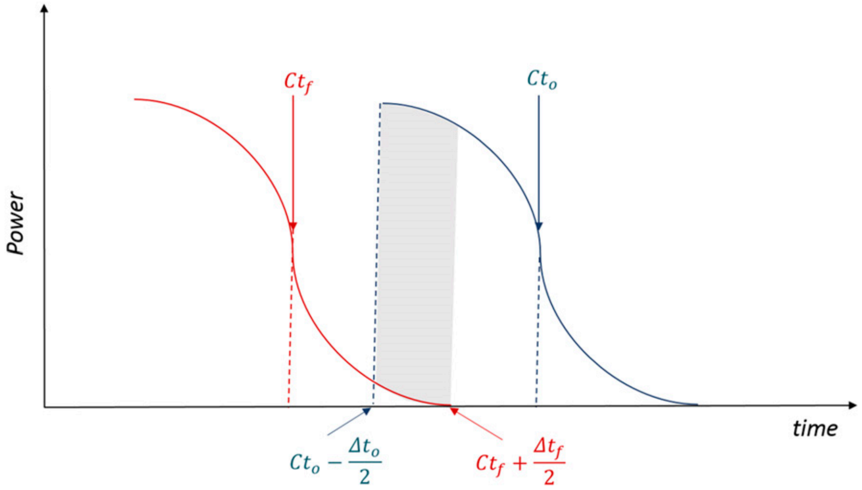

FIG. A2. Schematic representation of the second situation when the bonus will be applied to the model.

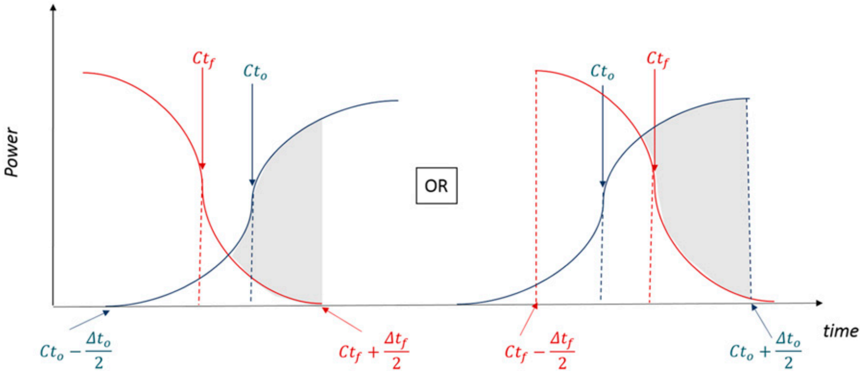
Power
Ct_f
Ct_o
OR
Ct_o
Ct_f
time
$Ct_o - \frac{\Delta t_o}{2}$
$Ct_f + \frac{\Delta t_f}{2}$
$Ct_f - \frac{\Delta t_f}{2}$
$Ct_o + \frac{\Delta t_o}{2}$

FIG. A3. Schematic representation of the third situation when the bonus will be applied to the model.

A feature of the curtailment bonus option is that the user can choose the weight to give to this bonus through a variable called bonusweight that ranges between 0 and 1 (with steps of 0.1), with 0 being equivalent to no bonus and 1 being the maximum. When a user runs the RT&M using the downloadable GUI, he or she can select what value to assign to the bonusweight in the appropriate box.

For the null (unmatched) events, curtailment can occur for scenarios 7 and 4, when the total power observed is greater than the forecast. However, even though for these two null case scenarios there may be value in curtailment, because these are missed events we still wish for the model to have a skill score close to zero. Therefore, for scenarios 7 and 4 we will multiply the value of bonusweight by a small value (0.1), so that the maximum score for these scenarios will never be greater than this small value. The scoring strategy for the null scenarios will then become as shown in Table A1.

Once we have assigned scores to the null scenarios, we can complete the entire table again, but taking into consideration that $\mathrm{MinAmpScore}_1$ of scenario 1 has to converge to the sum of the scores in scenarios 2 and 4 (as an observed and forecast up/up-ramp pair slide farther apart from one another in time, when they reach the window length WL, they will become an up/null pair and a null/up pair, and the sum of these two scenarios

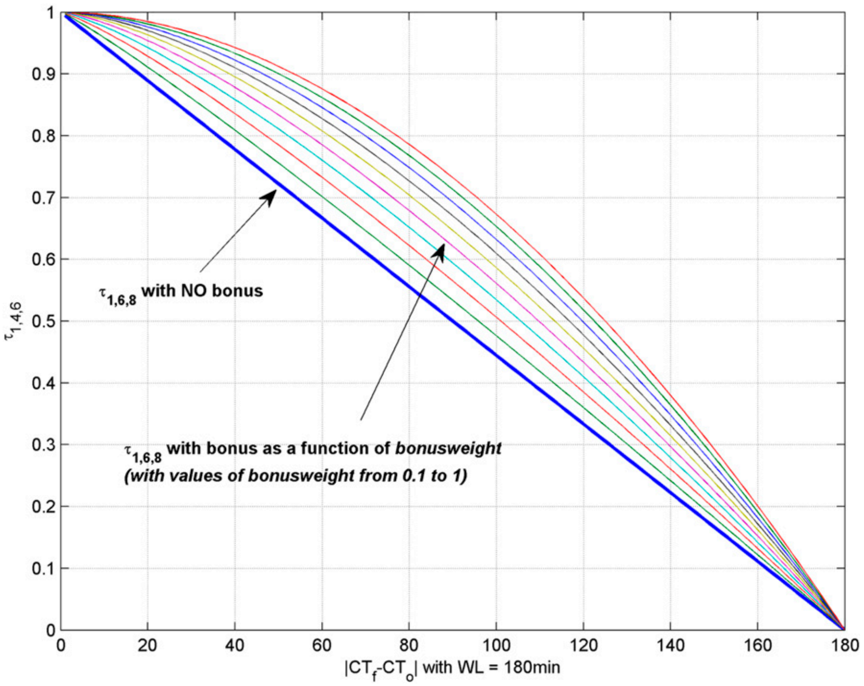

FIG. A4. The $\tau_{1,6,8}$ function with no bonus applied (thick blue line) and with the bonus (thin colored lines).

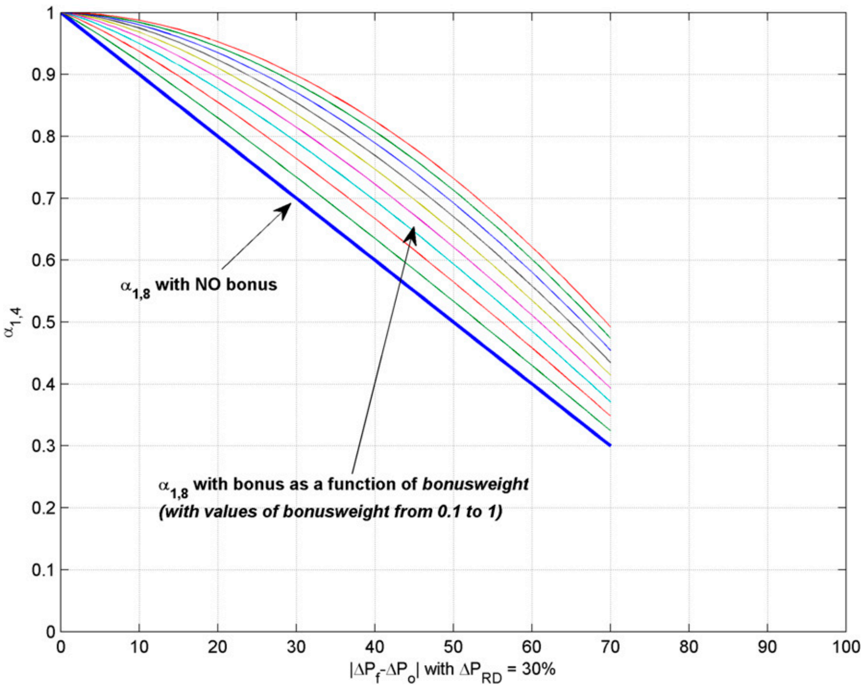

FIG. A5. As in Fig. A4, but for the $\alpha_{1,8}$ function.

should match the up/up pair at that point), and similar considerations can be repeated for MinAmpScore_3 of scenario 3, MinAmpScore_8 of scenario 8, and MinAmpScore_6 of scenario 6 as follows:

$$\begin{aligned}&\text{MinDistScore}_1 \rightarrow \text{Score}_2 + \text{Score}_4,\\&\text{MinDistScore}_3 \rightarrow \text{Score}_2 + \text{Score}_5,\\&\text{MinDistScore}_8 \rightarrow \text{Score}_7 + \text{Score}_5, \quad \text{and}\\&\text{MinDistScore}_6 \rightarrow \text{Score}_7 + \text{Score}_4.\end{aligned}$$

The scoring strategy for the eight scenarios will then become the one presented in Table A2. When the bonus is applied, not only is the range of scores changed, but the formulations are changed as explained below.

For the nonnull scenarios, bonuses will be applied in three different situations:

1) The first situation when the bonus will be applied to the model is in scenario 1, but only when the following conditions are met:
   - the central time of the forecast is later compared to the central time of the observations ($\text{Ct}_f > \text{Ct}_o$), and
   - the power change of the forecast is less than the power change of the observations ($\Delta p_f \le \Delta p_o$), and
   - the starting time of the forecast event happens before the end time of the observed event [$\text{Ct}_f - (\Delta t_f/2) < \text{Ct}_o + (\Delta t_o/2)$], so that we are sure there is a time during the forecast event when the actual wind power available is larger than the forecast, and, consequently, curtailment is possible. We note that it is possible for these criteria to not be met even though two ramps are matched.

   This situation is schematically presented in Fig. A1. In this case the gray area is the time during which the nonperfect forecast results in the actual wind power supply available (blue curve) being greater than the forecast power (red curve), and hence the amount required to balance demand, which can be alleviated by curtailing wind energy generation. From this figure we can see that if the forecast starting time is later compared to the final time of the observations, there will not be a time during which the observations are larger than the forecast and therefore the bonus cannot be applied.
2) The second situation when the bonus will be applied to the model is in scenario 8, but only when the following conditions are met:
   - the central time of the forecast is earlier compared to the central time of the observations ($\text{Ct}_f < \text{Ct}_o$), and
   - the power change of the forecast is less (more negative) than the power change of the observations ($\Delta p_f \le \Delta p_o$), and
   - the final time of the forecast event happens after the initial time of the observed event [$\text{Ct}_f + (\Delta t_f/2) > \text{Ct}_o - (\Delta t_o/2)$], so that we are sure there is a time during the forecast event when the observations are larger than the forecast, and, consequently, curtailment is possible.

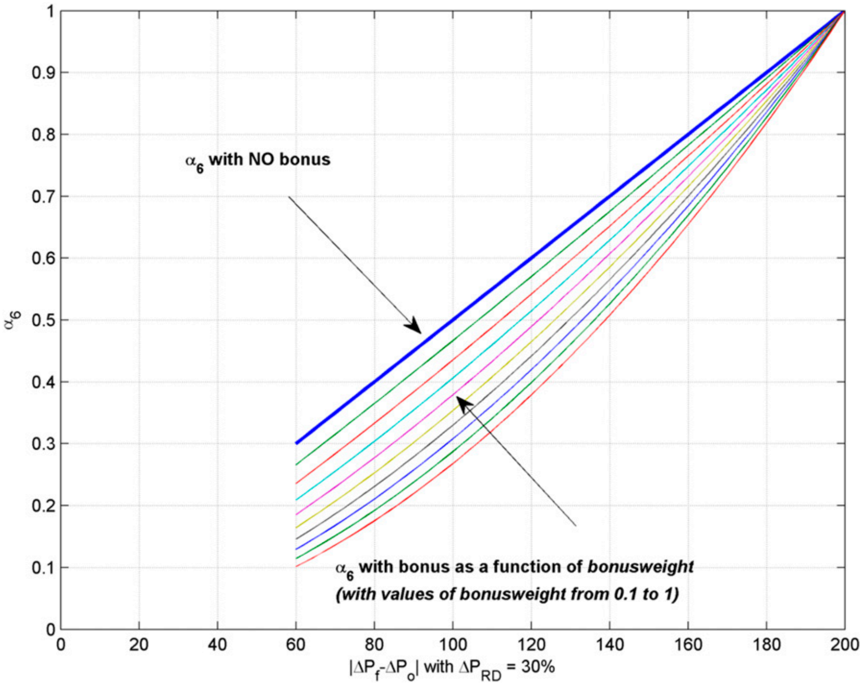

FIG. A6. As in Fig. A4, but for the α_6 function.

This situation is schematically presented in Fig. A2. In this case the gray area is again the time during which the nonperfect forecast can be alleviated by curtailing wind energy generation. Again, from this figure we can see that if the forecast final time is instead earlier compared to the initial time of the observations, there will not be a time during which the observations are larger than the forecast and the bonus cannot be applied.

3) The third and last situation when the bonus will be applied is in scenario 6, but only when the following conditions are met:
   - the central time of the forecast is earlier compared to the central time of the observations ($\mathrm{Ct}_f < \mathrm{Ct}_o$), and
   - the final time of the forecast event happens after the initial time of the observed event [$\mathrm{Ct}_f + (\Delta t_f/2) > \mathrm{Ct}_o - (\Delta t_o/2)$], or
   - the central time of the forecast is late compared to the central time of the observations ($\mathrm{Ct}_f > \mathrm{Ct}_o$), and
   - the initial time of the forecast event happens before the end time of the observed event [$\mathrm{Ct}_f - (\Delta t_f/2) < \mathrm{Ct}_o + (\Delta t_o/2)$], so that we are sure there is a time during the forecast event when the observations are larger than the forecast, and, consequently, curtailment is possible.

This situation is schematically presented in Fig. A3. In this case the gray area is again the time during which the nonperfect forecast can be alleviated by curtailing wind energy generation. Again, from this figure we can see that if the above conditions are not met, there will not be a time during which we can be sure the observed ramp event power is larger than the forecast and therefore the bonus cannot be applied.

To assign the bonuses to the model in the nonnull scenarios and in the instances where curtailment is possible, we need to modify the formulations for $\tau_{1,6,8}$, $\alpha_{1,8}$, and α_6.

The new formula for $\tau_{1,6,8}$ will become

$$\tau_{1,6,8} = \left(1 - \frac{|\mathrm{Ct}_f - \mathrm{Ct}_o|}{\mathrm{WL}} B_{\tau_{1,6,8}}\right),$$

with $B_{\tau_{1,6,8}} = (|\mathrm{Ct}_f - \mathrm{Ct}_o|/\mathrm{WL})^{\mathrm{bonusweight}}$ and bonusweight = 0:0.1:1.

So,

$$\tau_{1,6,8} \quad \text{with} \quad \text{BONUS} = \left[1 - \left(\frac{|\mathrm{Ct}_f - \mathrm{Ct}_o|}{\mathrm{WL}}\right)^{(1+\mathrm{bonusweight})}\right].$$

The $\tau_{1,6,8}$ function is presented in Fig. A4, when a value of WL = 180 min is chosen.

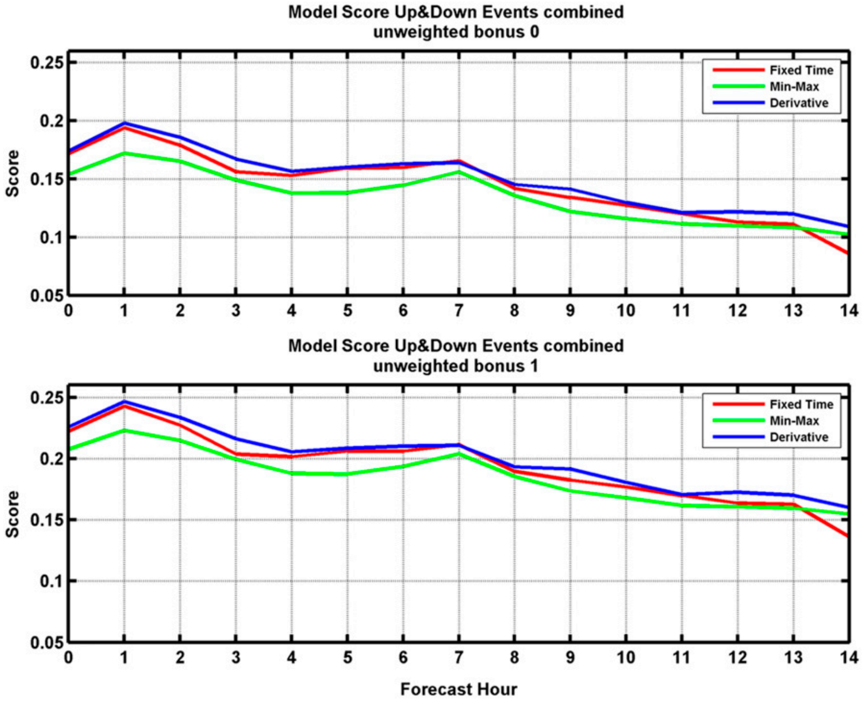

FIG. A7. Skill score results for the fixed-time ramp identification method (red), the min–max ramp identification method (green), and the explicit derivative ramp identification method (blue) with equal weighting of all matrix elements using the skill score in Table A2 with (top) bonusweight = 0 (equivalent to top panel in Fig. 9) and (bottom) when using bonusweight = 1.

The new formula for $\alpha_{1,8}$ will become

$$\alpha_{1,8} = (1 - |\Delta p_f - \Delta p_o|B_{\alpha_{1,8}}),$$

with $B_{\alpha_{1,8}} = |\Delta p_f - \Delta p_o|^{\text{bonusweight}}$ and bonusweight = 0:0.1:1.

So,

$$\alpha_{1,8} \text{ with BONUS} = [1 - |\Delta p_f - \Delta p_o|^{(1+\text{bonusweight})}].$$

The $\alpha_{1,8}$ function is presented in Fig. A5 in the case of $\Delta p_{\text{RD}} = 30\%$.

Finally, the new formula for α_6 will become

$$\alpha_6 = \left(\frac{|\Delta p_f - \Delta p_o|}{2} B_{\alpha_6}\right),$$

with $B_{\alpha_6} = (|\Delta p_f - \Delta p_o|/2)^{\text{bonusweight}}$ and bonusweight = 0:0.1:1.

So,

$$\alpha_6 \text{ with BONUS} = \left[\left(\frac{|\Delta p_f - \Delta p_o|}{2}\right)^{(1+\text{bonusweight})}\right].$$

The α_6 function is presented in Fig. A6 in the case of $\Delta p_{\text{RD}} = 30\%$.

Using a value of the bonusweight = 1, we repeated the same analysis presented in the body of the manuscript, when the stitching method is used to compare the model to the observations. The average score across the entire matrix is shown in the two panels of Fig. A7 for the three ramp definition methods, using an equal weighting of all the matrix elements and a value of bonusweight = 0 in the top panel (equivalent to the top panel in Fig. 9), and bonusweight = 1 in the bottom panel. We notice that because of the use of the bonus in circumstances when the nonperfect forecast results in wind curtailment, the score of the model is larger in the bottom panel of Fig. A7 compared to the top panel. The skill score of the RAP model remains positive for all forecast hours and for all three methods, and does decreases with forecast length.

REFERENCES

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. [Available online at https://hal-mines-paristech.archives-ouvertes.fr/hal-00765885/document.]

Bradford, K. T., R. L. Carpenter, and B. L. Shaw, 2010: Forecasting southern plains wind ramp events using the WRF model at 3-km. Preprints, *Ninth Annual Student Conf.*, Atlanta, GA, Amer. Meteor. Soc., S30. [Available online at https://ams.confex.com/ams/pdfpapers/166661.pdf.]

Bristol, E. H., 1990: Swinging door trending: Adaptive trend recording. *Proc. Int. Studies Association National Conf.*, New Orleans, LA, ISA, 749–753. [Available online at http://ebristoliclrga.com/PDF/SwDr.pdf.]

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, doi: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. [Available online at http://ceeesa.es.anl.gov/pubs/69166.pdf.]

Florita, A., B.-M. Hodge, and K. Orwig, 2013: Identifying wind and solar ramping events. *IEEE Green Technologies Conf.*, Denver, Colorado, IEEE, NREL/CP-5500-57447. [Available online at http://www.nrel.gov/docs/fy13osti/57447.pdf.]

Francis, N., 2008: Predicting sudden changes in wind power generation. *North American WindPower*, October, North American WindPower, Oxford, CT, 58 pp. [Available online at https://www.awstruepower.com/assets/Predicting-Sudden-Changes-in-Wind-Power-Generation_NAWP_Oct20082.pdf.)

Freedman, J., M. Markus, and R. Penc, 2008: Analysis of west Texas wind plant ramp-up and ramp-down events. AWS Truewind Tech. Rep., Albany, NY, 26 pp. [Available online at http://interchange.puc.state.tx.us/WebApp/Interchange/Documents/33672_1014_580034.PDF.]

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

Hamill, T. M., 2014: Performance of operational model precipitation forecast guidance during the 2013 Colorado Front Range floods. *Mon. Wea. Rev.*, **142**, 2609–2618, doi:10.1175/MWR-D-14-00007.1.

International Electrotechnical Commission, 2007: Wind turbines - Part 12-1: Power performance measurements of electricity producing wind turbines. IEC 61400-12-1, 90 pp.

Isaac, G. A., and Coauthors, 2014: The Canadian Airport Nowcasting System (CAN-Now). *Meteor. Appl.*, **21**, 30–49, doi:10.1002/met.1342.

Jacobs, A. J. M., and N. Maat, 2005: Numerical guidance methods for decision support in aviation meteorological forecasting. *Wea. Forecasting*, **20**, 82–100, doi:10.1175/WAF-827.1.

Kamath, C., 2010: Understanding wind ramp events through analysis of historical data. *Proc. Transmission and Distribution Conf. and Exposition*, New Orleans, LA, IEEE Power and Energy Society, doi:10.1109/TDC.2010.5484508.

Wandishin, M. S., G. J. Layne, B. J. Etherton, and M. A. Petty, 2014: Challenges of incorporating the event-based perspective into verification techniques. *Proc. 22nd Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., 4.4. [Available online at https://ams.confex.com/ams/94Annual/webprogram/Paper240097.html.]

Wilczak, J. M., L. Bianco, J. Olson, I. Djalalova, J. Carley, S. Benjamin, and M. Marquis, 2014: The Wind Forecast Improvement Project (WFIP): A public/private partnership for improving short term wind energy forecasts and quantifying the benefits of utility operations. NOAA Final Tech. Rep. to DOE, Award DE-EE0003080, 159 pp. [Available online at http://energy.gov/sites/prod/files/2014/05/f15/wfipandnoaafinalreport.pdf.]

——, 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, doi:10.1175/BAMS-D-14-00107.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 627 pp.

Wong, W.-K., C.-S. Lau, and P.-W. Chan, 2013: Aviation Model: A fine-scale numerical weather prediction system for aviation applications at the Hong Kong International Airport. *Adv. Meteor.*, **2013**, 532475, doi:10.1155/2013/532475.

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.

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

Zhang, J., A. Florita, B.-M. Hodge, and J. Freedman, 2014: Ramp forecasting performance from improved short-term wind power forecasting. *Int. Design Engineering Technical Conf./Computers and Information in Engineering Conf.*, Buffalo, NY, American Society of Mechanical Engineers, NREL/CP-5D00-61730. [Available online at http://www.nrel.gov/docs/fy14osti/61730.pdf.]