ANL/DIS-10-13

A Survey on Wind Power Ramp Forecasting

Decision and Information Sciences Division

About Argonne National Laboratory
Argonne is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC under contract DE-AC02-06CH11357. The Laboratory's main facility is outside Chicago, at 9700 South Cass Avenue, Argonne, Illinois 60439. For information about Argonne and its pioneering science and technology programs, see www.anl.gov.

Availability of This Report
This report is available, at no cost, at http://www.osti.gov/bridge. It is also available on paper to the U.S. Department of Energy and its contractors, for a processing fee, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
phone (865) 576-8401
fax (865) 576-5728
reports@adonis.osti.gov

ANL/DIS-10-13

A Survey on Wind Power Ramp Forecasting

by
C. Ferreira,[1] J. Gama,[1] L. Matias,[1] A. Botterud,[2] and J. Wang[2]
[1]Institute for Systems and Computer Engineering of Porto L.A. (INESC Porto L.A.)
[2]Decision and Information Sciences Division, Argonne National Laboratory

sponsored by
U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy,
Wind & Water Power Program

December 2010

CONTENTS

FIGURES

TABLE

ACRONYMS

3D	three-dimensional
AE	absolute error
ANFIS	Adaptive Neuro-Fuzzy Inference System
BS	Brier score
BSS	Brier skill score
CDF	cumulative density function
CSI	critical success index
EDS	extreme dependency score
EPS	ensemble prediction system
ERCOT	Electric Reliability Council of Texas
F	false alarm rate
FN	false negative, a miss
FP	false positive
GH	Garrad Hassan
H	hit rate
KSS	Hannsen & Kuipers Skill Score
MAE	mean absolute error
MSE	mean square error
MSEPS	Multi-Scheme Ensemble Prediction System
NCAR	National Center for Atmospheric Research
NWP	numerical weather prediction
OR	odds ratio
ORSS	odds ratio skill score
PCAP	purchase of ancillary services
RMSE	root mean square error
ROC	receiver operating characteristic
RPS	ranked probability score
RPSS	ranked probability skill score
RRE	ramp rate event
SCADA	supervisory control and data acquisition

Std	standard deviation
SVM	support vector machine
TN	true negative, i.e. non-occurring event
TP	true positive

Units of Measure

kW	kilowatt(s)
MW	megawatt(s)

ACKNOWLEDGMENTS

This report has been prepared by Argonne National Laboratory in collaboration with INESC Porto, Portugal. The INESC Porto team acknowledges the assistance of Vladimiro Miranda, Ricardo Bessa, Hrvoje Keko, Leonel Carvalho, Jean Sumaili, and Joana Mendes in the preparation of this report.

The authors acknowledge the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy through its Wind & Water Power Program for funding the research presented in this report under contract DE-AC02-06CH11357.

Argonne National Laboratory, December 10, 2010.

This page intentionally blank

ABSTRACT

The increasing use of wind power as a source of electricity poses new challenges with regard to both power production and load balance in the electricity grid. This new source of energy is volatile and highly variable. The only way to integrate such power into the grid is to develop reliable and accurate wind power forecasting systems. Electricity generated from wind power can be highly variable at several different timescales: sub-hourly, hourly, daily, and seasonally. Wind energy, like other electricity sources, must be scheduled. Although wind power forecasting methods are used, the ability to predict wind plant output remains relatively low for short-term operation. Because instantaneous electrical generation and consumption must remain in balance to maintain grid stability, wind power's variability can present substantial challenges when large amounts of wind power are incorporated into a grid system. A critical issue is ramp events, which are sudden and large changes (increases or decreases) in wind power. This report presents an overview of current ramp definitions and state-of-the-art approaches in ramp event forecasting.

1 INTRODUCTION

One of the major issues in wind power generation is dealing with ramp events. These events are characterized by sudden and large changes (increases or decreases) in wind power. To address these events, wind power operators, utilities, and system operators have to develop procedures that satisfy the electricity demand, as well as maximize both economic and environmental benefits. The sooner these events can be predicted, the more effective the procedures will be. To deal with a ramp-up event, a wind power producer can shut down turbines to avoid producing an excess of energy that cannot be compensated for by a sudden decrease in thermal generation, or it can increase its generation in agreement with the system operator and utilities. In the latter case, utilities can trade fossil energy costs by buying cheaper and clean wind energy. In a ramp-down event, the system operator can switch on fast hydro units or, if this procedure does not generate enough power to meet demand or is not available, the operator can use fossil energy turbines to meet the load (Parkes et al. 2006). If these measures are not sufficient, load curtailment must be adopted—a scenario that system operators obviously try to avoid.

In recent years, with the large-scale expansion of wind farms and turbine technology, the percentage of energy obtained from wind sources relative to the peak load is rapidly increasing. Thus, the demand for a more reliable wind power energy is driving the critical need to detect and predict ramp events (Kamath 2010). For instance, Francis (2008) reported a rapid and large ramp-down event that occurred in the Electric Reliability Council of Texas (ERCOT) operations area on February 26, 2008, that forced ERCOT to declare system emergency—a high-cost system condition.

Because this type of system condition can occur, both the accurate forecasting of ramp events and the quantification of ramp forecast accuracy are crucial to the large-scale integration of wind energy into electricity grids and to a better understanding of the risk involved in trades at times of high variability (Greaves et al. 2009).

One of the main problems in ramp forecasting is how to define a ramp. In fact, there is no standard formal definition (Kamath 2010; Focken and Lange 2008; Zheng and Kusiak 2009), and the existing literature reports different definitions, depending on, for instance, on the location and size of the wind farm.

For wind power forecasts, the existing models might be grouped under two different approaches (Monteiro et al. 2009):

1. *Physics-based models* are parametric models that are based on the physical characteristics of the weather and terrain. The main idea of these models is to translate and refine numerical weather prediction (NWP) forecasts into the wind power facilities' sites and to model local wind profiles. Moreover, these models use the wind farm theoretical power curve, or an estimated power curve, to forecast wind farm power output. For example, Greaves et al. (2009) and Focken and Lange (2008) use NWPs to produce forecasts of the power curves of the wind generation facilities.

2. *Statistical models* use historical wind power measurements, meteorological data, either NWPs or historical measurements, and machine learning algorithms to induce a predictive model. For example, Zheng and Kusiak (2009) use historical data collected by the supervisory control and data acquisition (SCADA) system as input for several regression models (e.g., regression trees) to predict power ramp rates.

This report presents the results from our study of the existing definitions of ramp events and methodologies to forecast ramp events. It is structured as follows. Section 2, on ramp event definitions, presents formal and informal definitions identified in the literature, as well as their characteristics and defined parameters. Section 3 presents some useful models to forecast ramps for different time horizons (i.e., short-term and long-term predictions). Section 4 summarizes the conclusions obtained from the study.

2 RAMP EVENT DEFINITIONS

Ramp forecasting is a relatively new research field. In order to study the ramp phenomena, it is important to define what could be considered a “ramp event.” According to the technical report for the Alberta Pilot Project (AWS Truewind 2008), a ramp occurs when there is a change in power output that has a large enough amplitude for a relatively short period of time. The same idea also appears in Greaves et al. (2009) and Cutler et al. (2007a). The expressions “swings,” “extreme events,” and “rapid changes” are also used synonymously with ramp events (Cutler et al. 2007b). These events can be detected locally at the wind farm level or in a wide geographical area that can include several wind farms. The events may cause severe grid management problems within the next few hours or days (Bradford et al. 2010).

Figure 1 illustrates the ramp definition presented in Greaves et al. (2009), which is a change in the wind farm power output that is at least 50% of the installed capacity and occurs within a time span of 4 hours or less.

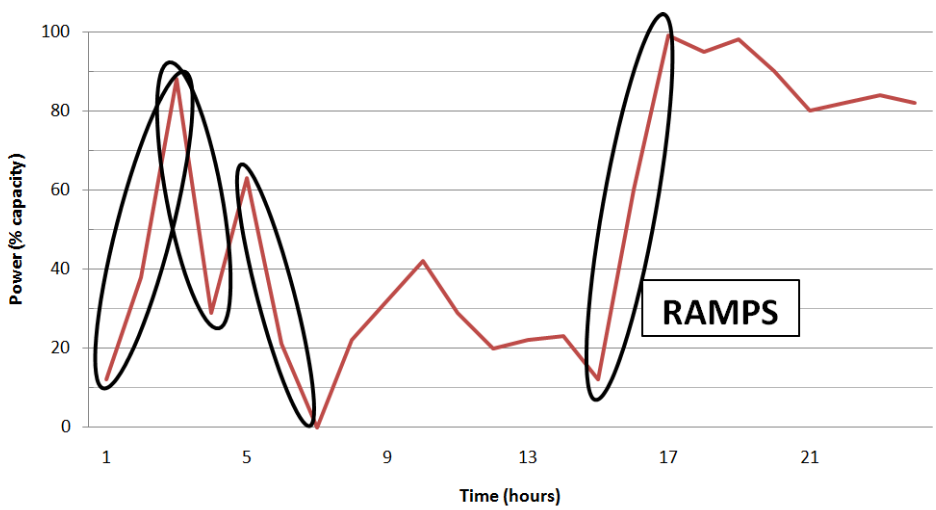

Figure 1 Ramp event definition: Change in power of at least 50% of the capacity over a maximum duration of 4 hours (Figure inspired by Parkes 2009; Greaves et al. 2009).

2.1 Characteristics for Ramp Definitions

Grimit and Potter (2008), Greaves et al. (2009), and Potter et al. (2009) outline several relevant characteristics for defining, characterizing, and identifying ramps. They state that to define a ramp, we have to determine values for its three key characteristics: direction, duration, and magnitude.

With respect to direction, there are two basic types of ramps: upward (or ramp-ups) and downward (or ramp-downs). Upward ramps are characterized by an increase of wind power, which might result from phenomena such as intense low-pressure systems (or cyclones), low-level jets, thunderstorms, wind gusts, or similar weather phenomena. The downward ramps result when there is a reduction in wind power (events that generally occur with the rapid slackening of a pressure gradient or the passage of a local pressure couplet) or when high-speed winds cause wind turbines to reach cut-out limits (typically 22 to 25 m/s) and shut down in order to protect the wind turbine from damage (Freedman et al. 2008). To be considered a ramp event, the minimum duration is assumed to be 1 hour in Potter et al. (2009); however, Kamath (2010) studies events in intervals of 5 to 60 minutes. The magnitude of a ramp event is typicallly represented by the percentage of the wind farm's nominal power (i.e., its name plate rating).

Duration and magnitude are usually related. For example, Potter et al. (2009) considers rapid ramp events to be when the change in power between two consecutive hours is greater than or equal to 10% of the nominal capacity of the wind farm. In addition, the AWS Truewind (2008) report suggests that:

- An important downward ramp occurs only if the power change in 1 hour is at least 15% of total capacity, and

- An important upward ramp occurs if the power change in 1 hour is at least 20% of total capacity.

Some ramp event definitions using these characteristics are presented in Section 2.2.

2.2 Ramp Definitions

Although it is easy to identify ramps visually, there is no consensus on the accepted formal definition of a ramp event (Kamath 2010; Focken and Lange 2008; Zheng and Kusiak 2009). This section presents four ramp event definitions. As stated above, a ramp event can be characterized according to three features: direction, magnitude, and duration. However, if we consider that the ramp magnitude values range from positive to negative, we can characterize a ramp by using only magnitude and duration features. The sign of the magnitude value can give us the ramp direction: positive magnitude values correspond to upward ramps, and negative magnitude values correspond to downward ramps.

In the definitions below, a ramp event can be identified according to the power signal *P(t)* and two user-defined parameters (one of the definitions requires only one parameter to identify a ramp). The first user-defined parameter Δ_t is related to the ramp duration and defines the size of the time interval considered to identify a ramp; it is usually measured in minutes or hours. Potter et al. (2009) and Zack et al. (2010) present some results that relate this parameter to the type and magnitude of identified ramps. The second user-defined parameter P_{val} is related to the ramp magnitude feature and provides a cut-off level on the power changes. The P_{val} parameter is usually defined according to the specific features of the wind farm site, and this threshold value is usually defined according to the amount of wind power capacity installed and as a percentage of the wind power nominal capacity or a specified number of megawatts (MW). Kamath (2010) claims that defining the P_{val} parameter according to the wind farm's nominal capacity can produce unreliable results. They consider that the nominal capacity of a wind farm is always changing, because at each moment, one or several units can be turned off. Their results are based on a simple analysis of historical measurements, in which the sensitivity of the two ramp definitions to each of the two parameters introduced above are studied: P_{val}, ranging from 150 to 600 MW, and Δ_t values, which varies between 5 and 60 minutes.

The first definition, provided below, was formally presented in Kamath (2010).

Definition 1: A ramp event is considered to occur at the start of an interval if the magnitude of the increase or decrease in the power signal at time Δ_t ahead of the interval is greater than a predefined threshold value, P_{val}:

$$|\mathrm{P}(\mathrm{t}+\Delta_{\mathrm{t}})-\mathrm{P}(\mathrm{t})|>\mathrm{P}_{\mathrm{val}} \tag{2.1}$$

This inequality considers only the values on the end points of the interval and ignores the ramps that occur in the middle. To address this issue, Kamath (2010) extends the previous definition.

Definition 2: A ramp event is considered to occur in a time interval Δ_t if the difference between the maximum and the minimum power output measured in that interval is greater than a threshold value, P_{val}:

$$\max(\mathrm{P}[\mathrm{t},\mathrm{t}+\Delta_{\mathrm{t}}])-\min(\mathrm{P}[\mathrm{t},\mathrm{t}+\Delta_{\mathrm{t}}])>\mathrm{P}_{\mathrm{val}} \tag{2.2}$$

This inequality considers the total magnitude of the power fluctuation through the interval. However, this definition does not consider the curve's slope (i.e., how quickly the power output decreases or increases). In order to analyze this important factor, we cannot consider an absolute threshold like P_{val}; we need a time-relative threshold.

A more elaborate definition considers the rate of change in power output over a period of time (Zheng and Kusiak 2009). To present this defnition, we define the "power ramp rate" or "slope" to be the rate of change of the power with respect to time. This measure is expressed in megawatts per minute (MWm^{-1}).

Definition 3: A ramp event is considered to occur if the ratio between the absolute difference of the power measured at two time points (the initial and final points of a time interval Δ_t) and the size of the time interval Δ_t is greater than a predefined reference value (i.e., the power ramp rate value or PRR_{val}):

$$\frac{|\mathrm{P}(\mathrm{t}+\Delta_{\mathrm{t}})-\mathrm{P}(\mathrm{t})|}{\Delta_{\mathrm{t}}}>PRR_{\mathrm{val}} \tag{2.3}$$

In the definitions presented in equations 2.1 and 2.3, we can easily identify the type of ramp: if $\mathrm{P}(\mathrm{t})>P(\mathrm{t}+\Delta_{\mathrm{t}})$, we are analyzing a downward ramp, otherwise, it is an upward one. In equation 2.2, on the other hand, this distinction is not so clear. In this latter case, we can identify the type of the ramp by using the relative position of the extreme time points within the interval. If the maximum power output occurs after the minimum power output, we get an upward ramp; otherwise, we get a downward ramp.

While the definitions above work directly with the wind power signal, other approaches transform the signal into a more appropriate representation. A usual transformation consists of considering k-order differences in the power amplitude. This strategy is used, for example, in Bossavy et al. (2010). In such a treatment, p_t is the wind power time series and p_t^f is the associated transformed signal that was obtained according to the formula:

$$\mathrm{p_t^f} = \mathrm{mean}\{\mathrm{p_{t+h}} - \mathrm{p_{t+h-n_{am}}}; \mathrm{h} = 1, \ldots, \mathrm{n_{am}}\} \tag{2.4}$$

In this formula, the parameter n_{am} stands for the number of averaged power differences to consider.

Definition 4: A ramp event is said to occur in an interval if the absolute value of the filtered signal p_t^f exceeds a given threshold value P_{val}:

$$|\mathrm{p_t^f}| > P_{val} \tag{2.5}$$

If required, the ramp time is considered to be the interval point for which the filtered signal has its maximum magnitude.

This last definition was introduced in Bossavy et al. (2010). Figure 2 presents a wind power time series and two filtered signals, one for each value of the n_{am} parameter; that is, $n_{am} = 2$ and $n_{am} = 5$. By setting the P_{val} threshold to be 25% of the wind farm nominal capacity, the authors identify ramp events to be the points in the filtered signal that are above the threshold line. Ramp timing is identified as the point at which the filtered signal achieves a local maximum.

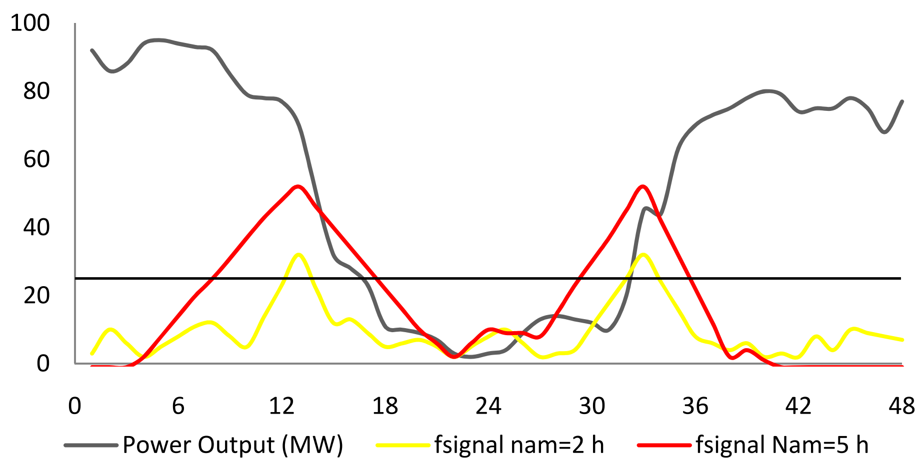

Figure 2 Example of wind power output time series and two transformed signals. By setting the threshold value to be 25% of the nominal power, we can identify the occurrence time of two ramp events, one at *t* = 13 hours and the other at *t* = 32 hours (Figure inspired by Bossavy et al. 2010).

In the following section, we analyze issues related to ramp forecasting and provide an overview of existing ramp forecasting models.

3 RAMP FORECASTING MODELS

Usually ramp forecasts deliver predictions for the next hours or days to the end users. Such information is used to support informed decisions about wind farm management, how much energy to trade or sell (Greaves et al. 2009), generation scheduling, etc. We can consider two main strategies:

- *Event detection models* use one of the previous ramp definitions over time series of wind power predictions or weather forecasts (Jørgensen and Mörlen 2008). In this case, the outputs of the forecast model are time-stamp series between 0 and 1, where 1 indicates the presence of a ramp, eventually paired with an associated probability.

- *Regression models* use historical data to predict ramps by using data mining techniques. Data mining algorithms learn a function of the form $\mathrm{y} = \mathrm{f}(\vec{\mathrm{x}})$ where $\vec{x}$ is a multivariate set of historical features and y is an output variable related to ramp events (Zheng and Kusiak 2009). In this case, the forecaster predicts a real number that is a function of the ramp magnitude.

3.1 Factors to Consider When Forecasting Ramps

Potter et al. (2009) and Bossavy et al. (2010) identify some errors that must be considered before one can start to build a ramp forecasting model:

- *Timing error (or phase error.)* Defined as an event whose magnitude is accurately predicted but that occurs at the "wrong" (unexpected) time.

- *Intensity error (or amplitude/magnitude error).* Defined as an event that is forecasted to occur at the "right" (expected) time but with the wrong magnitude.

- *Location error.* Defined as an error in the geographical location of the wind event. When forecasts for weather events are incorrectly located or when an event follows a different path than the one forecasted, the result might include timing and intensity errors.

Figure 3 plots both a "24-hour-ahead" forecast for the power output and the real value observed, thereby illustrating a timing error of 2 hours, as well as a magnitude error and a rate of change error.

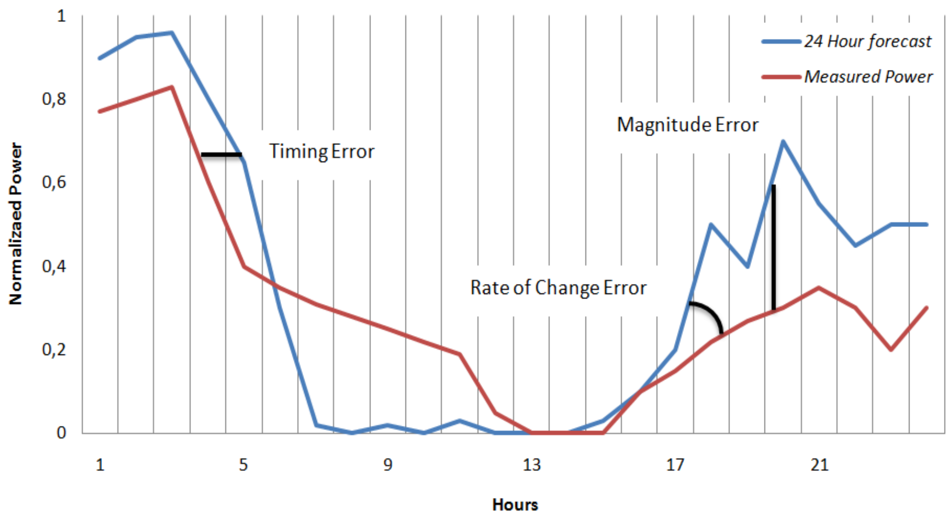

Figure 3 Plot illustrating three types of errors (timing, magnitude, and rate of change errors) between the 24-Hour-Ahead forecast and real values for the power output (Figure inspired by Potter et al. 2009).

3.2 Evaluation Metrics

We discuss two types of metrics: event detection metrics, which are used with event detection models, and predictive accuracy, which is used with regression models.

3.2.1 Metrics for Ramp Event Detection

This section presents metrics to evaluate event forecasters. First we present metrics to evaluate deterministic forecast systems, and then we present metrics used to evaluate probabilistic forecast systems.

The kappa coefficient (Cohen 1960) is a statistical measure of inter-rater agreement for qualitative (categorical) events. It is computed as:

$$k = \frac{P(a) - P(e)}{1 - P(e)} \tag{3.1}$$

where *P(a)* is the relative observed agreement among forecasters and *P(e)* is the hypothetical probability of chance agreement. The observed data are used to calculate the probabilities that each observer will randomly identify each category. If the forecasters are in complete agreement, then $k = 1$. If there is no agreement among the forecasters (other than what would be expected by chance), then k is <0.

Two other statistics that are widely used to evaluate the quality of deterministic forecast systems are precision and recall. Precision is defined as the ratio between the number of true positive events and the number of positive forecasts. Recall is defined as the ratio between the number of true positives and the number of observed positives. These statistics do not take into account the true negatives, which are positive aspects in predicting rare events like ramps.

To illustrate the computation of these metrics, Table 1 is a generic contingency table that summarizes the results of an event forecasting system. The four combinations of forecasts (yes or no) and observations (yes or no) are called the joint distribution; the total numbers of observed and forecasted occurrences and nonoccurences, which are shown on the lower and right sides of the contingency table, are called the marginal distribution.

Table 1 Contingency Table Representing Event Observation and Event Forecast

Event Forecast	Event Observation		Total
	Yes	*No*	
Yes	TP *(hits)*	FP *(false alarms)*	Forecast *Yes*
No	FN *(misses)*	TN	Forecast *No*
Total	Observed *Yes*	Observed *No*	N = TP + FP + FN + TN

TP = true positive; FP = false positive; FN = false negative; TN = true negative.

Formally, and using the illustrative table, we can write:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3.2}$$

Precision answers the question: What fraction of the predicted “yes” events really occurred?

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3.3}$$

Recall answers the question: What fraction of the observed “yes” events were correctly forecasted?

In order to assess the performance of a classification system, it can be useful to use a metric that combines precision and recall. The F_{score} is a weighted average combination (the harmonic mean) of precision and recall measures that gives equal weight to both measures, such that:

$$\text{F}_{\text{score}} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{3.4}$$

An F_{score} of 1 means that all events were detected, while a 0 score indicates that none of the events was detected. Other variants of this metric exist that can be used to consider unbalanced combinations of precision and recall.

In the context of ramp event detection, true negatives are irrelevant. Another useful metric is the critical success index (CSI), which is defined as:

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP+FN+FP}} \tag{3.5}$$

As in the case with precision and recall, the CSI metric takes values in the interval [0,1], where 1 means correct prediction. CSI measures the fraction of observed and/or forecasted events that were correctly predicted. It can be thought of as the accuracy when correct negatives have been removed from consideration; that is, CSI is only concerned with forecasts that count. Sensitive to hits, CSI penalizes both misses and false alarms. One of the projects that uses such a metric to evaluate ramp event forecast systems is described in Zack et al. (2010).

The bias score answers the question: How did the number of forecasted "yes" events compare with the number of observed "yes" events? It is defined as:

$$\text{Bias score} = \frac{\mathrm{TP+FP}}{\mathrm{TP+FN}} \tag{3.6}$$

The bias score measures the ratio of the frequency of forecast events to the frequency of observed events. It indicates whether the forecast system has a tendency to underforecast (bias score of <1) or overforecast (bias score of >1) events. The bias score does not measure how well the forecast corresponds to the observations; it only measures relative frequencies.

The Hanssen & Kuipers skill score (KSS), also known as Peirce's skill score or the true skill score (Hanssen and Kuipers 1965; Peirce 1884), is a widely used metric (see, for example, Bradford et al. 2010; Jørgensen and Mörlen 2008) that takes into account all of the elements of the contingency table. It measures the ability to separate the "yes" events from the "no" events. The KSS can be defined by means of both the hit rate $\left(H = \frac{TP}{TP+FN}\right)$ and the false alarm rate $\left(F = \frac{FP}{FP+TN}\right)$ as follows:

$$\mathrm{KSS} = \mathrm{H} - \mathrm{F} = \frac{\mathrm{TP*TN-FP*FN}}{\mathrm{(TP+FN)(FP+TN)}} \tag{3.7}$$

The KSS takes values in the interval [–1,1], where 0 indicates no skill and 1 is the perfect score.

When predicting rare events having large TN values, the KSS approaches the value of the hit rate, thus becoming vulnerable to hedging, a strategy that consists of always forecasting event occurrence. This score is more appropriate to verify frequently occurring events.

A special purpose score aimed at verifying predictions of rare events is the extreme dependency score (EDS) (Coles et al. 1999). This metric does not account for either the false positive alarms (FPs) or the nonoccurring events (TNs) but considers the sample size (n) (i.e., the total number of cases), as follows:

$$\mathrm{EDS} = \frac{2\log((\mathrm{TP+FN})/\mathrm{n})}{\log(\mathrm{TP/n})} - 1 \tag{3.8}$$

While this metric has some helpful properties, it does not tend to zero for vanishing events, and it is not explicitly dependent on the bias, although it is sensitive to hedging. The EDS score range is [–1,1], where –1 is the worst score and 1 is the perfect score. The EDS takes a value of –1 when the base rate $\left(BR = \frac{TP+FN}{n}\right)$ takes a value of 1, and it takes a value of 1 when the hit rate (H) takes a value of 1. It answers the question: What is the association between forecasted and observed rare events?

In Ghelli and Primo (2009), the EDS is used to investigate the performance of NWP in predicting rare precipitation events. The EDS score prove to be effective in identifying rare events, although some care should be taken with regard to hedging. Our recommendation is that the EDS should not be the only score used to assess the performance of a forecasting model and that clear improvements could be achieved if the EDS could be combined with the false alarm rate and/or the bias score metrics.

One other metric that is suitable for analyzing rare events and is less sensitive to hedging is the odds ratio (OR) (Jørgensen and Mörlen 2008; Stephenson 2000). This metric can easily be defined in terms of the hit rate (H) and false alarm rate (F),

$$\mathrm{OR} = \frac{\mathrm{H}/(1-\mathrm{H})}{\mathrm{F}/(1-\mathrm{F})} \tag{3.9}$$

The OR formula presents the ratio between the odds of making a hit to the odds of making a false alarm. The score of OR ranges from 0 to ∞. The perfect OR score is ∞, and scores higher than 1 mean that the hit rate exceeds the false alarm rate. This measure answers the question: What is the ratio between the odds of making a good forecast and the odds of making a bad forecast?

In Yule (1900), the odds ratio skill score (ORSS) is introduced in order to measure the association in contingency tables. The ORSS is a function of the odds ratio whose value ranges from –1 to 1. The value 1 corresponds to a perfect forecast skill, whereas the value 0 corresponds to a null skill. The ORSS can be defined in terms of the OR as:

$$\mathrm{ORSS} = \frac{\mathrm{OR}-1}{\mathrm{OR}+1} \tag{3.10}$$

The ORSS answers the question: How much gain was obtained by using the forecasting system compared with using a random prediction strategy?

Some ORSS properties are presented in Stephenson (2000), including, among others, the nondependence of the marginals and therefore the difficulty of hedging the score. Moreover, the author shows that the combination of ORSS, KSS, and the bias score form a complete set to describe a 2×2 contingency table.

3.2.2 Probabilistic Forecast

The metrics presented above are among those used to evaluate deterministic forecasts; however, by applying simple procedures, they could be used to evaluate probabilistic forecasting systems as well. A probabilistic forecast assigns a probability to the prediction. One evident advantage of probabilistic forecasts is the introduction of a degree of freedom, by, for example, using a threshold regarding the probability of whether to decide the ocurrence of events. One technique that can be used to choose the optimal threshold is the receiver operating characteristic (ROC) curve.

3.2.2.1 ROC Space

The ROC curve is a graphical plot of sensitivity versus specificity for a binary classifier system, when its discrimination threshold is varied. The ROC is obtained by plotting the fraction of true positives versus the fraction of false positives as the criterion threshold changes (Hand 2009). The best possible prediction method would yield a point in the upper left corner—or coordinate (0, 1) of the ROC space—representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The diagonal of the ROC space corresponds to a random guess. Points above the diagonal correspond to predictions better than does a random choice, while points below correspond to predictions worse than does a random guess. A deterministic forecast defines a point in the ROC space. A probabilistc forecast defines a trajectory in the ROC space, where each point in the trajectory corresponds to a different probability threshold. Moreover, in a classification problem, each ROC curve is independent of the class distribution and error cost, which is the cost of predicting the wrong class (Provost and Fawcett 1997). We can use the ROC curve to find the optimal operating point (i.e., a probability threshold) under varying costs of making incorrect predictions, FPs, and misses (FNs) (see Figure 4).

3.2.2.2 Metrics for Probabilistic Forecasts

For forecasts that assign a probabilty to each event, a more informed metric might be used. The Brier score (Brier 1950) is a score function that measures the accuracy of a set of probability assessments. It measures the average squared deviation between predicted probabilities for a set of events and their outcomes. It is computed as:

$$\mathrm{BS} = \frac{1}{\mathrm{N}}\sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{F_t} - \mathrm{O_t})^2 \tag{3.11}$$

where F_t is the probability that was forecasted, O_t is the actual outcome of the event at instance t (0 if it does not happen and 1 if it happens), and N is the number of forecasting instances. A lower score represents higher accuracy.

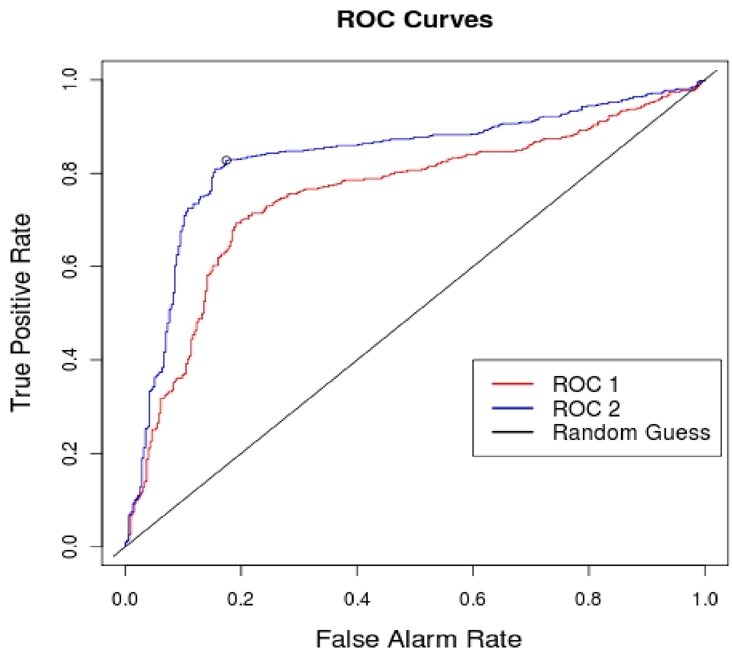

Figure 4 Representation of ROC curves for two forecasting systems, and the diagonal of the ROC space (black line) which corresponds to a random guess. The small circle identifies the best operating point, based on user-defined error costs, of the forecasting system associated with the blue curve.

Bossavy et al. (2010) compute BS and the Brier skill score (BSS) to compare the performance of the proposed forecast system against a reference methodology, a climatology model. The BSS is computed according to the formula:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{ref}}} \tag{3.12}$$

where $\mathrm{BS_{ref}}$ and BS stand for the Brier score of the reference model and the Brier score of the proposed model, respectively. The BSS is a particular case of the ranked probability score (RPS) (Brier 1950; Murphy 1971).

The RPS is a score used to assess the performance of multicategory probabilistic forecasting systems. The RPS compares the forecast cumulative density function (CDF) of the probabilistic forecast against the corresponding observed CDF. If we consider the discrete case, we can define:

$$\mathrm{RPS} = \frac{1}{\mathrm{K}-1}\sum_{\mathrm{k}=1}^{\mathrm{K}}(\mathrm{F_k} - \mathrm{O_k})^2 \tag{3.13}$$

where K is the number of forecasting categories and $\mathrm{F_k} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{f_i}$ and $\mathrm{O_k} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{o_i}$ are the k components of the forecast and observed CDF, respectively. The f_i is the forecast probability of an event to occur in category i, and o_i is a binary valued variable that takes a value of 1 if the event is observed in category i; otherwise, it takes a value of 0.

Zack et al. (2010) compute the RPS and the ranked probability skill score (RPSS) (Weigel et al. 2007) to evaluate the developed probabilistic ramp rate forecasts. The RPSS is computed in order to compare the forecasts against the reference model, a climatology model. The RPSS is defined as:

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS_{ref}}} \tag{3.14}$$

where $\mathrm{RPS_{ref}}$ and RPS stand, respectively, for the ranked probability score of the reference model and the ranked probability score of the proposed model.

By using each of the above metrics, we can now quantitatively assess the value of a forecast system. There are other techniques that might be used to visualize, and hence further analyze, the quality of probabilistic forecasts. Reliability diagrams, presented in Bröcker and Smith (2007), can be used to analyze the properties of a probabilistic forecasting system. The reliability diagrams plot the observed relative frequency against the forecasted relative frequency. We consider Y_i the actual outcome of the event at instance i (0 if it does not happen and 1 if it happens) and X_i the forecast probability that Y_i will be equal to 1. If we partition the forecast probabilities into k disjoint bins B_k, we define $\mathrm{I_k} = \{\mathrm{i}; \mathrm{X_i} \in \mathrm{B_k}\}$ to be the set of events with a probability value in the B_k bin. In addition, we define the observed relative frequency f_k and the forecast relative frequency r_k to be, respectively:

$$\mathrm{f_k} = \frac{\sum_{\mathrm{i} \in \mathrm{I_k}} \mathrm{Y_i}}{\#\mathrm{I_k}}, \tag{3.15}$$

$$\mathrm{r_k} = \frac{\sum_{\mathrm{i} \in \mathrm{I_k}} \mathrm{X_i}}{\#\mathrm{I_k}} \tag{3.16}$$

where $\#\mathrm{I_k}$ is the number of instances having a forecast probability value in the bin B_k.

The reliability diagram plots f_k against r_k. Figure 5 reproduces a reliabily diagram presented in Potter et al. (2009) to evaluate a probabilistic ramp rate event (RRE) forecast system. We can see that by not considering phase errors, the RRE performance (blue line) aligns closely in terms of forecast reliability to the climatology forecast. If we allow phase errors, by considering a time window of 3 hours, the probabilistic forecasting systems track more closely with the perfect reliability.

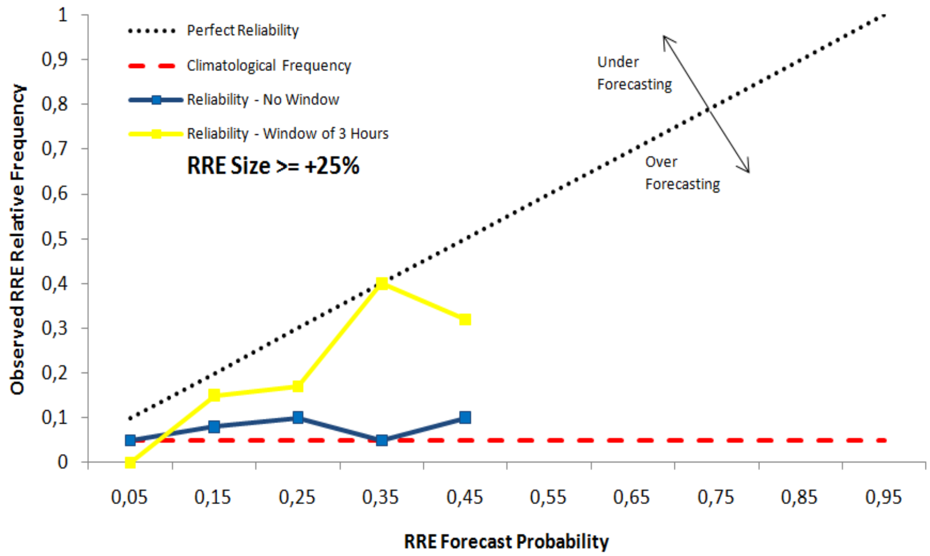

Figure 5 Reliability diagram from RRE forecast system showing the reliability of forecasts for ramps greater than or equal to 25% of nameplate capacity (Figure inspired by Potter et al. 2009).

3.2.3 Metrics to Assess Forecast Accuracy

Approaches based on data mining map the ramp prediction to a regression problem. The output is a real number, and the predictive accuracy is a function of the difference between the forecasted value and the observed value . Potter et al. (2009) observe that metrics based on mean square error (MSE) are not appropriate for ramp forecasting assessment. The MSE, root mean square error (RMSE), and other MSE-based metrics tend to over-penalize large errors. This occurs as a result of the squaring of each term, which effectively weights large errors more heavily than small ones. This property, which is undesirable in many applications, such as in ramp detection (Jørgensen and Mörlen 2008), has led researchers to use such alternatives as the mean absolute error (MAE) or those based on the median.

Two main metrics, the MAE and the standard deviation (Std) of the absolute error, AE, were used to measure the prediction accuracy of different regression algorithms in Zheng and Kusiak (2009). Small values of MAE and Std imply superior prediction performance. In fact, MAE and Std based on absolute error are widely used in the wind industry. These metrics are formally defined in the following formulas:

$$\mathrm{AE}(t) = \left|\hat{y}(t) - y(t)\right| \tag{3.17}$$

$$\mathrm{MAE} = \frac{\sum_{t=1}^{N} \mathrm{AE}(t)}{N} \tag{3.18}$$

$$\mathrm{Std} = \sqrt{\frac{\sum_{t=1}^{N} (\mathrm{AE}(t) - \mathrm{MAE})}{N - 1}} \tag{3.19}$$

where $\hat{y}(t)$ is the prediction value, *y(t)* is the observed value at instance t, and N is the number of prediction points.

3.2.4 Forecasts and Economic Value

The metrics presented in the previous two sections are used to quantify a forecasting system's accuracy. This section considers the economic performance of different types of forecasting systems, focusing on the integration of ramp event forecasting systems into the operation of the electricity market. Considering that no forecasting system works as a standalone component, the development and integration of any system should take into account other components, such as electricity generation, reserves, electricity costs, and the like.

In Grimit and Potter (2008) and Potter et al. (2009), a cost function is defined in order to compare the economic performance (i.e., the optimum application range) of different types of event forecasting systems. The comparison is made by estimating the total cost of ancillary services over time as a function of the ratio between the cost of pre-purchasing ancillary services and the cost of not protecting (the partial cost of advance purchase; PCAP), as follows:

$$\mathrm{Cost(PCAP)} = \mathrm{H} + \mathrm{F} + \frac{\mathrm{M}}{\mathrm{PCAP}} \tag{3.20}$$

The predictions of each system were cast into contingency table categories of hit or true positive *H*; a miss or false negative *M*; a false alarm or false positive *F*; and correct negatives, for ramp event sizes of at least 10% of installed capacity between two consecutive hours. Each of these forecast predictions has an associated cost. (Because the cost associated with the correct negatives is 0, these values do not appear in the above formula.)

The goal is to find the PCAP intervals where each forecast system outperforms the others (i.e., where each event forecast system obtains the minimum expenditure(s) on ancillary services). Three forecasting systems were compared: a climatology-based forecast, the probabilistic rapid ramp event (RRE) forecast, and a deterministic system. Note that the hits, misses, and false positive counts themselves are also a function of the PCAP value for the probabilistic RRE forecast. When using this procedure, the authors found four PCAP intervals, each of which is associated with a protection strategy, as shown in Figure 6.

The figure reports two extreme cases. When the PCAP ratio is a small value, i.e. when the cost of purchasing ancillary services is much smaller than the cost of not protecting, then the strategy is always one of protecting. In contrast, if the PCAP ratio is high, then the strategy is never to protect against ramp events. In the middle of these two extreme cases, the minimum cost is achieved by using the RRE probabilistic forecast system or a deterministic forecast system.

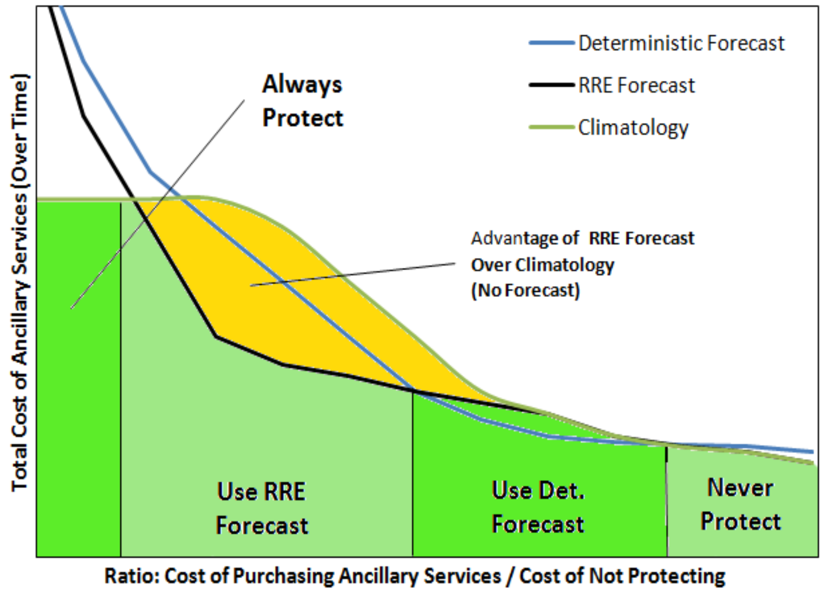

Figure 6 Conceptual diagram showing the relative value of the RRE forecast compared with a deterministic forecast and climatological frequency (Figure inspired by Potter et al. 2009).

3.3 Methodologies and Systems to Forecast Ramp Events

To address the ramp event forecast problem, a set of issues needs to be considered. Of these, some of the most important are defining the forecast goal, the time scale, the aggregation methodology, and the forecasting techniques to be used. This section briefly introduces each of these issues and and presents relevant work in the literature that addresses the ramp event forecasting problem. We emphasize that ramp event forecasting is a new and evolving field, and that new ideas and approaches are proposed and tested at a rapid pace.

3.3.1 Issues in Ramp Forecasting

When addressing a ramp event forecasting problem, there are two major factors to be considered. The first is the goal of the forecast: to predict wind speed or to predict wind power output. In the latter, we directly obtain wind power output predictions, whereas in the former, we must use turbine models (power curves) to convert wind speed forecasts into wind power output (Negnevitsky and Johnson 2008). The other major factor is the time horizon of the predictions. It is well known that the reliability of forecasts decreases as the time horizon increases (Cutler et al. 2007b). Therefore, the stochastic nature of ramp events makes it almost impossible to generate reliable ramp event forecasts for horizons longer than 48 hours.

Depending on the goal and the time horizon, we must carefully choose the most suitable data. We can use NWPs, meteorologial measurements, and/or power measurements, among others. Negnevitsky and Johnson (2008) argue that NWP-based models have proven to be unsuitable for time horizon predictions shorter than 6 hours. Moreover, they state that for these short horizons, it is best to use historical measurements. We can combine different sources of data; for instance, we can use NWPs from different providers and individual or aggregated wind farm data. Zack et al. (2010) report that for very short-term predictions (i.e., mainly for ramp event detection), NWP provides little value. This issue results from the low frequency of updates (typically issued only every 6 hours) and from the low vertical and horizontal resolutions of the NWPs. Sørensen et al. (2008) model power fluctuations of a large, off-shore wind farm by using two models: one that models each individual turbine and another that models the wind farm by using an averaged representative turbine. By using the representative turbine, the authors obtain a good compromise between accuracy and simulation time. The effectivness of these aggregation strategies is also heavily dependent on a farm's location. Kariniotakis et al. (2004) evaluate the performance of 11 forecasting systems on sites located in four different countries. The sites were chosen in order to cover a wide range of weather conditions and terrain complexity. The authors conclude that the performance of the models heavily depends on the resolution of the NWPs and the complexity of the terrain (e.g., orography). The higher the terrain complexity, the higher the resolution requirements.

While there is a wide range of forecasting techniques—classification or regression, supervised or nonsupervised, linear or nonlinear, parametric or nonparametric—there is still no overall best technique for ramp forecasting. Zack et al. (2010) and Negnevitsky and Johnson (2008) observe that the forecasting technique should be selected according to the nature of the time series data. For example, for a time series that is steady or involves small random variations, a persistence model will be appropriate for very short-term prediction horizons. On the other hand, if the time series data have strong variations or high volatility, then it could be effective to use a nonlinear model, such as Adaptive Neuro-Fuzzy Inference System (ANFIS), even for very short-term predictions. Negnevitsky and Johnson (2008) also propose an online methodology that uses a prior performance analysis of the most interesting models. The online method identifies the concept drift in the time series and changes the forecasting technique accordingly.

Another relevant issue is the type of predictions (either deterministic or probabilistic) produced by the forecasting system. Currently, most of the available forecasting systems provide deterministic forecasts, also known as spot or point forecasts. Juban et al. (2007) state that this approach removes any uncertainty concerning the predicted values. Hence, hard constraints are added to their decision-making applications. The authors propose a probabilistic prediction model, based on kernel density estimators, that provides a probability density function for each time horizon. Furthermore, it has been shown in Grimit and Potter (2008), Potter et al. (2009) and Pinson (2006) that by using probabilistic forecasting systems, higher economic benefits can be obtained. The economic benefits of using a special-purpose, probabilistic ramp event forecasting system are analyzed in Grimit and Potter (2008) and Potter et al. (2009).

3.3.2 Ramp Event Forecasting Models

In Greaves et al. (2009), researchers from Garrad Hassan (GH) present results of an investigation that studies methodologies to predict ramp events, including the temporal uncertainty associated with such events.

In this work, the authors define a ramp event to be a power output change that exceeds the nameplate capacity by more than 50% and that occurs over a period of 4 hours or less. Regarding the methodology used to compute forecast uncertainty, and despite the vagueness of the discussion, the authors combine multiple NWP inputs, statistical processing, and adaptive algorithms. In order to evaluate the methodologies developed, the authors define two metrics, forecast accuracy and ramp capture; these metrics correspond to the precision and recall metrics, respectively, defined in Section 3.2.1. These metrics do not consider the true negative events. Moreover, the authors include the phase error when identifying a ramp event. A true forecast is defined to be a forecast ramp with a measured ramp of the same direction within ± 12 hours of the time of the forecast ramp. The authors stated that a range of ± 12 hours is the maximum time difference to give sufficient data for temporal uncertainty analysis whilst maintaining a realistic connection between forecast and measured ramp events.

To evaluate historical data of the proposed methodologies from GH services, they used data from individual wind farms in the United Kingdom (UK) and United States, as well as data from a portfolio of wind farms in Ireland. The data include GH power forecasts and data used by the GH forecaster to predict power output, including NWP and measured power. The proposed methodology was used to predict ramps in short and medium time horizons (i.e., 3- to 24-hour-ahead predictions). As expected, in general, precision and recall take higher values for short-time-ahead predictions. Moreover, in the portfolio of wind farms, recall takes higher values than those measured on individual farms, and, on the contrary, precision takes lower values on the portfolio. These results can be explained by the small number of large ramps that occur in a portfolio and by the fact that portfolio forecasts are more accurate. Furthermore, the same definition of a ramp event used to identify ramps in individual farms was applied to identify ramps in the portfolio of wind farms, which resulted in a much smoother signal. Greaves et al. (2009) observe that the performance of the methodology is sensitive to wind farm location and that ramp events occur more frequently in the wind farms located in the United States, all of which are located in one U.S. state.

The authors also analyze the method's sensitivity to the use of different sources of NWP data. The presented results show that the precision is higher when NWPs from different sources are combined. Moreover, the combination of NWPs from different sources always yields the best recall values when compared with using only a single NWP source.

Zheng and Kusiak (2009) combine feature selection and five data-mining algorithms to study a 10- to 60-minute-ahead prediction of power ramp rates. In this study, the authors use time-series power data obtained from a SCADA system that controls 100 turbines. The authors use the mean, standard deviation, maximum and the minimum wind speed of a turbine, and they also use the measured wind farm power and the power ramp rate of the wind farm. The power ramp rate is defined as the 10-min absolute difference in wind farm power (similar to definition 2.3 presented in Section 2.2). The training examples correspond to one month's worth of data, with a 10-minute time step. Considering that the huge number of predictors can degrade the performance, a boosting tree algorithm is used to select the most interesting features. The selected features are used to train five data mining algorithms: Multilayer Perceptron, Support Vector Machines (SVM), Random Forest, Classification and Regression Trees, and Pace Regression. All of the algorithms were trained for predicting a power ramp rate at 10 and

20 minutes ahead. Within these time horizons, the SVM algorithm produced results with the highest accuracy. Moreover, the SVM algorithm was used to forecast 30, 40, 50, and 60 minutes ahead. The MAE of the SVM algorithm ranges from 243 kW/minute for the 10-minute forecast to 459 kW/minute for the 60-minute forecast. The quality of the results was assessed by computing the mean absolute error and standard deviation. An issue arising from this work is that the data used to train a model predictor are the same, despite the different ranges of prediction horizons.

Bossavy et al. (2010) introduce a methodology that identifies ramps by mapping the initial wind power series into a signal that results from computing the average of time power differences (see equation 2.4). This methodology can be used to detect ramps with different resolutions and requires only the definition of a single parameter. The authors propose two probabilistic forecasting methods, grounded in this methodology, that aim to predict wind power output by using ramp information and to predict ramp timing. The first method that predicts wind power output uses information extracted from the translated space, ramp intensity, and ramp forecast time information. The other method that predicts ramp timing translates the signal of each of the members of an ensemble of wind power curves and then uses the ensemble votes to define a confidence interval for the ramp timing.

The above methods were evaluated empirically. The first method was used to predict wind power 48 hours ahead in two European wind farms. In this experiment, Bossavy et al. (2010) map the hourly measured data from the two farms and use that data as input to a quantile regression forest algorithm. Comparison against a basic method that does not use ramp information is unable to clarify the contribution of the methodology presented. The proposed methodology offers better results in the higher quantiles but less accurate results in the lower quantiles. The second method was used to predict the time of ramp occurrence 80 hours ahead. This experiment uses power measurements from three wind farms located in France and 51 meteorological forecasts of wind speed provided by the ensemble prediction system (EPS) model of the European Centre for Medium-Range Weather Forecasts. The performance of this method is compared against a climatology algorithm using the BSS. The quality of the results is related to the number of ensemble members predicting the interval time of ramp occurrence and the size of the time intervals. The larger the number of ensemble predictors and the wider the time interval, the more reliable the results.

Zack et al. (2010) present the methodologies developed by AWS Truewind (2008) to predict wind ramps between 0 and 6 hours ahead. The system capabilities include: a probabilistic ramp rate forecasting module that can predict ramp rate probabilities for different time resolutions and a wide range of ramp rates; a hybrid deterministic-probabilistic ramp event forecasting module that has deterministic values as its output; a confidence interval for the events satisfying the ramp event definition; and the ability to predict the average power production for 15-minute intervals. For this last task, the system uses ramp event forecasts and a deterministic-probabilistic approach to provide a confidence band for the predictions.

This system has four main components. The first is a high-resolution atmospheric analysis system that uses three-dimensional (3D) variational calculus analysis to create a 3D representation of the patterns of temperature, pressure, wind, and moisture atmospheric variables. This component uses surface weather observations (on-land and off-shore), vertical data profiles from weather balloons (radiosondes), and radial wind estimates from Doppler radars, as well as satellite imagery (infrared and ultraviolet) information. The second component is the use of a special NWP model, originally developed to predict severe thunderstorms and tornados, to generate high-resolution atmospheric forecast simulations in a rapid cycle mode that produces a 13-hour ahead forecast every 2 hours. The third major component is a methodology that associates algorithms with ramp types. Using 1 year's worth of data, the method identifies types of ramps and selects the algorithm having the better performance for each type of ramp. The fourth component is a set of regime-based statistical models that generate the deterministic or probabilistic forecasts by using data and information from the previous three components. The component learns models for significantly different weather conditions. At the forecasting time, the system identifies weather conditions and, according to the authors, selects the most suitable prediction model. The skills of this system were evaluated on an ERCOT area. The authors—by using (a) different metrics to assess the performance of probabilistic ramp rate forecasts, (b) MAE and RMSE to evaluate average power production, (c) CSI to evaluate ramp occurrence, and (d) RPSS to compare two forecast methodologies—proved that they could obtain better results (by 25%) than did the climatology reference model. Furthermore, it is reported that the quality of the ramp event forecasts is highly dependent on the size of the time interval considered to detect a ramp event. The longer the time interval, the higher the accuracy.

WEPROG is a Danish company that provides real-time uncertainty weather forecasts (WEPROG undated). WEPROG developed a special-purpose tool that provides forecast for wind ramps. The tool uses data from its Multi-Scheme Ensemble Prediction System (MSEPS), a system that uses 75 forecasts, from the application of a perturbation procedure over a single-kernel NWP model, to produce predictions with uncertainty, for several weather parameters. In a research project conducted in Alberta, Canada (Jørgensen and Mörlen 2008), MSEPS ensembles were used to identify extreme ramp events. In this project, the researchers define an extreme ramp event as a change in power of more than 80 MW over a 1-hour period and study important issues in ramp forecasting. Among others, the following important points were identified:

1. Ensembles that are RMSE optimal are too smooth to be useful for predicting extreme ramp events.

2. The higher the phase error allowed to identify an extreme ramp event, the higher the skill of the forecasting system. In this study, phase errors ranging between 0 and 4 hours were studied.

3. The diversity in the ensemble forecast is an important measure for the uncertainty of amplitude and the phase of steep ramps.

4. Approximately 90% of the extreme ramp events were forecasted, with the correct amplitude, by at least one ensemble member. Nevertheless, 20% of these events were detected with a timing error ranging from 1 to 3 hours.

5. The spatial resolution of the NWPs is directly related to the performance of the system, with higher spatial resolutions resulting in better forecasting skills. In this study, metrics

such as accuracy, hit rate, false alarm rate, KSS, and OR were computed. More recent commercial versions of the MSEPS include a ramp rate prediction module.

Recently, Pyle (2010) reported that Xcel Energy, a U.S. energy provider, along with Vaisala, a Finland-based company working on developing electronic measurement systems for environmental use, and the U.S.-based National Center for Atmospheric Research (NCAR) formed a joint collaboration to develop observation equipment that will capture high-resolution data and methodologies to provide advanced notification of wind ramps and decision support information. The high-resolution data, collected from a network of observation points, will be used as input to a mesoscale modeling system. These modeled data will then be used as input to a statistical and pattern recognition system that predicts wind ramp events. The final system, which was planned to be functional in 2010, is expected to predict wind ramps between 15 minutes to 3 hours ahead and to have ramp features, such as ramp timing, magnitude, and type (ramp up or down), as its outputs.

4 CONCLUSIONS

Accurate ramp event forecasting is an important and challenging topic. Ramp forecasting can become an important tool to help maintain the stability of the electrical grid. The research addressing this problem is rapidly increasing. This report focuses on recent literature concerning ramp event definitions, ramp event forecasting, and metrics to evaluate ramp event forecasting. Moreover, it identifies literature that reports on recent advances in ramp event forecasting. We should point out that many of the current breakthroughs and methodologies were developed in the context of commercial wind forecasting systems, which may explain the limited information available.

Ramp event detection is a problem that has emerged fairly recently, as signified by the fact that it has no established standards, including definitions, forecasting methods, and evaluation metrics. Furthermore, the few existing methodologies report results that tend to be unreliable and of low accuracy. These observations indicate that the field is wide open to research that pushes beyond the current state of the art. More work is therefore needed to develop new methods for probabilistic ramp forecasting and to design evaluation strategies for such forecasts.

This page intentionally blank

5 REFERENCES

AWS Truewind, 2008, "AWS Truewind's Final Report for the Alberta Forecasting Pilot Project," Alberta, Canada.

Bossavy, A., R. Girard, and G. Kariniotakis, 2010, "Forecasting Uncertainty Related to Ramps of Wind Power Production," in *Proceedings of the European Wind Energy Conference & Exhibition*, Warsaw, Poland, April.

Bradford, K.T., R.L. Carpenter, and B.L. Shaw, 2010, "Forecasting Southern Plains Wind Ramp Events Using the WRF Model at 3-km," in *Proceedings of the AMS Student Conference,* Atlanta, Georgia.

Brier, G.W., 1950, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78(1):1–3.

Bröcker, J., and L.A. Smith, 2007, "Increasing the Reliability of Reliability Diagrams," *Weather and Forecasting* 22(3):651–661.

Cohen, J., 1960, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* 20(1):37–46.

Coles, S., J. Heffernan, and J. Tawn, 1999, "Dependence Measures for Extreme Value Analyses," *Extremes* 2(4):339–365.

Cutler, N., M. Kay, H. Outhred, and I. McGill, 2007a, "High-Risk Scenarios for Wind Power Forecasting in Australia," in *Proceedings of the European Wind Energy Conference & Exhibition (EWEC),* Milan, Italy.

Cutler, N., M. Kay, K. Jacka, and T.S. Nielsen, 2007b, "Detecting, Categorizing, and Forecasting Large Ramps in Wind Farm Power Output Using Meteorological Observations and WPPT," *Wind Energy* 10(5):453–470.

Focken, U., and M. Lange, 2008, "Wind Power Forecasting Pilot Project in Alberta," energy & meteo systems, Oldenburg, Germany.

Francis, N., 2008, "Predicting Sudden Changes in Wind Power Generation," in *North American WindPower*, October, p. 58.

Freedman, J., M. Markus, and R. Penc, 2008, "Analysis of West Texas Wind Plant Ramp-up and Ramp-down Events," AWS Truewind LLC, Albany, N.Y.

Ghelli, A., and C. Primo, 2009, "On the Use of the Extreme Dependency Score to Investigate the Performance of an NWP Model for Rare Events," *Meteorological Applications* 16(4)537–544.

Greaves, B., J. Collins, J. Parkes, and A. Tindal, 2009, "Temporal Forecast Uncertainty for Ramp Events," in *Proceedings of the European Wind Energy Conference & Exhibition (EWEC)*, Marseille, France.

Grimit E. and Potter C., 2008, "A Prototype Day-Ahead Forecast System for Rapid Wind Ramp Events," in *Proceedings of Windpower 2008 Conference and Exhibition*, Houston, TX.

Hand, D.J., 2009, "Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve," *Machine Learning* 77(1):103–123.

Hanssen, A.W., and W.J.A. Kuipers, 1965, "On the Relationship between the Frequency of Rain and Various Meteorological Parameters," *Mededelingen van de Verhandlungen* 81(2–15).

Jørgensen, J.U., and C. Mörlen, 2008, "AESO Wind Power Forecasting Pilot Project," Technical Report 2008, WEPROG, Ebberup, Denmark.

Juban, J., N. Siebert, and G.N. Kariniotakis, 2007, "Probabilistic Short-term Wind Power Forecasting for the Optimal Management of Wind Generation," pp. 683–688 in *Proceedings of the* IEEE *PowerTech Conference*, Lausanne, Switzerland, July.

Kamath, C., 2010, "Understanding Wind Ramp Events through Analysis of Historical Data," pp. 1–6 in *Proceedings of the IEEE PES Transmission and Distribution Conference and Expo*, New Orleans, La., April.

Kariniotakis, G., I. Martí, D. Casas, P. Pinson, T.S. Nielsen, H. Madsen, G. Giebel, J. Usaola, I. Sanchez, A.M. Palomares, R. Brownsword, J. Tambke, U. Focken, M. Lange, P. Louka, G.Kallos, C. Lac, G. Sideratos, and G. Descombes, 2004, "What Performance Can Be Expected by Short-Term Wind Power Prediction Models Depending on Site Characteristics?," in *Proceedings of the European Wind Energy Conference & Exhibition (EWEC)*, London, UK, November.

Monteiro, C., R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, 2009, *Wind Power Forecasting: State-of-the-Art 2009*, ANL/DIS-10-1, Argonne National Laboratory, Argonne, IL.

Murphy, A.H., 1971, "A Note on the Ranked Probability Score," *Journal of Applied Meteorology* 10(1):155–156.

Negnevitsky, M., and P. Johnson, 2008, "Very Short Term Wind Power Prediction: A Data Mining Approach," pp. 1–3 in *Proceedings of the Power and Energy Society General Meeting — Conversion and Delivery of Electrical Energy in the 21st Century*, Pittsburgh, Pa.

Parkes, J., 2009, "Temporal Forecast Uncertainty of Ramp Events," in *Proceedings of the European Wind Energy Conference & Exhibition (EWEC) 2009*, Marseille, France.

Parkes, J.R., J. Wasey, and A.J. Tindal, 2006, "Wind Energy Trading Benefits through Short Term Forecasting," in *Proceedings of the European Wind Energy Conference & Exhibition (EWEC),* Athens, Greece.

Peirce, C.S., 1884, "The Numerical Measure of the Success of Predictions," *Science* 4(93):453–454.

Pinson, P., 2006, "Estimation of the Uncertinity in Wind Power Forecasting," Ph.D. Dissertation, École des Mines de Paris.

Potter, C.W., E. Grimit, and B. Nijssen, 2009, "Potential Benefits of a Dedicated Probabilistic Rapid Ramp Event Forecast Tool," pp. 1–5 in *Proceedings of the IEEE Power Systems Conference and Exposition–PSCE '09*, Seattle, Wash., March.

Provost, F., and T. Fawcett, 1997, "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions," pp. 43–48 in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)*, Newport Beach, Calif., August.

R. Pyle, R., 2010, "Wind Ramp Prediction—Improved Predictability for Wind Energy Production," Jan. [http://www.vaisala.com/en/press/vaisalanews/vaisalanews182/Pages/vn182_wind_ramp_prediction.aspx].

Sørensen, P., N.A. Cutululis, A Vigueras-Rodríguez, H. Madsen, P. Pinson, L.E. Jensen, J. Hjerrild, and M. Donovan, 2008, "Modelling of Power Fluctuations from Large Offshore Wind Farms," *Wind Energy* 11(1):29–43.

Stephenson, D.B., 2000, "Use of the 'Odds Ratio' for Diagnosing Forecast Skill," *Weather and Forecasting* 15(2):221–232.

Weigel, A.P., M.A. Liniger, and C. Appenzeller, 2007, "Generalization of the Discrete Brier and Ranked Probability Skill Scores for Weighted Multimodel Ensemble Forecasts," *Monthly Weather Review* 135(1):118–124.

WEPROG, undated, home page [http://www.weprog.com].

Yule, G.U., 1900, "On the Association of Attributes in Statistics," *Philosophical Transactions of the Royal Society of London* 194:257–319.

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", *Proceedings of the European Wind Energy Conference & Exhibition (EWEC),* Warsaw, Poland, April.

Zheng, H., and A. Kusiak, 2009, "Prediction of Wind Farm Power Ramp Rates: A Data-Mining Approach," *Journal of Solar Energy Engineering* 131(3): 031011-1–031011-8.

Decision and Information Sciences Division
Argonne National Laboratory
9700 South Cass Avenue, Bldg. 221
Argonne, IL 60439-4844

www.anl.gov