LLNL-JRNL-562015

Using Data Mining to Enable Integration of Wind Resources on the Power Grid

C. Kamath, Y. J. Fan

June 20, 2012

Statistical Analysis and Data Mining

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Using Data Mining to Enable Integration of Wind Resources on the Power Grid

Chandrika Kamath and Ya Ju Fan

Abstract—As renewable resources, such as wind, start providing an increasingly larger percentage of our energy needs, we need to improve our understanding of these resources so we can manage them better. The intermittent nature of the power generation makes it challenging for control room operators to schedule wind energy while balancing the load on the grid. Forecasts of the energy to be generated by a wind farm in the hours ahead tend to be inaccurate, even under normal conditions. The problem is exacerbated during ramp events, where the generation changes by a large amount in a small time. In this paper, we analyze historical data to determine ways in which data mining techniques can enable the integration of wind energy into the grid. Our results indicate that we can use feature selection methods to identify important weather variables associated with ramp events and inaccurate forecasts, thus reducing the number of data streams an operator must monitor. In addition, we can use ensembles of decision trees to predict days likely to have ramp events or inaccurate forecasts, thus providing grid operators additional information they can use to make well-informed decisions on scheduling wind energy.

Index Terms—wind energy, ramp events, energy forecasts, dimension reduction, classification

1 Introduction

A desire for energy independence from fossil fuels, along with various climate change initiatives, have resulted in an increasing interest in renewable sources of energy, such as wind and solar. However, to manage their successful integration into the power grid, we need to understand the intermittent nature of these resources. In the case of wind energy, we can have days when the wind does not blow and the generation is low and flat, or it may be high and flat on days when the wind speed is at a sustained high level for most of the hours in the day. Or, the generation may be high in the early hours, drop down to near zero by noon, and rise again in the late evening.

Scheduling an intermittent resource, such as wind, is usually done through forecasts. Control room operators typically use a 0- to 6-hour ahead forecast to determine the amount of energy to schedule for the hours ahead. The forecasts are updated hourly and appropriate changes made to the energy scheduled for the hours ahead. Additional fine tuning is done in real time so that the load and the generation are balanced at all times. These wind power generation forecasts are obtained from numerical weather prediction simulations which predict the wind speeds for a time horizon of up to ten days. The wind speed is then converted into wind power generation using either a statistical approach based on training with measurement data or a physical approach based on a detailed physical description of the lower atmosphere, combined with the power curve of the wind turbine [1], [2].

In a typical scenario, a control room operator uses the forecast to schedule the energy to be contributed by the wind farms within the balancing area being managed by the electric utility. If the forecast is accurate, there are no issues in scheduling. However, forecasting wind speed accurately using numerical weather prediction models can be difficult [3], especially in regions where the terrain is complex and the meteorological processes controlling the wind speed are difficult to model. When the forecast is inaccurate, the operators might look at the actual wind power generation for the previous hours or days, and, based on their prior experience, as well as the current weather conditions, appropriately schedule the wind energy for the upcoming hour. This is understandably difficult under normal operating conditions, but more so during ramp events, where the energy generated suddenly increases or decreases rapidly in response to changes in wind velocity.

In the case of positive ramps, where the wind energy increases by a large amount over a short period, the operators must either reduce other generation or sell the excess energy so that the load and generation remain balanced. This is a challenge if the positive ramp had not been forecast and it is not possible either to reduce the other generation or to sell the excess energy at short notice. Positive ramps can also cause problems if the transmission lines cannot handle the sudden increase in energy. A possible solution is to "spill" wind by letting the wind energy go to waste, but this is not an option if the wind energy is being traded in a market where it is a "must take" resource. In case of a negative ramp event, the operators must have enough backup power to keep the load balanced. Having this additional back-

- *Chandrika Kamath and Ya Ju Fan are with the Lawrence Livermore National Laboratory, Livermore, CA 94551. E-mail: kamath2,fan4@llnl.gov*

up might not be cost-effective, especially if a negative ramp is predicted but does not occur. The problem is further aggravated if the energy markets levy a high penalty when a wind farm is unable provide the energy scheduled.

These inaccurate forecasts and ramp events were not an issue in the past when the percentage of energy from wind sources, relative to the peak load, was small. For example, in 2006, the California Independent System Operator (CaISO) managed over 2200 MW of wind power generation, which was only 4% of the total generation resources in the area [4]. The Tehachapi area in Southern California, which is one of the largest wind power generation areas under CaISO and provides electricity through Southern California Edison (SCE), had 740 MW installed capacity. At this capacity, the ramp events are relatively small. So, it is relatively easy to keep the load balanced, and the generation required to back up wind power is small as well.

However, with increasing wind penetration, the size of the ramp events has also increased. For example, the installed wind capacity in the mid-Columbia Basin, a region in the Bonneville Power Administration balancing area (BPA [5]), at the end of 2008 was ~1500 MW. By the end of 2011, this had increased to ~4000 MW, making it a significant percentage of the peak load of ~8,000 MW. At this capacity, the wind ramps can be quite large, changing by nearly 1000 MW in an hour [6], and it becomes more of a challenge to balance the load and the generation. The control room operators and schedulers now have to monitor the wind power generation more closely, and plan appropriately for inaccurate forecasts as well as positive and negative ramp events.

Figure 1 shows a sample of our data to illustrate the problem. Panel (a) shows the actual generation (dotted curve, bottom) and the actual load (continuous line, top) for the first week of January 2008 for BPA. Note the daily periodicity in the load curve and the intermittency in the wind power generation. Panel (b) shows the actual generation for BPA for the month of January 2008. Note that some days have no generation, while on other days, it can go over 1000 MW. Panel (c) shows the actual (continuous line) vs. forecast (dotted line) wind power generation for the first week of January 2008 for the Antelope wind farm in the Tehachapi Pass region. Note that the forecast may track the shape of the actual generation, but at a different magnitude, and can sometimes be quite different from the actual generation. There is a negative ramp that occurs in the morning of January 6, which is much sharper than the drop forecast. Panel (d) shows the actual generation for SCE for the month of January 2008. This is the sum of the generation of two wind farms in the region: Antelope and Vincent. Note that the maximum generation is lower than in BPA.

There are several ways in which we can better manage the challenges resulting from the integration of intermittent resources, such as wind, on the power grid. We can improve the accuracy of forecasts either through the assimilation of higher quality data obtained by the judicious placement of appropriate sensors [7] or through improved models [8]. We can also analyze historical data to understand wind power generation better, and in the process, use the insights obtained to enable control room operators to make well informed scheduling decisions especially on days when there are ramp events or the forecasts are inaccurate.

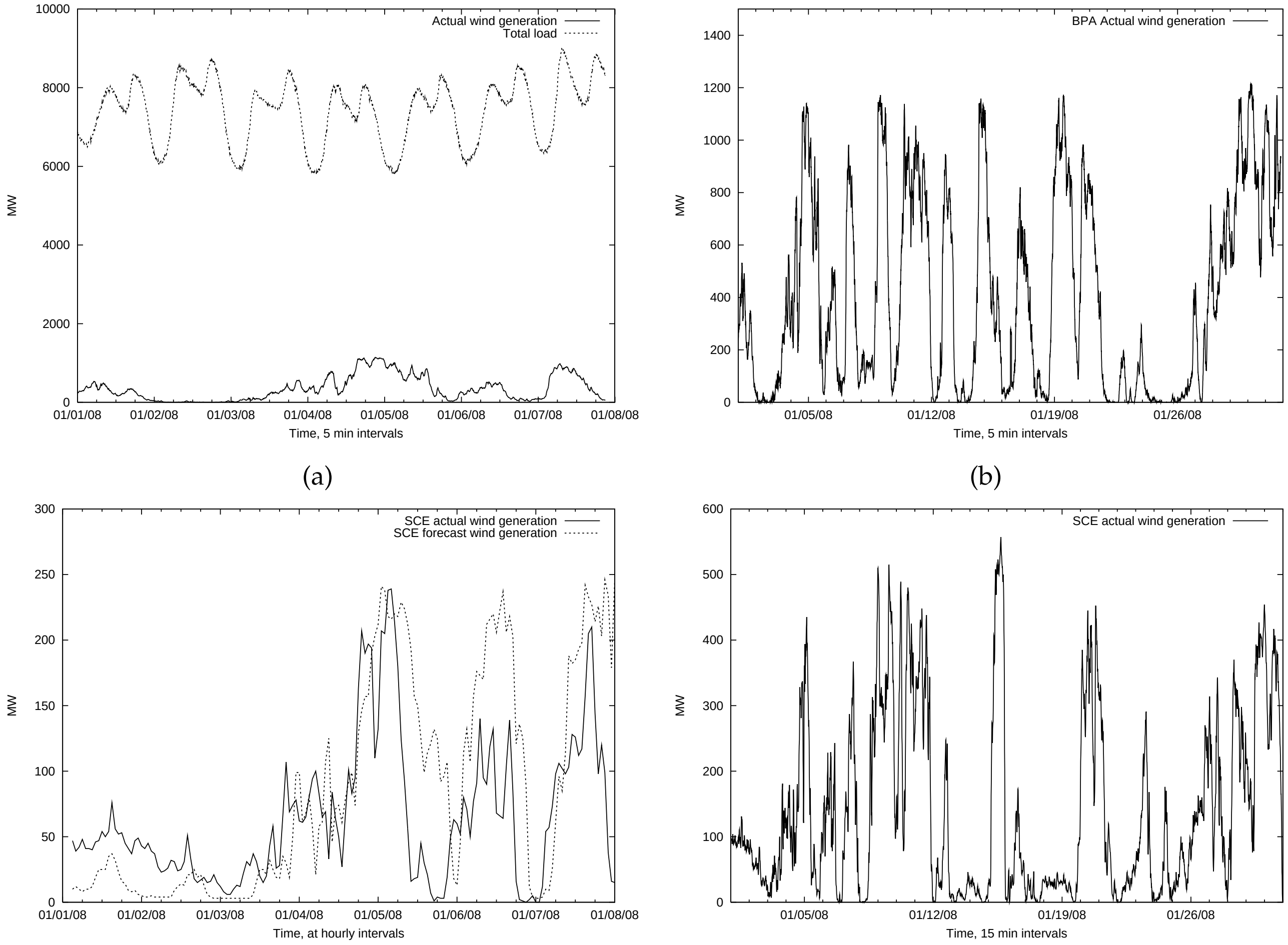
In this paper, we investigate the use of dimension reduction and classification techniques to analyze historical data from the wind farms in the Tehachapi Pass and mid-Columbia Basin regions to determine if we can:

- identify the weather conditions associated with days with ramp events and build models to predict if a day will have ramp events.
- identify the weather conditions associated with days with inaccurate forecasts and build models to predict if a day will have an inaccurate forecast.

If successful, this will reduce the number of weather variables the control room operators must monitor and provide them some advance notice of days likely to have ramp events or inaccurate forecasts.

This paper builds on prior work in several ways. In [6], [9], we first studied ramp events in the Tehachapi Pass and mid-Columbia Basin regions to evaluate different definitions of these events. We also extracted simple statistics to address questions such as, do ramps occur more frequently in the mornings or evenings, do the negative ramps occur as frequently as the positive ramps and are they as severe, and do severe ramps occur rarely or are they relatively frequent? Next, in follow-on work [10], we used feature selection techniques to identify weather conditions that were associated with days with ramp events. A control room operator could then monitor only those weather variables related to ramps and not be overwhelmed by irrelevant weather information. This paper extends our previous work as follows. First, we complement the use of feature selection methods for identifying important weather variables associated with ramp events by investigating linear and non-linear transform-based, dimension reduction techniques. Second, we use classification techniques to determine if the weather variables can be used to build accurate predictive models for days with ramp events. Finally, we extend our work to include not just ramp events, but also days with inaccurate forecasts.

This paper is organized as follows. First, in Section 2, we present the wind generation and weather data available for our analysis. Next, in Section 3, we describe how we can use the above data to address the two questions posed earlier; (i) identifying important variables associated with ramp events and inaccurate forecasts and (ii) building predictive models. We describe our analysis approach in Section 4 followed by experimental results and discussion in Section 5. Related work is described in Section 6 and we conclude with some thoughts on future work.



(a)

(b)

(c)

(d)

Fig. 1. (a) The load (top curve) and the wind power generation (bottom curve) for the first week of January 2008 for the BPA balancing area. Data are available every 5 minutes. (b) The wind power generation for January 2008 for the BPA balancing area. Data are available every 5 minutes. (c) Forecast (solid line) vs. actual (dotted line) generation for the Antelope wind farm in SCE for the first week of January 2008. Data are hourly. (d) The wind power generation for January 2008 for the Tehachapi Pass area. Data are available every 15 minutes.

2 Description of the Data

We apply our analysis techniques to data from the Tehachapi Pass in Southern California (referred to as the SCE data) and the mid-Columbia Basin region (referred to as the BPA data). In light of the increasing wind power generation in the last few years, we focus our analysis on data from the recent past as these are likely to be more relevant. For the SCE region, we use data from 2007-2008 and for the BPA region, we use data from 2007-2009. Our datasets will use the prefix BPA or SCE to indicate the region, followed by a 'G' or 'W' to indicate generation or weather, and then a number or a letter reflecting the time between samples. Note that the data available are frequently averages over data collected at a higher frequency, for example, data at 5 minute intervals are actually averages of data collected at intervals of 1 minute. As is often the case with observed data from sensors, these data are of varying quality, with noise and many missing values. They are also sampled at different frequencies, which may be an issue if the sampling frequency is not sufficiently high to capture the event of interest, for example, ramp events that occur over 30 minutes. One of our challenges is to appropriately pair the wind power generation data with a similarly sampled weather dataset, and address the missing values suitably, to achieve our analysis goals and improve our understanding of issues related to integrating wind energy on the power grid.

2.1 Wind power generation data

We have three datasets available for the wind power generation in mid-Columbia Basin and the Tehachapi Pass:

- **BPA 5 minute data (BPA-G-5):** The BPA data available for the period 2007-2009 are the total generation from all the wind farms in the BPA balancing area [11], sampled at 5 minute intervals. There are missing values in the data - if values were missing for one or two consecutive intervals, they were filled-in using interpolation, while longer periods were replaced by ”-9999” to indicate such values for future processing. In addition, to reduce the noise in the wind energy data, we smoothed the original data by two applications of a mean filter of size 3. We refer to this dataset as BPA-G-5.

- **SCE 15 minute data (SCE-G-15):** The Tehachapi Pass wind generation data are sampled more coarsely than the Columbia Basin data. These data are available at 15 minute intervals for the wind farms in the Vincent and Antelope regions of the Tehachapi Pass. As these regions are close by, their wind power generation is very similar, and we consider the sum of the generation in our analysis. The generation from the Antelope region occasionally had small negative values which were replaced by zero before being added to the data from the corresponding interval from the Vincent region. Unlike the data from Columbia basin, no smoothing was used, as it would have adversely affected the calculation of 30 and 60 minute ramps [6]. We refer to this dataset as SCE-G-15.
- **SCE 60 minute data (SCE-G-60):** For the SCE region, we also have hourly data on both the forecast and actual generation for 2007-2008. These data were from a different source than the 15 minute data (see Section 8 for source information). There were several periods of 3 hours each, where there were missing values for both the forecast and actual generation. These were replaced by interpolated values. There were also time periods where the actual generation was reported as 0.0, though the forecast was ~200 MW. A comparison with the 15 minute data, SCE-G-15, indicated that the values were in error. We also found that the actual generation in this dataset was very close to the hourly generation derived from SCE-G-15. So, for consistency of analysis, we replaced all values of the actual generation by the hourly generation derived from the 15 minute data. These data are analyzed without any smoothing. We refer to this dataset as SCE-G-60.

2.2 Weather data

There are several weather data sources available for use in our analysis. These data are available at different temporal resolutions from several meteorological towers in and around the two regions of interest.

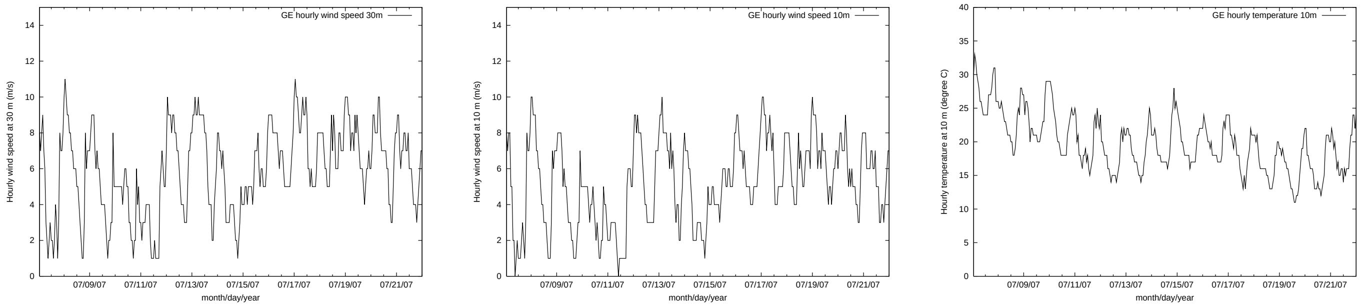
- **SCE hourly data (SCE-W-H):** The first weather dataset, which is from the same source as the SCE-G-60 data, is from seven towers in the Tehachapi Pass. These data are hourly data and different variables are available from each station as summarized in Table 1. We refer to these data as SCE-W-H. They are of relatively poor quality, with some repetitions (contiguous values representing the variables for the same hour), missing time periods (where rows representing time periods were missing from the data file), and missing variables (where the missing values were replaced by a suitable number, such as -99 or -9.9). To clean the data, we removed the duplicate rows and added rows for missing time periods, using the appropriate “missing” value for the variables. After this, we found that nearly 46% of the rows in the data had one or more variables with a “missing” value. Figure 2 shows some of the hourly weather variables in the Tehachapi Pass region.
- **SCE daily average data (SCE-W-D):** In addition to the hourly weather data for Tehachapi Pass, we also have daily averages obtained from the Remote Automated Weather Station (RAWS). These data are available from the Western Regional Climate Center (http://wrcc.dri.edu). There are several variables available from each station, as listed in Table 2. Note that the variables at these stations are different from the variables in the SCE-W-H data.
  Figure 3(a) shows the different weather stations in the Tehachapi Pass region. We started by considering the ones near the wind farms and selected those with no missing values for the variables. The three sites that met our criteria, along with their latitude/longitude, are Jawbone (35.294722,-118.226389); Bearvalley (35.139722, -118.625); and Piutes (35.431667, -118.329722), with Tehachapi Pass located at (35.102222, -118.282778). We also found that some variables (barometric pressure and the average, maximum, and minimum soil temperature) were missing for all days at all stations and therefore, removed from consideration. We refer to this dataset as SCE-W-D. Figure 4 shows some of the daily averaged weather variables in the Tehachapi Pass region.
- **BPA daily average data (BPA-W-D):** Using a similar approach, we identified four RAWS locations in the mid-Columbia Basin, at Locks (45.669444, 121.881667); Patjens (45.322222, 120.925); Umatilla NWR (45.916667, 119.566667); Wasco (45.61,121.33) (see Figure 3(b)). The variables at these stations are listed in Table 2. As in the case of SCE-W-D data, we do not consider barometric pressure and the average, maximum, and minimum soil temperature, as these are missing for all days. We refer to this dataset as BPA-W-D.

Selecting the weather datasets for use in our work was a challenge as we were constrained to using publicly available data from existing weather stations. As a result, the locations of the stations, and the variables they measure, may not be the optimal ones for use in our analysis. On one hand, the expansion of wind farms without a corresponding increase in the number of weather stations has resulted in an acknowledged lack of weather data suitable for accurate prediction of wind speeds. At the same time, given that the wind farms are often spread out over a large area (see, for example, Figure 3(b)), it is unclear what should be the appropriate locations for

Name of station (initials)	Latitude/Longitude	Variables
Sky River (SR)	35.260980 N 118.246958 W	wspeed30, wdir30
GE Wind (GE)	35.081558 N 118.37193 W	wspeed30, wdir30, wspeed10, wdir10, temp30, temp10, pressure
STP (ST)	35.134747 N 118.458735 W	wspeed30, wdir30, wspeed10, wdir10, temp30, temp10, pressure
Morewind (MW)	35.052237 N 118.285772 W	wspeed30, wdir30
Oak Creek M11-13 (OC1)	35.041738 N 118.370612 W	wspeed30, wdir30
Oak Creek M54-25 (OC2)	35.031377 N 118.346922 W	wspeed30, wdir30
The Rock (TR)	35.12956 N 118.349897 W	wspeed30, wdir30

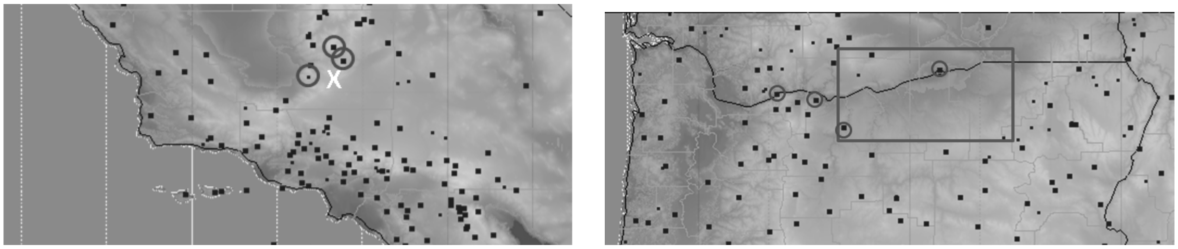
TABLE 1

The hourly weather data available at the seven meteorological stations in the Tehachapi area. The variables are measured at heights of either 10m or 30m (as indicated by the two numerals at the end of the variable name), and include wind speed, wind direction, temperature and pressure. Note that not all variables are available at all stations.



(a) (b) (c)

Fig. 2. Some examples of the hourly weather data available for the SCE region. (a) The hourly wind speed at 30m for GE from SCE-W-H. (b) The hourly wind speed at 10m for GE from SCE-W-H. (c) The hourly temperature at 30m for GE from SCE-W-H. These hourly data are for a period of 15 days in July 2007.



(a) (b)

Fig. 3. The locations of the WRCC weather stations (indicated by the small black squares). (a) The Southern California region, where the white cross indicates the Tehachapi Pass area. The three circles indicate the specific sites chosen in our analysis: Bearvalley, Jawbone, and Piutes. (b) The Oregon-Washington border region, where the square box indicates the region of the wind farms in the BPA BA. The four circles indicate the specific sites chosen in our analysis: Locks, Patjens, Umatilla, and Wasco.

weather stations and what variables should be measured at each, so these variables will improve our predictions. We justify our choice of weather datasets as these are the same weather stations that are available to the control room operators. Our goal is to understand to what extent we can exploit the existing weather information available to us, while recognizing that the sensor data from the weather stations may not be optimal. We also observe that we use the actual weather data in our analysis, though in practice, any predictions would be done using forecast values for the weather variables.

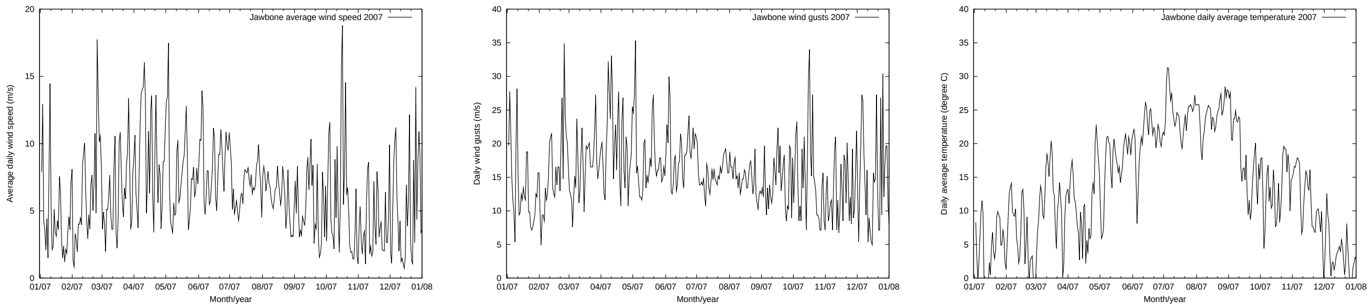
2.3 Processing the weather data

For the daily and hourly weather data used in our analysis, we observed that certain variables were likely to be correlated, for example, the air temperature and the fuel temperature, or the two definitions of growing degree days, or the air temperature average, maximum, and minimum. In addition, the variables could be correlated across the meteorological sites as well, for example, two

Solar Rad. total kW-hr/m2	Speed average m/s	Wind dir vector deg
Speed Gust m/s	ASCE Et. total mm	Penman Et. total mm
Air temp Average deg C	Air temp Maximum deg C	Air temp Minimum deg C
Fuel Temp Average deg C	Fuel Temp Maximum deg C	Fuel Temp Minimum deg C
Soil temp Average deg C	Soil temp Maximum deg C	Soil temp Minimum deg C
Relative humidity average percent	Relative humidity maximum percent	Relative humidity minimum percent
Heating Degree Days	Cooling Degree Days	Growing Degree Days Base 40
Growing Degree Days Base 50	Precipitation Total mm	Barometric Pressure Average mbar

TABLE 2

The daily averages for the above weather variables (along with their units) are available from the Remote Automated Weather Stations (RAWS) at the Western Regional Climate Center.



(a) (b) (c)

Fig. 4. Some examples of the daily weather data available for the SCE region. (a) The daily wind speed for 2007 for Jawbone from SCE-W-D. (b) The daily wind gusts for 2007 for Jawbone from SCE-W-D. (c) The daily average temperature for 2007 for Jawbone from SCE-W-D.

nearby sites might have correlated air temperatures. We identified such variables using the Pearson correlation coefficient and removed them from consideration. Given two vectors, $\mathbf{x}$ and $\mathbf{y}$, each of length n, the Pearson correlation coefficient between them is given by

$$\frac{1}{n}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sigma(x)\ \sigma(y)}$$

where x_i is the i-th element of the vector $\mathbf{x}$, $\bar{x}$ is its mean value, and $\sigma(x)$ is its standard deviation. In our analysis, we considered two variables to be correlated if they had a coefficient greater than 0.75.

For the daily average weather variables, after the removal of correlated variables at each site, we were left with the following seven variables in the RAWS datasets: total solar radiation, average wind speed, wind direction, wind speed gusts, average air temperature, average relative humidity, and the precipitation. Concatenating them for the 3 and 4 stations in the SCE and BPA areas gave us a total of 21 and 28 daily weather variables, respectively. We could have pre-processed the data further to remove potential outliers and variables correlated across sites. This was not done as we wanted to ensure that small scale weather phenomena, which might affect one site, but not another nearby site, would be included in the analysis. Also, some variable values which appeared as outliers, were not really outliers, such as the few days when precipitation at a site was high.

The hourly weather data (SCE-W-h) was also processed in a similar manner to identify correlated variables. However, since there are fewer variables for each station and the stations are close together, we considered all 24 variables from Table 1 in identifying correlations. As a result of this analysis, we were left with the following seven uncorrelated variables: SR-wspeed30, SR-wdir30, ST-wdir30, ST-wdir10, GE-temp30, st-pressure, and GE-pressure.

3 Formulating the problems

In our analysis, we are interested in understanding if weather conditions can be associated with, and used to predict, ramp events and inaccurate wind power generation forecasts. We next show how we can use the data described in Section 2 to meet these goals.

3.1 Calculating ramp events

For both the BPA and SCE regions, we define a ramp event [6], of magnitude Tr in MW, to occur between time intervals T and $(T+\Delta T)$ if

$$max(MW[T,T+\Delta T]) - min(MW[T,T+\Delta T]) > Tr.$$

The ramp is considered to be a negative ramp, if the maximum occurs before the minimum, and is positive otherwise. Then, we identify days when ramp events have occurred using the following values for the parameters:

- **The time interval ΔT:** We considered two cases - 30 minutes and 60 minutes as these are durations typically considered for ramps.
- **The threshold Tr:** The choice of the threshold was non-obvious. We started by selecting an absolute threshold of 120 MW and 240 MW for the 30 minute and 60 minute ramps, respectively, for the BPA data. An absolute threshold made sense as it gives operators a sense of how much back-up generation they need or how much they should reduce other generation. However, in the case of BPA, as the installed wind capacity increased substantially during the analysis period, use of a fixed threshold resulted in many more ramps being identified during the latter part of the period. So, a day with certain weather conditions early in the analysis period may have no ramps, while a day with similar weather conditions later in the analysis period could have many ramps. To avoid this unintended consequence of the increase in installed capacity on our analysis, we used a percentage of the installed capacity on any day as the threshold.

  We considered thresholds of 10% and 12% of capacity for 30 minute ramps and 15% and 20% of capacity for 60 minute ramps for both regions. For the Tehachapi Pass, where the installed capacity was constant at 740 MW over the analysis period, this results in the use of 75 MW and 90 MW thresholds for 30 minute ramps and 115 MW and 150 MW thresholds for 60 minute ramps. For the Columbia basin region, the installed capacity, which is available only from October 2007 onwards, ranges from a low of 922 MW to a high of 2617 MW at the end of the analysis period. These percentages of installed capacity were chosen so that the results at low capacity were not only close to our choice of absolute thresholds, but also led to a moderate number of days with ramps so that we had roughly equal number of days with and without ramps. A threshold set too low (or too high) would have led to too many (or too few days) with ramp events.

Having identified the ramp events in the data, we next had to assign a label to each day. We considered a day to have a ramp event if any one of the intervals during the day was part of a ramp event, regardless of its sign. This option resulted in a two-class problem, where a day either had a ramp or not. We could have also considered this as a four class problem, where a day was assigned a label based on whether it had no ramps, only positive ramps, only negative ramps, or both positive and negative ramps. This, and other options with multiple classes based on the severity or the number of ramps in a day, were not considered further as they led to too few examples of each of the classes.

3.2 Calculating forecast accuracy

We can use the hourly actual and forecast generation for SCE to calculate the error in the forecast for each hour. Then, an hour is assigned the label 1, indicating accurate forecast, if the error is less than an hourly threshold, and zero otherwise. We can also calculate the daily error by taking the sum of the absolute error over the hourly intervals for each day. Each day can then be assigned a label of 1, indicating an accurate forecast, if the error is less than a daily threshold, and zero otherwise.

In both cases, we need to determine an appropriate threshold. Since the maximum generation at the SCE site (the sum of the generation from the Antelope and Vincent wind farms) is ~500 MW, we select a threshold that is 10% or 15% of this maximum. In other words, an hourly error of 50 or 75 MW is considered as tolerable in the context of balancing the load with generation; a higher error would be considered an inaccurate forecast. Converting this to a daily threshold results in a daily error of 1200 or 1800 MW.

3.3 Pairing the wind power generation and weather data

We address the questions raised in Section 1 by appropriately pairing the the wind power generation and the weather data described in Sections 2.1 and 2.2, respectively. We do this by considering the suitability of the data available, the sampling rate, and the questions we are trying to address.

But first, a few observations on the quality of the data are helpful in understanding the pairing of the datasets. We found that the wind generation data are of higher quality than the weather data. They are at a higher resolution, being sampled at 5 minute, 15 minute, or 60 minute intervals. There are few missing values, and when the values are missing, it is only one or two consecutive values which can be easily filled in by interpolation. In contrast, the weather data, regardless of whether they are from RAWS or other meteorological stations, have a large number of missing values over a long duration, making it infeasible to use interpolation to fill in these values. These data are also at a coarser resolution, which may not necessarily be a drawback. If the data are sampled at a finer resolution (say, at 5 minute intervals), then, for variables such as wind speed, we may need to incorporate a lag into the analysis to account for the wind farm being far away from the station. This lag could change over time, making the alignment difficult.

We also observe that the hourly forecast and actual wind power generation data, along with the hourly weather data, are available only for SCE. The variables in the hourly weather data (SCE-W-H) are also different

Problem Description	Sites	Number of instances	Number of features	Parameters	Class Distribution
1. Associating daily weather conditions with days ramp events: pairs SCE-G-15 with SCE-W-D and BPA-G-5 with BPA-W-D.	SCE	731	21	(a) 30 min, 10%	58%
				(b) 30 min, 12%	72%
				(c) 60 min, 15%	58%
				(d) 60 min, 20%	77%
	BPA	819	28	(a) 30 min, 10%	50%
				(b) 30 min, 12%	66%
				(c) 60 min, 15%	46%
				(d) 60 min, 20%	70%
2. Associating daily weather conditions with days with inaccurate forecasts: pairs SCE-G-60 with SCE-W-D.	SCE	731	21	(a) 1200 MW	38%
				(b) 1800 MW	69%
3. Associating hourly weather conditions with hours with inaccurate forecasts: pairs SCE-G-60 with SCE-W-H	SCE all	9330	7	(a) 50MW	54%
				(b) 75 MW	69%

TABLE 3

Summary of the problems considered: for each problem, the table presents the number of instances, the number of features, the parameters used in generating the class labels, and the class distribution. All problems are two-class problems, and the class distribution indicates the percentage of the more desirable class. This is the percentage of days without ramp events or the percentage of time intervals with accurate forecasts.

from the ones in the daily RAWS data (SCE-W-D and BPA-W-D) .

Based on the above, we considered the following pairing of the data, resulting in three problems:

- **Associating daily weather averages with days with ramp events:** First, as described in Section 3.1, we extract 30 and 60 minutes ramp events from the wind power generation data which is available at 5 minute and 15 minute intervals for BPA and SCE, respectively. Then, we associate with each day a class label indicating if it does or does not have ramp events of a certain magnitude and duration. This information is paired with the daily weather data and used to associate weather variables with days with ramp events. By considering ramp events at the resolution of a day, we can exploit the weather information in the RAWS data that are available as daily averages. This problem pairs SCE-G-15 with SCE-W-D and BPA-G-5 with BPA-W-D.
- **Associating daily weather averages with days with inaccurate forecasts:** We can use a similar argument to take days with inaccurate forecasts for the SCE region (for which we have both the hourly forecast and the actual generation) and identify the daily average weather conditions that are associated with such days. This would pair SCE-G-60 and SCE-W-D.
- **Associating hourly weather conditions with hours with inaccurate forecasts:** We have thus far focused on using the daily weather data as they are of better quality and measure a varied set of variables. However, it is natural to ask if we can exploit the hourly weather data available for SCE (SCE-W-H), despite the large number of missing values. Considering only those hours which have no missing weather variables, we created a dataset of 9330 instances out of a total of 17547. This problem pairs SCE-G-60 with SCE-W-H.

Table 3 summarizes the datasets for the problems considered, including the number of instances, the numher of features, the parameters considered, and the distribution of classes. All problems are two-class problems, with the instances labeled 0 and 1, where 1 refers to the more favorable or desirable case, that is, instances without ramp events or with accurate forecasts.

4 Analysis Approach

In our work, we are interested in using data mining techniques to determine if we can provide control room operators information they can use to make well informed decisions on days with ramp events or inaccurate forecasts. We identified two ways in which we can help. First, we can identify important weather variables associated with ramp events and inaccurate forecasts, thus reducing the number of data streams that the operators have to monitor. This obviously points to the use of dimension reduction techniques, specifically, feature selection methods. Second, we can predict days likely to have ramp events or inaccurate forecasts, so the operators can be better prepared. This points to the use of classification methods. We next briefly describe the dimension reduction and classification methods used in our work.

4.1 Dimension reduction

One of our goals is to determine which of the many weather variables at the different sites in a region are associated with wind related events in that region. If we can determine a small set of such variables, then the control room operators need only monitor this small set, reducing their data overload. Feature selection techniques are a natural solution to this problem. In our work, we focus on "filter" methods [12]. They select features based on properties we would expect of good feature subsets, such as class separability or high correlation with the target. They are also computationally less expensive than the "wrapper" methods which evaluate the subset selected using the classifier; however, this may lead to the filter methods producing less accurate results when the subset of features is used in classification. We consider the following four filter methods in our analysis:

- **Distance filter:** The distance filter calculates the class separability of each feature using the Kullback-Leibler (KL) distance between histograms of feature values. For each feature, there is one histogram for each class. In our two class problem, if a feature has a large distance between the histograms for the two classes, then the feature is likely to be an important feature. If, on the other hand, the histograms overlap, then the feature is unlikely to be helpful in differentiating between days with and without ramp events. We discretized numeric features using $\sqrt{|D|/2}$ equally-spaced bins, where $|D|$ is the size of the data. The histograms are normalized by dividing each bin count by the total number of elements to estimate the probability that the j-th feature takes a value in the i-th bin of the histogram given a class n, $p_j(d = i|c = n)$. For each feature j, we calculate the class separability as

$$\Delta_j = \sum_{m=1}^{c} \sum_{n=1}^{c} \delta_j(m, n),$$

where c is the number of classes (= 2 for our problem) and $\delta_j(m, n)$ is the KL distance between histograms corresponding to classes m and n:

$$\delta_j(m, n) = \sum_{i=1}^{b} p_j(d = i|c = m) \log\left(\frac{p_j(d = i|c = m)}{p_j(d = i|c = n)}\right),$$

where b is the number of bins in the histograms.

The features are ranked simply by sorting them in descending order of the distances Δ_j (larger distances mean better separability).

- **Chi-squared filter:** This filter computes the Chi-square statistics from contingency tables for every feature. The contingency tables have one row for every class label and the columns correspond to possible values of the feature as shown in Table 4.1, adapted from [13]. Numeric features are represented by histograms, so the columns of the contingency table are the histogram bins.

Class	f1=1	f1=2	f1=3	Total
0	31 (22.5)	20 (21)	11 (18.5)	62
1	14 (22.5)	22 (21)	26 (18.5)	62
Total	45	42	37	124

TABLE 4

A 2×3 contingency table, with observed and expected frequencies (in parenthesis) of a fictitious feature f1 that takes on 3 possible values (=1, 2, and 3).

The Chi-square statistic for feature j is

$$\chi_j^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

where the sum is over all the cells in the $r \times c$ contingency table, where r is the number of rows and c is the number of columns; o_i stands for the observed value (the count of the items corresponding to the cell i in the contingency table); and e_i is the expected frequency of items calculated as:

$$e_i = \frac{(\text{column total}) \times (\text{row total})}{\text{grand total}}$$

The variables are ranked by sorting them in descending order of their χ^2 statistics.

- **Stump filter:** This filter is derived from the process of building a decision-tree classifier. Decision trees split the data by examining each feature and finding the split that optimizes an impurity measure. To search for the optimal split of a numeric feature x, the feature values are sorted ($x_1 < x_2 < ... < x_n$) and all intermediate values $(x_i + x_{i+1})/2$ are evaluated as possible splits using a given impurity measure. The features are then ranked according to their optimal impurity measures.

There are several options we can use for the impurity measure. In our work, we use the Gini index [14] which is based on finding the split that most reduces the node impurity, where the impurity for a c class problem is defined as follows:

$$L_{Gini} = 1.0 - \sum_{i=1}^{c} (L_i/|T_L|)^2$$

$$R_{Gini} = 1.0 - \sum_{i=1}^{c} (R_i/|T_R|)^2$$

$$\text{Impurity} = (|T_L| * L_{Gini} + |T_R| * R_{Gini})/n$$

where $|T_L|$ and $|T_R|$ are the number of examples, L_i and R_i are the number of instances of class i, and L_{Gini} and R_{Gini} are the Gini indices on the left and right side of the split, respectively.

A stump is a decision tree with only the root node; the stump filter ranks features using the same process as the one used to create the root node.

- **ReliefF:** The original Relief algorithm [15], was proposed for a 2-class problem, though it has since been extended to multi-class problems and regression [16]. It estimates the importance of features based on how well they distinguish between instances near to each other. It essentially maintains a quality estimate for each feature. Given a randomly selected instance, it finds the nearest instance from the same class (called a hit) and of a different class (called a miss). If a feature has different values for the random instance and the hit, it is not a desirable feature and its quality is reduced. Similarly, a feature which has different values for the random instance and the miss, can be considered a useful feature and its quality estimate is increased. The process is repeated for different random instances; in our work, we use all instances. We can also use the k nearest hits and misses (the ReliefF algorithm), instead of a single hit and miss. We experimented with $k = 3, 10, \mathrm{and} 20$; the performance tends to be similar, so we include the results for just $k = 3$.

In our earlier work [10], we focused exclusively on feature selection methods as these give results which are interpretable. We did not consider transform-based approaches, such as principal component analysis (PCA), to reduce the dimension as the transformation makes it difficult to identify which of the original variables are the important ones. However, given the recent research in non-linear dimension reduction techniques, it is obvious to ask if the reduced dimensional representation obtained by such methods could give better prediction accuracy relative to the feature selection methods. With this intent, we also considered the following techniques:

- **Isomap:** This method [17] preserves pairwise geodesic distances between data points. It starts by constructing an adjacency graph that determines which points are neighbors in the input space. These neighbors can be either the k-nearest neighbors or points which lie within an ϵ-neighborhood. Next, the geodesic distances between all pairs of points are estimated by computing their shortest path distances over the graph. Let $D_G = \{d_G(i,j)\}_{i,j=1,\ldots,n}$ be the matrix of geodesic distances, where $d_G(i,j)$ is the distance between point i and j. Isomap then constructs an embedding in a d-dimensional Euclidean space such that the pair-wise Euclidean distances between points in this space approximate the geodesic distances evaluated in the input space. Let $D_Y = \{d_Y(i,j)\}_{i,j=1,\ldots,n}$ be the Euclidean distance matrix and $d_Y(i,j) = \|Y_i - Y_j\|_2$. The goal is to minimize the cost function $\|\tau(D_G) - \tau(D_Y)\|_2$, where the function τ performs double centering on the matrix to support efficient optimization. The optimal solution can be found by solving the eigendecomposition of $\tau(D_G)$, where the matrix D_G is a dense matrix. The Y coordinates are then computed based on the d largest eigenvalues and their corresponding eigenvectors.
- **Locally Linear Embedding (LLE):** This method [18] preserves the reconstruction weights ω_{ij} that are used to describe a data point X_i as a linear combination of its neighbors $X_j, j \in \mathcal{N}(i)$, where $\mathcal{N}(i)$ denotes the set of points that are neighbors of point i. The optimal weights for each i are obtained by minimizing the cost function,

$$\min_{\omega} \quad \{\|X_i - \sum_{j \in \mathcal{N}(i)} \omega_{ij} X_j\|^2 \Big| \sum_{j \in \mathcal{N}(i)} \omega_{ij} = 1\}.$$

  LLE assumes that the manifold is locally linear and hence the reconstruction weights are invariant in the low-dimensional space. The embedding Y of LLE is obtained from the eigenvectors corresponding to the smallest d nonzero eigenvalues of the embedding matrix, defined as $M = (I - W)^{\mathsf{T}}(I - W)$, where W is the reconstruction weight matrix whose element $W_{ij} = 0$ if $j \notin \mathcal{N}(i)$; $W_{ij} = \omega_{ij}$ otherwise. I is an identity matrix and $(I - W)$ is a sparse, symmetric matrix. Similar to Isomap, LLE constructs the adjacency graph of data points, but it is less sensitive to outliers as it preserves solely local properties. The sum to one constraint on the weights forces the reconstruction of each data point to lie in the subspace spanned by its nearest neighbors [19].
- **Laplacian Eigenmaps:** Laplacian eigenmaps provide a low-dimensional representation of the data in which the weighted distances between a data point and other points within an ϵ-neighborhood (or its k-nearest neighbors) are minimized [20]. The distances to the neighbors are weighted according to the Laplacian operator $W_{ij} = e^{-\frac{\|X_i - X_j\|^2}{t}}$, or the sparse adjacency matrix W whose element W_{ij} is 1 if X_i and X_j are neighbors and 0 otherwise. Here, $t = 2\sigma^2$, where σ is the standard deviation of the Gaussian kernel. The representation of Y is computed by solving the generalized eigenvector problem: $(S - W)v = \lambda S v$, where $S_{ii} = \sum_j W_{ij}$. Only the eigenvectors (v) corresponding to the nonzero eigenvalues (λ) are used for the embedding.

To complement the above non-linear dimension reduction techniques, we also included PCA, a linear transform-based method.

4.2 Classification

The use of feature selection techniques to identify the important features enables the control room operators to focus on a few weather variables instead of being overwhelmed by multiple variables from different stations in the vicinity of the wind farms. While these insights are useful, we may well ask if it is possible to use the weather variables to predict days likely to have ramp events or intervals where the forecast is likely to be inaccurate. To build these predictive models, we consider ensembles of decision trees. Our choice of decision trees stems from the fact that they are interpretable and therefore, easier for application scientists to understand and incorporate into their work.

In our work, we used three ensemble based methods: adaboost [21], bagging [22], and ASPEN, a method based on approximate splits [23]. This approach creates ensembles by introducing randomization at each node of the tree in two ways. It first randomly samples the examples at a node and selects a fraction (we use 0.7) for further consideration. Then, for each feature, instead of sorting these examples based on the values of the feature, it creates a histogram, evaluates the splitting criterion (we use Gini [14]) at the mid-point of each bin of the histogram, identifies the best bin, and then selects the split point randomly in this bin. The randomization is introduced both in the sampling and in the choice of the split point. The use of the histograms and the smaller samples speeds up the creation of each tree in the ensemble.

5 Experimental Results and Discussion

We next present the results of the application of the various dimension reduction methods and ensemble-based classifiers to the problems described in Table 3. The ensemble-based methods all use 11 trees in the ensemble and the results are evaluated using the percentage error rate obtained for five-fold cross validation repeated five times. The dimension reduction results include the top variables identified by the different feature selection methods as important as well as the variation in the percentage error rate for the ASPEN method as we use the first d features in constructing the classifier. We selected the ASPEN method as it gave better accuracy than Adaboost or Bagging.

To evaluate the dimension reduction methods, we also included an additional feature which is a noise variable. In feature selection methods, when we consider the list of features ordered by importance, any variables ranked lower than the noise variable will be unimportant. In transform-based methods, the inclusion of the noise variable gives us an estimate of how robust the methods are to irrelevant features.

For the non-linear dimension reduction methods, we experimented with several different parameters and chose the ones which gave good results overall. The three methods - Isomap, LLE, and Laplacian eigenmaps - all work better when we consider the points in an ϵ-neighborhood instead of the k-nearest neighbors. However, this required experimenting with the values of ϵ as small values could result in no neighbors for many of the points.

We next present the results for the three different problems outlined in Table 3 and discuss these results for each problem, followed by some overall observations in Section 5.4.

To place the experimental results in context, we recall our goals in this analysis: first, to determine if we can identify important weather variables associated with wind ramp events and inaccurate forecasts, and second, to investigate if we can use these variables to predict the ramp events and inaccurate forecasts. Our intent is to provide control room operators additional information they can use to make better decisions in scheduling wind power generation.

5.1 Results: Problem 1

In this section, we present the results for Problem 1 described in Table 3, where we consider daily weather conditions in the BPA and SCE regions in the context of ramp events. Each region has four datasets corresponding to two time intervals and two thresholds used in the calculation of the ramp events. The results for SCE are presented in Figure 5 and Table 5 for the 30 minute ramps and Figure 6 and Table 6 for the 60 minute ramps. The corresponding information for BPA is summarized in Figure 7 and Table 7 for the 30 minute ramps and Figure 8 and Table 8 for the 60 minute ramps.

For the SCE data, when we compare the three different ensemble methods - Adaboost, Bagging, and ASPEN - we find the percentage error rate for all four datasets to be the lowest for the ASPEN approach. While the error rate may appear to be high, for example, 24.33% for 30 min ramps with 10% threshold or 26.98% for 60 min ramps with 15% threshold, we need to evaluate this in the context of the problem being addressed. Recall that the weather data are relatively low quality, so their predictive ability for ramp events is not expected to be very high. However, we are interested in determining if we can do better than a random guess (50% accuracy for our two class problem) or the accuracy obtained by assigning the majority class based on the class distribution in Table 3. Our results in Figures 5 and 6 indicate that we certainly meet the former criterion. If we consider the latter criterion, we find that for the lower values of threshold, we obtain an error rate of 24-28%, which is better than an error rate of 42% if we were to assign all cases to the majority class. For the larger thresholds, the reduction in error rate from 23-28% to 18-22% is not as large, but is non-negligible. This indicates that there is some benefit to be gained by using the weather conditions to predict the ramp events.

We next consider the feature selection results for SCE presented in Tables 5 and 6. For both the 30 min and 60 min ramp events, we observe that three of the feature selection methods - distance, chi-squared, and stump filters - all tend to select common variables as the top seven important features. These features include the average relative humidity and average air temperature at the Bearvalley and Jawbone stations and the wind speed at the Jawbone and Piutes stations. We also found that the wind speed and wind gusts at Bearvalley ranked among the least important variables. A closer inspection indicated that many of these values were erroneous. For example, of the total of 731 days in the study, 112 days had speed gusts in Bearvalley of 44.70 m/s, indicating an inoperable or faulty sensor. In contrast, we observe that the ReliefF approach, with $k = 3$, tends to select somewhat different features as important, such as the solar radiation and the noise feature. It also includes the erroneous feature of wind gusts at Bearvalley. We conclude that the distance, chi-squared, and stump filters are more robust than ReliefF, and the important features identified by these methods are useful in helping control room operators reduce the number of data streams they must monitor when they consider weather data in making their scheduling decisions.

Finally, for SCE, we consider how the percentage error rate of ASPEN varies as we use the top d features identified by the different dimension reduction methods (Figures 5 and 6). We observe that the feature selection methods, including ReliefF, tend to give lower error rates than the non-linear dimension reduction methods, sometimes even below that obtained by using all features (the horizontal line in the plots). This indicates that some of the features are irrelevant. The noise feature is typically the feature identified as least important by the distance, chi-squared, and stump filters. Note that even though the noise feature is sometimes considered as important by ReliefF, the in-built dimension reduction in the creation of the decision tree, ignores this feature when other more important features are present. We also observe that the error rate with PCA is usually higher than with the non-linear dimension reduction methods.

For the BPA results, we have similar observations. ASPEN tends to give more accurate results than Adaboost or Bagging. The distance, chi-squared, and stump filters tend to identify common weather variables as important, though these variables are different from those for SCE data in the Tehachapi Pass area. This is expected as the meteorological processes in the two regions are very different. As with the SCE dataset, the ReliefF algorithm identifies the noise variable as important; it also selects a different set of weather variables in comparison to the other feature selection methods.

For BPA, we also observe that the reduction in percentage error rate is higher for the lower threshold ramps, when we consider the error made by assigning all instances to the majority class. Use of the decision tree ensemble is better than assigning a class randomly for all the four datasets for this region.

And finally, for the BPA datasets, as with the SCE datasets, the feature selection techniques tend to give lower error rates than the non-linear dimension reduction techniques, which, in turn, are better than PCA. The feature selection methods also reduce the error rate to below what can be obtained by using all the features.

5.2 Results: Problem 2

In this section, we present the results for Problem 2 in Table 3, where we consider daily weather conditions in the SCE region associated with inaccurate forecasts. There are two datasets corresponding to different thresholds used in the calculation of the accuracy of the forecasts. The results are presented in Figure 9 and Table 9.

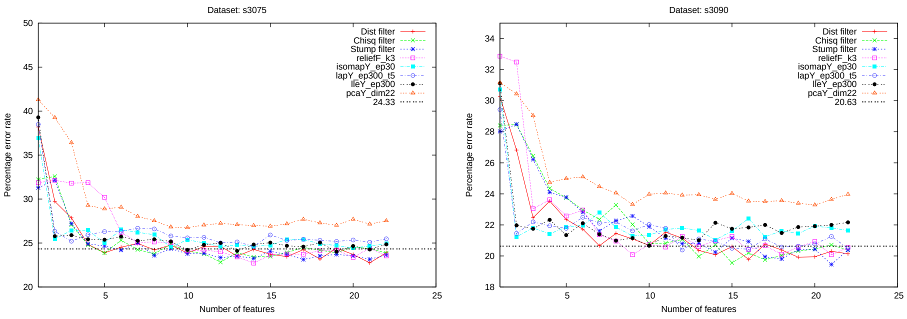
These results reflect the observations made for the SCE datasets for problem 1, where we considered ramp events instead of inaccurate forecasts. The accuracy with ASPEN is better than with Adaboost or Bagging. The use of decision trees is better than making a random call for the class label or selecting the majority class, indicating that the weather conditions have some predictive value in identifying days with inaccurate forecasts.

We again observe that the three filter methods - distance, chi-squared, and stump filters - identify common variables as important, while the ReliefF method selects less important variables, such as solar radiation, noise, or wind gusts at Bearvalley (a variable with many erroneous values). However, unlike problem 1, the filter methods and the non-linear dimension reduction methods give similar results as we consider the variation in percentage error rate with the number of features. Since PCA did not perform well in problem 1, we did not apply it to the problem 2 dataset. We also observe that the results using fewer features are not much better than using all the features.

5.3 Results: Problem 3

In this section, we present the results for Problem 3 in Table 3, where we consider hourly weather conditions in the SCE region associated with inaccurate forecasts. There are two datasets corresponding to different thresholds used in the calculation of the accuracy of the forecasts. The results are presented in Figure 10 and Tables 10. Recall that the weather conditions for this problem are hourly data from a different set of weather stations.

These results reflect the observations made for the SCE datasets for problems 1 and 2. The accuracy with ASPEN is better than with Adaboost or Bagging. However, the use of decision trees does not provide as large an improvement in comparison with making a random call for the class label or selecting the majority class. For the smaller threshold dataset, the improvement is much smaller than before, while for the larger threshold dataset, there is no improvement. This is likely due to the fact that hourly weather data tend to be more noisy as they do not benefit from the smoothing effect of averaging that is used to generate the daily weather data. It is also possible that the weather variables (which are different from, and fewer in number, than those used in



Method	30 min, 10%	30 min, 12%
Adaboost	28.33 (0.55)	22.68 (0.71)
Bagging	25.07 (0.67)	22.35 (0.30)
ASPEN	24.33 (0.45)	20.63 (0.39)

(a) (b) (c)

Fig. 5. Results for problem 1, SCE, 30 minute ramps: Percentage error rate for the ASPEN method for (a) 10% threshold; (b) 12% threshold using the top k features identified by various dimension reduction methods. The horizontal line is the percentage error rate using all features. (c) Percentage error rate (standard error) for different ensemble methods using all features.

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
J_speed_g	B_rhumid_avg	J_rhumid_avg	B_rhumid_avg
B_rhumid_avg	J_rhumid_avg	B_rhumid_avg	noise
J_rhumid_avg	J_speed_g	J_speed_g	P_solar_rad
B_atemp_avg	B_atemp_avg	B_atemp_avg	J_solar_rad
J_precip	J_atemp_avg	J_atemp_avg	J_rhumid_avg
J_atemp_avg	P_speed_g	B_precip	J_speed_g
P_speed_g	P_dir	P_speed_g	B_solar_rad

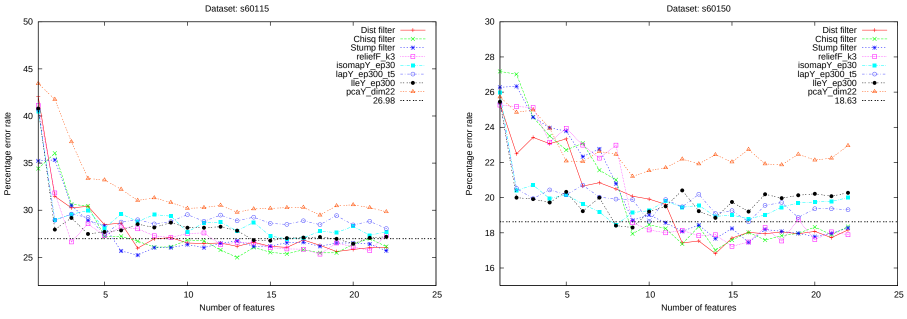
(a)

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
J_speed_g	B_rhumid_avg	J_rhumid_avg	noise
J_speed_avg	J_rhumid_avg	B_rhumid_avg	J_speed_avg
B_rhumid_avg	J_speed_g	J_speed_g	J_atemp_avg
J_rhumid_avg	B_atemp_avg	B_atemp_avg	B_rhumid_avg
B_atemp_avg	J_atemp_avg	J_atemp_avg	P_atemp_avg
J_dir	P_dir	B_precip	B_speed_g
J_atemp_avg	P_speed_avg	P_speed_avg	J_solar_rad

(b)

TABLE 5

Results for problem 1, SCE, 30 minute ramps: The seven top-ranked variables using (a) 10% and (b) 12% thresholds.



Method	60 min, 15%	60 min, 20%
Adaboost	27.92 (2.36)	21.39 (0.81)
Bagging	27.56 (0.28)	19.75 (0.57)
ASPEN	26.98 (0.44)	18.63 (0.25)

(a) (b) (c)

Fig. 6. Results for problem 1, SCE, 60 minute ramps. Percentage error rate for the ASPEN method for (a) 15% threshold; (b) 20% threshold using the top k features identified by various dimension reduction methods. The horizontal line is the percentage error rate using all features. (c) Percentage error rate (standard error) for different ensemble methods using all features.

problems 1 and 2) are not as predictive of hours with inaccurate forecasts.

We again observe that the three filter methods - distance, chi-squared, and stump filters - identify common variables as important, while the ReliefF method selects less important variables, such as noise or GE_temp30. In contrast, the other filter methods all rank noise as the least important variable, and when the noise variable is excluded, give better accuracy than using all the features.

5.4 General observations

We next make some general observations about the results presented in the previous sections, both from a data mining viewpoint and in the context of the insights obtained in enabling the integration of an intermittent resource, such as wind, on the power grid.

First, we observe that the ASPEN method tends to give better accuracy than Adaboost and Bagging. Second, in many of the datasets, especially those with a large number of features, it is possible to improve classification accuracy by using fewer than all the features. This indicates that some weather variables are irrelevant to

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
J_speed_g	B_rhumid_avg	B_rhumid_avg	J_speed_g
B_rhumid_avg	J_rhumid_avg	J_rhumid_avg	B_rhumid_avg
J_rhumid_avg	J_speed_g	J_speed_g	P_solar_rad
J_dir	J_dir	J_speed_avg	noise
J_speed_avg	B_atemp_avg	J_dir	J_solar_rad
B_dir	B_dir	B_atemp_avg	J_rhumid_avg
B_atemp_avg	J_speed_avg	J_atemp_avg	J_atemp_avg

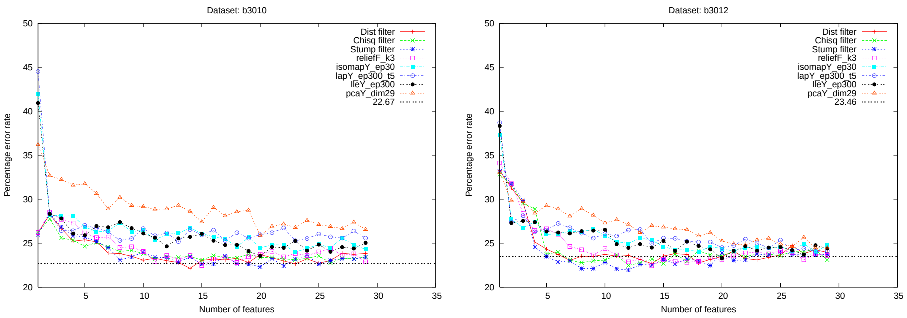
(a)

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
J_speed_g	B_rhumid_avg	B_rhumid_avg	noise
B_atemp_avg	J_rhumid_avg	J_rhumid_avg	B_rhumid_avg
B_rhumid_avg	B_atemp_avg	B_atemp_avg	B_speed_g
J_rhumid_avg	J_speed_g	J_atemp_avg	P_solar_rad
J_atemp_avg	J_atemp_avg	J_speed_g	P_rhumid_avg
J_speed_avg	P_atemp_avg	P_atemp_avg	J_atemp_avg
P_atemp_avg	P_dir	B_precip	J_solar_rad

(b)

TABLE 6

Results for problem 1, SCE, 60 minute ramps. The seven top-ranked variables for 60 min ramps using (a) 115 MW and (b) 150 MW thresholds for Tehachapi Pass.



(a)

(b)

Method	30 min, 10%	30 min, 12%
Adaboost	24.09 (1.54)	26.92 (0.70)
Bagging	24.29 (0.14)	24.44 (0.37)
ASPEN	22.67 (0.40)	23.46 (0.23)

(c)

Fig. 7. Results for problem 1, BPA, 30 minute ramps: Percentage error rate for the ASPEN method for (a) 10% threshold; (b) 12% threshold using the top k features identified by various dimension reduction methods. The horizontal line is the percentage error rate using all features. (c) Percentage error rate (standard error) for different ensemble methods using all features.

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
W_speed_g	W_speed_g	W_speed_g	W_speed_g
P_speed_g	W_speed_avg	W_speed_avg	W_dir
W_speed_avg	P_speed_g	U_speed_g	P_dir
W_dir	W_dir	P_speed_g	noise
U_speed_g	U_speed_g	W_dir	U_speed_g
P_speed_avg	U_speed_avg	P_speed_avg	L_speed_g
L_dir	L_dir	L_dir	P_speed_g

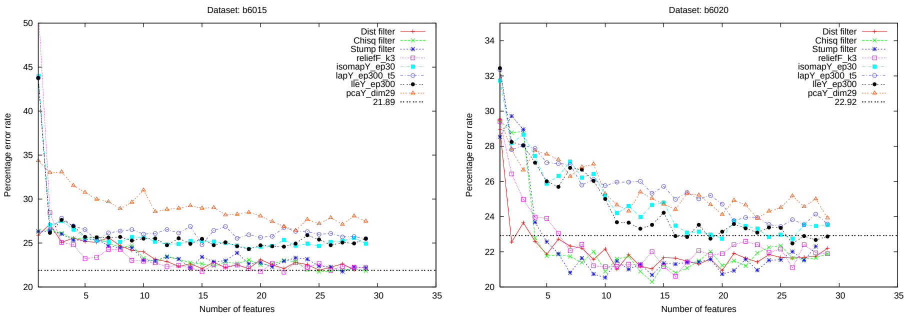
Distance filter	Chi-squared filter	Stump filter	ReliefF filter
W_speed_g	W_speed_g	W_speed_g	W_dir
P_speed_g	P_speed_g	P_speed_g	P_dir
P_speed_avg	U_speed_g	U_speed_g	W_speed_g
W_speed_avg	W_dir	W_speed_avg	L_dir
U_speed_g	W_speed_avg	U_speed_avg	L_rhumid_avg
P_dir	P_dir	W_dir	L_speed_g
W_dir	U_speed_avg	L_precip	W_rhumid_avg

TABLE 7

Results for problem 1, BPA, 30 minute ramps. The seven top-ranked variables for 30 min ramps using (a) 10% and (b) 12% thresholds for mid-Columbia Basin.

ramp events or inaccurate forecasts. Third, the linear and non-linear transform methods are not as accurate as the feature selection methods. The latter are also computationally inexpensive and, by identifying a subset of the original variables, produce results which are interpretable. And finally, the ReliefF algorithm, though a feature selection method, tends to select features which are ranked low by the other methods. We expect that this, and the poor performance of non-linear dimension reduction methods, is partly the result of the use of the nearest neighbors in these algorithms. Identifying the nearest neighbors can be problematic in high-dimensional spaces [24].

Our goal in this work was to investigate if we could provide control room operators additional information they can use to make better decisions in scheduling wind power generation, especially during ramp events and inaccurate forecasts. There are three main observations from this study that are useful to operators. First, the feature selection methods, by ranking the weather variables in order of importance, indicate to the control room operators which variables they should monitor. Second, the plots of the error rates of the decision tree ensembles indicate that the lowest error rate is often obtained using far fewer weather variables than available. This, in effect, implies that the control room operators can reduce the



Method	60 min, 15%	60 min, 20%
Adaboost	23.92 (0.23)	24.46 (0.44)
Bagging	23.80 (0.41)	22.48 (0.47)
ASPEN	21.89 (0.39)	22.92 (0.47)

(a) (b) (c)

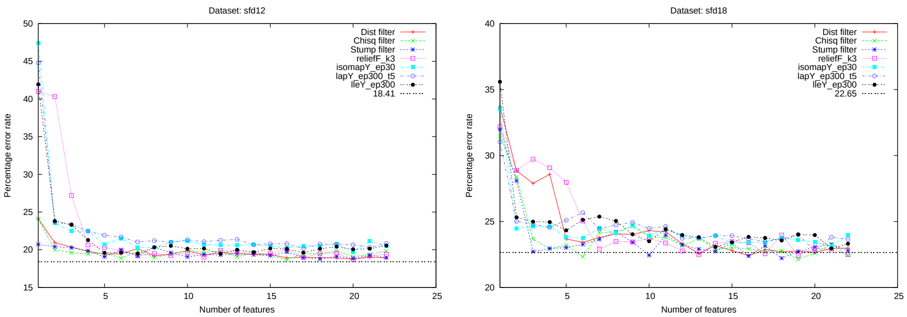
Fig. 8. Results for problem 1, BPA, 60 minute ramps: Percentage error rate for the ASPEN method for (a) 15% threshold; (b) 20% threshold using the top k features identified by various dimension reduction methods. The horizontal line is the percentage error rate using all features. (c) Percentage error rate (standard error) for different ensemble methods using all features.

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
W_speed_g	W_speed_g	W_speed_g	noise
P_speed_g	W_speed_avg	W_speed_avg	W_speed_g
W_speed_avg	W_dir	W_dir	U_speed_g
W_dir	P_speed_g	P_speed_g	L_speed_g
U_speed_g	U_speed_g	U_speed_g	U_speed_avg
P_speed_avg	L_dir	U_speed_avg	P_dir
L_dir	U_speed_avg	L_dir	P_speed_g

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
W_speed_g	W_speed_g	W_speed_g	P_dir
W_speed_avg	P_speed_g	P_speed_g	W_speed_g
P_speed_g	U_speed_g	U_speed_g	W_dir
P_speed_avg	W_speed_avg	W_speed_avg	U_speed_g
U_speed_g	P_dir	L_precip	noise
L_speed_avg	U_speed_avg	U_speed_avg	L_dir
U_atemp_avg	W_dir	W_dir	L_rhumid_avg

TABLE 8

Results for problem 1, BPA, 60 minute ramps. The seven top-ranked variables for 30 min ramps using (a) 15% and (b) 20% thresholds for mid-Columbia Basin.



Method	1200 MW	1800 MW
Adaboost	24.03 (0.13)	26.27 (1.21)
Bagging	21.23 (0.31)	24.22 (0.47)
ASPEN	18.41 (0.35)	22.66 (0.38)

(a) (b) (c)

Fig. 9. Results for problem 2, SCE. Percentage error rate for the ASPEN method for (a) 1200 MW and (b) 1800MW using the top k features identified by various dimension reduction methods. The horizontal line is the percentage error rate using all features. (c) Percentage error rate (standard error) for different ensemble methods using all features.

number of weather data streams they need to monitor. Third, the use of decision tree classifiers to predict days with ramp events or inaccurate forecasts, gives better results than a random guess. While additional weather stations, with more accurate measurements, could certainly improve the predictions, our experiments show that data mining techniques could be beneficial even with the current weather data. However, for the analysis using hourly weather data to predict inaccurate forecasts, we found that more data, of higher quality, are required when we consider analysis at a lower temporal resolution.

6 Related Work

The process of generating wind power forecasts often involves machine learning techniques, used either by themselves or in conjunction with numerical weather prediction codes. These codes do not provide perfect predictions as the atmosphere is a highly non-linear chaotic system [25]. To improve the prediction, forecasters use several approaches [8]. In some cases, they may use more than one code and identify an optimal combination based on a classification of the weather situation. Or they may address systematic forecast errors using linear regression techniques. Classification techniques have also

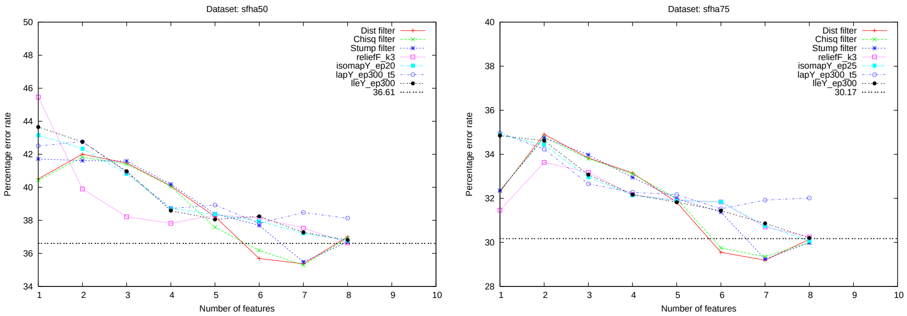
Distance filter	Chi-squared filter	Stump filter	ReliefF filter
J_speed_avg	J_speed_avg	J_speed_g	P_solar_rad
J_speed_g	J_speed_g	J_speed_avg	noise
J_dir	J_dir	J_dir	B_rhumid_avg
P_dir	B_dir	B_dir	J_speed_g
B_dir	P_dir	P_dir	J_solar_rad
P_speed_avg	B_humidity_avg	P_speed_avg	J_speed_avg
P_speed_g	P_speed_avg	B_rhumid_avg	B_speed_g

(a)

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
J_speed_avg	J_speed_g	J_speed_g	J_speed_g
J_speed_g	J_speed_avg	J_speed_avg	J_speed_avg
J_dir	B_rhumidity_avg	J_rhumidity_avg	noise
P_speed_avg	J_rhumidity_avg	B_rhumidity_avg	B_dir
B_rhumidity_avg	P_dir	P_speed_avg	B_speed_g
J_rhumidity_avg	P_speed_avg	P_dir	P_solar_rad
P_dir	J_dir	J_dir	B_rhumid_avg

(b)

TABLE 9
Results for problem 2, SCE. The seven top-ranked variables for (a) 1200 MW and (b) 1800 MW thresholds for Tehachapi Pass.



Method	50 MW	75 MW
Adaboost	38.30 (0.14)	33.38 (0.22)
Bagging	37.67 (0.08)	31.65 (0.07)
ASPEN	36.61 (0.19)	30.17 (0.13)

(a) (b) (c)

Fig. 10. Results for problem 3, SCE. Percentage error rate for the ASPEN method for (a) 50 MW threshold; (b) 75 MW threshold using the top k features identified by various dimension reduction methods. The horizontal line is the percentage error rate using all features. (c) Percentage error rate (standard error) for different ensemble methods using all features.

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
ST_wdir10	ST_wdir10	ST_wdir30	GE_temp30
ST_wdir30	ST_wdir30	ST_wdir10	SR_wspeed30
SR_wdir30	SR_wdir30	SR_wdir30	ST_wdir30
SR_wspeed30	SR_wspeed30	SR_wspeed30	ST_wdir10

(a)

Distance filter	Chi-squared filter	Stump filter	ReliefF filter
ST_wdir10	ST_wdir10	ST_wdir10	GE_temp30
ST_wdir30	ST_wdir30	ST_wdir30	noise
SR_wdir30	SR_wdir30	SR_wdir30	SR_wspeed30
SR_wspeed30	SR_wspeed30	SR_wspeed30	ST_wdir10

(b)

TABLE 10
Results for problem 3, SCE. The seven top-ranked variables for (a) 50 MW and (b) 75 MW thresholds for Tehachapi Pass.

been proposed to convert the wind speed and direction generated by the numerical weather prediction codes into the power forecast from a wind farm [26], [27]. These techniques help to address the uncertainty in the wind speed and direction obtained from the codes and the nonlinearity of the power curve used to convert the speed into power. A similar approach is used in [28], where data mining techniques are used in two ways: to directly predict the power from weather data and to use the weather data to predict the wind speed and then generate the wind power. To reduce the number of weather variables, principal component analysis is first applied to the weather data. Alternately, one could use time series analysis to predict wind power, as described in [29], where the focus is on predicting ramp events.

Our work presented in this paper differs from the above in one crucial way. We are not using data mining techniques to either predict, or improve the prediction of, wind power forecasts from a wind farm. Instead, we assume that the forecasts are already provided and investigate how we can use data mining techniques, applied to historical data, to provide insights the control room operators can use to make better scheduling decisions. Our approach is applicable when the forecasts available are not accurate enough, resulting in a need for additional information.

7 Conclusions and Future Work

In this paper, we discussed how we can use data mining techniques to provide control room operators additional information they can use to make informed decisions in scheduling wind energy on the power grid. Our focus was on days with inaccurate forecasts and ramp events as these conditions pose the greatest challenge to operators. Using data from Tehachapi Pass and mid-Columbia Basin, we showed that feature selection techniques and predictive models, such as decision trees, can provide useful insights.

There are several ways in which we can extend this work. From an algorithm viewpoint, we can investigate some of the newer feature selection methods, such as those based on a spectral approach [30], and explore the use of cost-sensitive learning [31] to address the unequal costs of different types of errors. As the cost information is unlikely to be available precisely, working with cost intervals [32] might be more appropriate in our problem. From the application viewpoint, we can apply the techniques to data over longer time periods to determine how the results are affected by changing weather patterns over several years as well as the increase in installed wind capacity. The longer time series would also allow us to evaluate if the accuracy of predictions is seasonal, for example, is the accuracy of prediction of ramp events in the summer months the same as in the winter months? Finally, as more sensors are deployed to improve the forecasts of wind power generation, we can also exploit the data from these sensors in predictive analysis.

8 Acknowledgment

LLNL-JRNL-562015: This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

This work was part of the WindSENSE project at Lawrence Livermore National Laboratory, funded by the Wind and Hydro Technology program at DOE. We would like to thank Jim Blatchford (CaISO), Robert Farber (SCE), John Manobianco (MESO), Dora Nakafuji (HECO), John Pease (BPA), Scott Winner (BPA), and John Zack (MESO) for technical discussions and insights into issues with integration of wind energy on the power grid. We gratefully acknowledge Min-Lin Cheng (SCE), Robert Farber (SCE), John Pease (BPA), John Zack (MESO), and the Western Regional Climate Center for access to the data used in this study. We also thank the reviewers for their suggestions on alternative algorithms to explore.

References

[1] M. Lange, “On the uncertainty of wind power predictions,” in *Proceedings, AWEA Special Topic Conference: "The Science of Making Torque from Wind*, 2004, available at http://energymeteo.com/media/DELFT2004_uncertainty.pdf.

[2] M. Lange and U. Focken, *Physical Approach to Short-Term Wind Power Prediction*. Spinger, 2005.

[3] C. Monteiro *et al.*, “Wind power forecasting: State-of-the-art 2009,” Argonne National Laboratory, Tech. Rep., November 2009.

[4] D. Hawkins and M. Rothleder, “Evolving role of wind forecasting in market operation at the CaISO,” in *Proceedings, IEEE Power Systems Conference and Exposition*, 2006, pp. 234–238, dOI: 10.1109/PSCE.2006.296304.

[5] “Bonneville power administration wind power web page,” http://www.bpa.gov/corporate/WindPower/.

[6] C. Kamath, “Using simple statistical analysis of historical data to understand wind ramp events,” Lawrence Livermore National Laboratory, Tech. Rep., February 2010, available at http://ckamath.org/publications_by_project.

[7] J. Zack, E. J. Natenberg, S. Young, J. Manobianco, and C. Kamath, “Application of ensemble sensitivity analysis to observational targeting for short term wind speed forecasting,” Lawrence Livermore National Laboratory, Tech. Rep., February 2010, available at http://ckamath.org/publications_by_project.

[8] M. Lange and U. Focken, “New developments in wind energy forecasting,” in *Proceedings, IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, 2008, pp. 1–8.

[9] C. Kamath, “Understanding wind ramp events through analysis of historical data,” in *Proceedings, IEEE PES Transmission and Distribution Conference*, April 2010, available at http://ckamath.org/publications_by_project.

[10] ——, “Understanding wind ramp events through analysis of historical data,” in *Proceedings, IEEE Power and Energy Systems, Transmission and Distribution Conference and Exposition*, March 2011.

[11] “BPA wind projects map: Current and proposed wind project interconnections to BPA transmission facilities,” http://www.transmission.bpa.gov/PlanProj/Wind/documents/map-BPA_wind_interconnections.pdf.

[12] C. Kamath, *Scientific Data Mining: A Practical Perspective*. Society for Industrial and Applied Mathematics (SIAM), 2009.

[13] S. H. Huang, “Dimensionality reduction on automatic knowledge acquisition: a simple greedy search approach,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1364–1373, 2003.

[14] L. Breiman, J. Friedman, R. A. Olshen, and C. Stone, *Classification and Regression Trees*. Boca Raton, Florida: CRC Press, 1984.

[15] K. Kira and L. Rendell, “A practical approach to feature selection,” in *Proceedings of the Ninth International Conference on Machine Learning*, 1992, pp. 249–256.

[16] M. Robnik-Šikonja and I. Kononenko, “Theoretical and empirical analysis of ReliefF and RReliefF,” *Machine Learning*, vol. 53, pp. 23–69, 2003.

[17] J. B. Tenenbaum, V. Silva, and J. C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[18] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, pp. 2323–2326, 2000.

[19] L. K. Saul, S. T. Roweis, and Y. Singer, “Think globally, fit locally: Unsupervised learning of low dimensional manifolds,” *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[20] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, June 2003.

[21] Y. Freund and R. Schapire, “Experiments with a new boosting algorithm,” in *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996, pp. 148–156.

[22] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 26, no. 2, pp. 123–140, 1996.

[23] C. Kamath, E. Cantú-Paz, and D. Littau, “Approximate splitting for ensembles of trees using histograms,” in *Proceedings, Second SIAM International Conference on Data Mining*, 2002, pp. 370–383.

[24] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is “nearest neighbor” meaningful?” in *Proceedings of the 7-th International Conference on Database Theory, Lecture Notes in Computer Science*, vol. 1540. Springer-Verlag, 1999, pp. 217–235.

[25] T. N. Palmer, “Predicting uncertainty in forecasts of weather and climate,” ECMWF Technical Memorandum No. 294, Tech. Rep., November 1999.

[26] L. Fugon, J. Juban, and G. Kariniotakis, “Data mining for wind power forecasting,” in *European Wind Energy Conference and Exhibition (EWEC)*, 2008.

[27] A. Bossavy, R. Girard, and G. Kariniotakis, “Forecasting ramps of wind power production with numerical weather prediction ensembles,” *Wind Energy*, 2012, to appear.

[28] A. Kusiak, H. Zheng, and Z. Song, “Wind farm power prediction: A data-mining approach,” *Wind Energy*, vol. 12, pp. 275–293, 2009.

[29] H. Zareipour, D. Huang, and W. D. Rosehart, “Wind power ramp events classification and forecasting: A data mining approach,” in *Proceedings, IEEE PES Annual General Meeting*, July 2011.

[30] Z. A. Zhao and H. Liu, *Spectral Feature Selection for Data Mining*. CRC Press, 2011.

[31] C. Elkan, “The foundations of cost-sensitive learning,” in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973–978.

[32] X.-Y. Liu and Z.-H. Zhou, “Learning with cost intervals,” in *Proceedings of the ACM SigKDD Conference on Knowledge Discovery and Data Mining*, 2010, pp. 403–412.