LLNL-CONF-457176

Associating Weather Conditions with Ramp Events in Wind Power Generation

C. Kamath

September 27, 2010

IEEE Power Systems Conference and Exposition
Phoenix, AZ, United States
March 20, 2011 through March 23, 2011

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.

Associating Weather Conditions with Ramp Events in Wind Power Generation

Chandrika Kamath, *Senior Member, IEEE*

***Abstract*—As the percentage of wind energy on the power grid increases, the intermittent nature of this energy source can make it difficult to keep the generation and the load balanced. While wind speed forecasts can be helpful, they can often be inaccurate. In such cases, we are interested in providing the control room operators additional relevant information they can exploit to make well informed scheduling decisions. In this paper, we investigate if weather conditions in the region of the wind farms can be effective indicators of days when ramp events are likely. Using feature selection techniques from data mining, we show that some variables are more important than others and offer the potential of data-driven predictive models for days with ramp events.**

***Index Terms*—wind energy, ramp events, weather conditions, feature selection.**

I. Introduction

As renewable resources, such as wind, start providing an increasing percentage of our energy needs, we need to understand these resources better so we can successfully manage their integration into the grid. A key challenge with wind energy is that it is intermittent. We can have days when the wind does not blow, as well as time intervals when there is a sudden sharp increase or decrease in the wind speed over a short period of time, leading to ramp events in the power generated. It can also be difficult to forecast wind speed accurately using numerical weather prediction models [1], especially in regions where the terrain is complex and the meteorological processes controlling the wind speed are difficult to model.

These issues are not a major problem when the percentage of wind energy within a balancing area is small. However, as the percentage increases, it becomes increasingly difficult for control room operators to schedule wind energy and keep the load balanced. The problem is further aggravated if the energy markets levy a high penalty when a wind farm is unable provide the energy scheduled.

In a typical scenario, a control room operator schedules wind energy using a forecast of the energy expected in the next hour from the wind farms in the balancing area. This forecast is based on data from meteorological towers in the region or derived through numerical weather prediction models in combination with meteorological data. The forecasts are usually provided for several hours ahead and updated hourly. If the forecast is accurate, there are no issues in scheduling the wind energy. When the forecast is inaccurate, the operators might look at the actual wind 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.

In the case of positive ramps, where the wind energy increases by a large amount over a short period, the operators must reduce other generation to keep the load balanced. This is a challenge if the positive ramp had not been forecast and the other generation cannot be reduced at short notice. In case of a negative ramp event, the operators must have enough backup power to keep the load balanced. Having this additional back-up might not be cost-effective, especially if a negative ramp is predicted but does not occur.

These issues point to several ways in which we can make it easier to schedule wind energy. For example, we can improve the forecasts provided to the operators by using better computational models, higher resolution models, or by driving the models using more appropriate weather data [2]. Or, we can provide appropriate additional information so the operators can make better informed scheduling decisions when the forecast is inaccurate. Typically, control room operators have access to different meteorological data from several weather stations, though not all of these data are useful or relevant. If we can identify the ones which are associated with extreme events for wind farms in a specific region, then the operators can monitor only those data streams for use in scheduling decisions.

In this paper, we describe how we can analyze historical data to determine if there are specific weather conditions which are associated with ramp events. In Section II, we describe the wind and weather data and outline the techniques for selecting key weather variables in Section III. Section IV describes the results using our test-bed data sets and we conclude in Section V with a summary and ideas for future work.

II. Description of the data

We conduct our analysis using wind energy and weather data from two regions - the Tehachapi Pass in Southern California and the Columbia Basin region on the Oregon-Washington border. The wind generation data are available at 15 minute intervals for the Tehachapi Pass and at 5 minute intervals for the Columbia Basin region. In contrast, the weather data are available at different temporal resolutions from several meteorological towers in the two regions.

A key issue in our work is the temporal resolution we should use for the analysis. If we use a very fine granularity, say every 5-10 minutes, it has certain implications which must be addressed for the analysis to give meaningful results. The

C. Kamath is with the Lawrence Livermore National Laboratory, Livermore, CA 94551 USA e-mail: kamath2@llnl.gov (see http://people.llnl.gov/kamath2).

meteorological towers for which weather data are available are usually not at the site of the wind farms, but several miles away. So, we need to account for a lag (or lead) between the weather measurements and the ramp events, which can be difficult as the lag or lead may vary with the weather conditions. In addition, the weather measurements tend to be quite noisy, and any analysis done using fine granularity data are likely to be error-prone. So, in this initial study, we focus on weather data available as daily averages and associate them with days which either have or do not have ramps of a certain magnitude occurring over a specified time interval. The use of daily averages smooths out the error in the weather measurements and avoids the problem of accounting for a lag or lead between the wind speed at the site of a tower and at the site of the wind farm. In addition, the daily weather summary data are publicly available, while the finer temporal granularity data are not.

A. Wind generation data

We conduct our study using actual wind generation data from wind farms in Tehachapi Pass (Southern California) for the years 2007-2008 and Columbia Basin (border of Oregon and Washington) for the years 2007-2009 [3], [4]. We chose data from the recent past as any analysis of these data is likely to be more relevant. Also, the last few years have seen a large increase in installed wind power, which makes this analysis timely. For example, in the Bonneville Power Administration (BPA) balancing area, which includes the Columbia Basin wind farms, the installed wind capacity has increased from 700 MW in 2006-2007 to over 1300 MW in 2008 and more than 2600 MW in 2009 [5], [6].

The Columbia Basin data available for the period 2007-2009 are the total generation from all the wind farms in the BPA balancing area [7], 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.

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 Vincent and Antelope regions. As these regions are close by, their wind generation is very similar, and we consider the sum of the generation in our analysis. Also, 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 [4].

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

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

Thus, to identify days when ramp events have occurred, we need to select the following:

- **The time interval ΔT:** We considered two cases - 30 minutes and 60 minutes as these are durations typically considered for ramps.
- **The threshold Tr:** This choice was harder. We first selected an absolute threshold of 120 MW and 240 MW for the 30 minute and 60 minute ramps, respectively, for the Columbia Basin 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 Columbia Basin, 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.
- **A way to determine if a day has ramps or not:** We considered several options. The first was to consider 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. The second option 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, were not considered further as they led to too few examples of each of the class labels indicating days with different types of ramps.

B. Weather data

There are several weather data sources available for use in our analysis. In our work, we used the publicly available data from the Western Regional Climate Center (http://wrcc.dri.edu). Specifically, we used the Remote Automated Weather Station (RAWS) data for Oregon

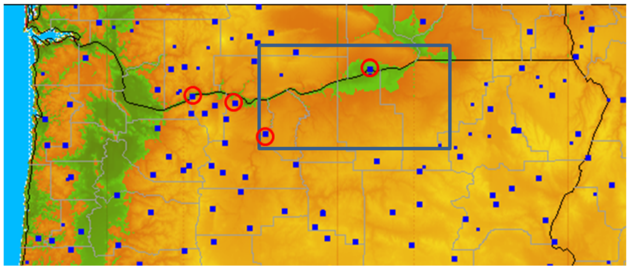


Fig. 1. The Oregon-Washington border region, where the square box indicates the region of the wind farms in the BPA BA. The small squares indicate the meteorological tower locations from WRCC. The four circles indicate the specific sites chosen in our analysis, which are at the following latitude/longitude: Locks (45.669444, 121.881667); Patjens (45.322222, 120.925); Umatilla NWR (45.916667 119.566667); Wasco (45.61,121.33).

(http://www.raws.dri.edu/orF.html) and Southern California (http://www.raws.dri.edu/scaF.html) for the Columbia Basin and Tehachapi Pass regions, respectively.

For each region, we started by considering weather stations near the area of the wind farms. These stations may not be at the most favorable locations for use in analysis of wind-related events as the locations were not selected for this purpose. However, since we are using rather coarse temporal granularity of a day, we wanted to investigate if the data from these stations could still be useful in our analysis. In addition, as the data from weather stations are often of poor quality, with many missing values, and the wind farms are often spread out over a large area, it is unlikely that we will always have access to the most appropriate weather data for our analysis. We also observe that we use the actual weather data, though in practice, current and forecast weather data would need to be used as indicators of ramp events.

Appendix A gives the list of 28 variables available for a weather station. Note that some variables, such as the day of the year or the day of the run (that is, the row number in the data file) are irrelevant and are removed. Some variables (barometric pressure and the average, maximum, and minimum soil temperature) were missing for all days and therefore, removed as well.

Once the initial cleanup was done, the data from each site were further analyzed to determine how many values were missing. For the Columbia Basin region, four sites (Locks, Patjens, Umatilla, and Wasco) had no missing values and were considered in our analysis (see Figure 1). For the Tehachapi Pass region, three sites (Bearvalley, Jawbone, and Piutes) met our criterion of no missing values and were therefore used in the analysis (see Figure 2). For each region, the variables from the selected weather stations were appended to form one long vector which represented the values of the weather conditions in the region for that day.

III. Feature selection techniques

The basic idea in our work is to determine which of the many weather variables at the different sites in a region are associated with ramp events in the wind energy generated in that region. If we can determine a small set of such variables, then the control room operators need only monitor this small set. Further, these variables can also be used to build classification models, such as decision trees or neural networks, which can then be used to predict days likely to have ramp events.

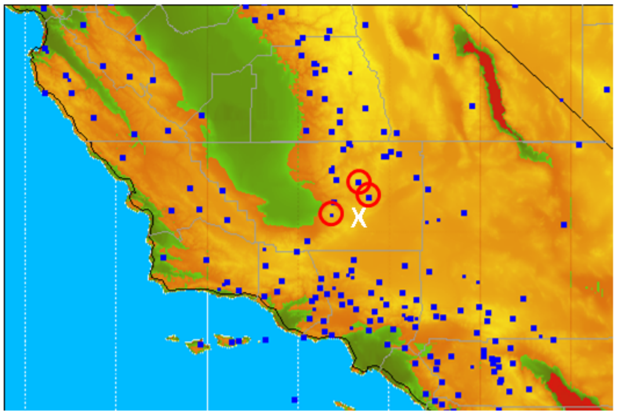


Fig. 2. The Southern California region, where the white cross indicates the Tehachapi Pass area. The small squares indicate the meteorological tower locations from WRCC. The three circles indicate the specific sites chosen in our analysis, which are at the following latitude/longitude: Jawbone (35.294722,-118.226389); Bearvalley (35.139722, -118.625); and Piutes (35.431667, -118.329722), with Tehachapi Pass located at (35.102222, -118.282778).

The task of identifying the key weather variables associated with ramp events is one of dimension reduction, where the dimension, which is the number of variables, is reduced so that only the "important" ones are retained. This topic has been extensively studied in the data mining literature [8]. While transform-based approaches, such as principal component analysis, are an option for reducing the dimension, the reduced dimensional representation is in the form of linear combinations of the original variables, making it difficult to identify which of the original variables are the important ones.

So, we focus on feature selection techniques, where a subset of the original variables (or features) is identified as being relevant to the target variable or the class (in our problem, this would be the occurrence, or lack of occurrence, of ramp events). We consider techniques called "filter" methods in machine learning [8]. These are independent of any classifier which may be used subsequent to the selection of the features. 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 three 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

TABLE II
PERCENTAGE OF DAYS WITH RAMP EVENTS OF A SPECIFIC DURATION THAT EXCEED A SPECIFIC MAGNITUDE (IN MW AS A PERCENTAGE OF THE INSTALLED CAPACITY) FOR THE TEHACHAPI PASS AND COLUMBIA BASIN REGIONS.

Region	30 min. 10%	30 min. 12%	60 min. 15%	60 min. 20%
Tehachapi Pass	42%	28%	43%	23%
Columbia Basin	50%	34%	54%	29%

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

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

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:

$$\delta_j(m, n) = \sum_{i=1}^{b} p_j(d = i|c = m) \log \left(\frac{p_j(d = i|c = m)}{p_j(d = i|c = n)} \right), \tag{2}$$

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 (see table III, adapted from [9]). Numeric features are represented by histograms, so the columns of the contingency table are the histogram bins.

TABLE I
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).

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

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 [10] 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.

IV. EXPERIMENTAL RESULTS

We combine the weather data for each day (described in Section II-B) with a class label derived from the wind-generation data (Section II-A) that indicates if the day has at least one occurrence of a ramp event. For each of the two regions under consideration (Tehachapi Pass and Columbia Basin), we have four files, corresponding to 30 minute ramp events at 10% and 12% installed capacity and 60 minute ramps at 15% and 20% installed capacity. The Tehachapi Pass files include weather data from three meteorological sites and cover the time period 2007-2008 (731 days). The Columbia Basin files have data from four meteorological sites and cover the time period from October 2007 through December 2009 (819 days). Table IV summarizes the percentage of days in each file which have ramp events of a certain magnitude and duration.

A. Initial Processing

We first observed that for each of the weather sites, once we removed the irrelevant variables (the first four in the list in Appendix A), several of the remaining variables are correlated, for example, the air temperature and the fuel temperature, or the two definitions of growing degree days, or the heating degree days and the air temperature average, maximum, and minimum. In addition, the variables could be correlated across the meteorological sites as well, for example, two nearby sites might have correlated air temperatures.

We addressed the issues of correlated variables by removing them before the feature selection. These variables were identified using the Pearson correlation coefficient. 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.

Once the highly correlated variables within a site were removed, the remaining variables (see Appendix B) were used in feature selection. This resulted in a total of 28 and 21 variables for the Columbia Basin and Tehachapi Pass sites, respectively.

In addition, 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.

We also introduced into each dataset a column containing a random noise variable which is uniformly distributed in the interval [0,1]. Variables ranked lower than the noise variable are discarded as they are unlikely to be relevant. In addition, the noise variable can also serve as an indicator of the confidence we can have in the results of a feature selection algorithm - an algorithm which ranks the noise variable relatively high, while other algorithms rank it much lower for the same data set, might indicate the inappropriateness of the algorithm for that data set.

B. *Results for Columbia Basin*

Tables III and IV list the top seven variables identified by each of the three methods for the Columbia Basin region for 30 and 60 minute ramps, respectively. The variables associated with the four weather sites of Locks, Patjens, Umatilla and Wasco, are represented using the prefixes L_, P_, U_, and W_, respectively.

We make several observations on these tables. First, all three methods find certain common variables to be important (these are indicated by bold letters in the tables). The number of these variables can range from 4 to 6, depending on the duration and strength of the ramp event. Of the remaining variables, often two of the methods rank them in the top seven. Second, we observe that most of the top seven variables are quantities related to wind speed, representing the average wind speed, the speed gusts, and the wind direction. Third, certain weather sites tend to occur more frequently in the top seven variables; these are Wasco and Patjens, with Locks occurring rarely.

TABLE III

SEVEN OF THE TOP-RANKED VARIABLES FOR 30 MIN RAMPS USING (TOP) 10% AND (BOTTOM) 12% THRESHOLDS FOR COLUMBIA BASIN. VARIABLES IN BOLD ARE COMMON ACROSS ALL THREE METHODS.

Distance filter	Chi-squared filter	Stump filter
W_speed_g	**W_speed_g**	**W_speed_g**
P_speed_g	**W_speed_avg**	**W_speed_avg**
W_speed_avg	**P_speed_g**	**U_speed_g**
W_dir	**W_dir**	**P_speed_g**
U_speed_g	**U_speed_g**	**W_dir**
P_speed_avg	U_speed_avg	P_speed_avg
L_dir	**L_dir**	**L_dir**

Distance filter	Chi-squared filter	Stump filter
W_speed_g	**W_speed_g**	**W_speed_g**
P_speed_g	**P_speed_g**	**P_speed_g**
P_speed_avg	**U_speed_g**	**U_speed_g**
W_speed_avg	**W_dir**	**W_speed_avg**
U_speed_g	**W_speed_avg**	U_speed_avg
P_dir	P_dir	**W_dir**
W_dir	U_speed_avg	L_precip

TABLE IV

SEVEN OF THE TOP-RANKED VARIABLES FOR 60 MIN RAMPS USING (TOP) 15% AND (BOTTOM) 20% THRESHOLDS FOR COLUMBIA BASIN. VARIABLES IN BOLD ARE COMMON ACROSS ALL THREE METHODS.

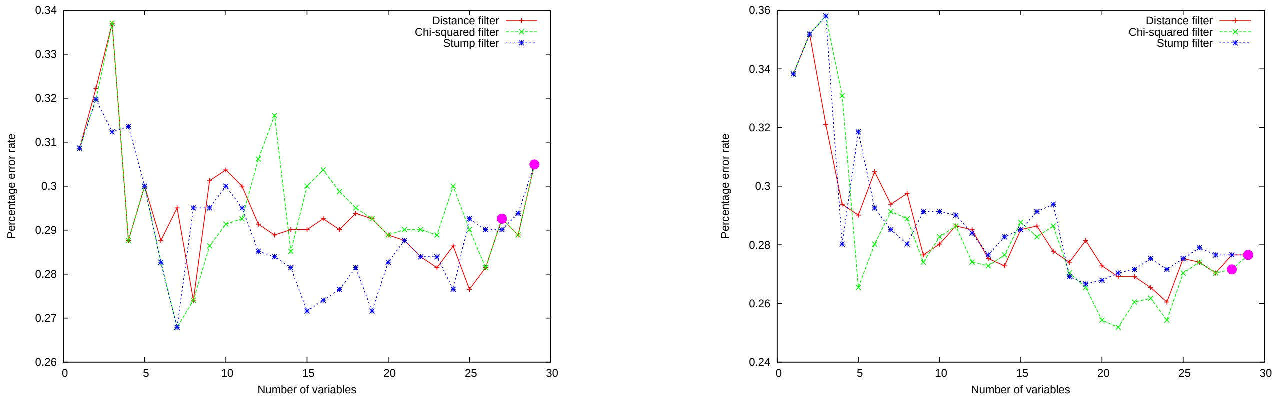
Distance filter	Chi-squared filter	Stump filter
W_speed_g	**W_speed_g**	**W_speed_g**
P_speed_g	**W_speed_avg**	**W_speed_avg**
W_speed_avg	**W_dir**	**W_dir**
W_dir	**P_speed_g**	**P_speed_g**
U_speed_g	**U_speed_g**	**U_speed_g**
P_speed_avg	**L_dir**	U_speed_avg
L_dir	U_speed_avg	**L_dir**

Distance filter	Chi-squared filter	Stump filter
W_speed_g	**W_speed_g**	**W_speed_g**
W_speed_avg	**P_speed_g**	**P_speed_g**
P_speed_g	**U_speed_g**	**U_speed_g**
P_speed_avg	**W_speed_avg**	**W_speed_avg**
U_speed_g	P_dir	L_precip
L_speed_avg	U_speed_avg	U_speed_avg
U_atemp_avg	W_dir	W_dir

We also observed that certain variables such as solar radiation are usually ranked low, indicating that they do not need to be monitored closely.

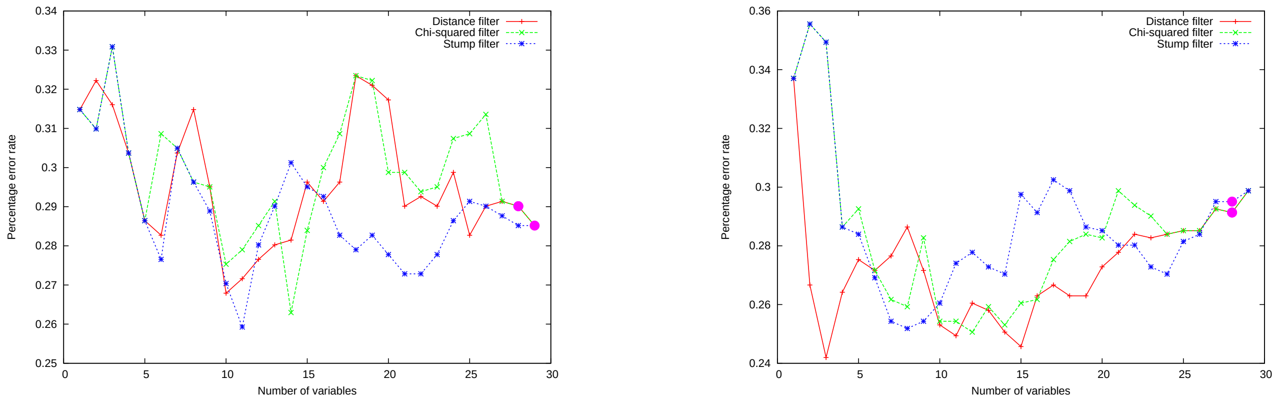
As we considered the top-ranked variables that are associated with ramp events of various durations and magnitudes, it was obvious to ask if we could build a predictive model (such as a decision tree or neural network) which could predict if a day was likely to have ramp events. Figures 3 and 4 show how the error rate of a decision tree, created using the Gini impurity measure, changes for the three feature selection algorithms as we use only the top k important variables to build the tree. This error rate was obtained using 10-fold cross-validation.

These results show that as the number of variables is increased, the error rate first reduces as the important discriminative variables are added to the subset considered. It then often increases with the addition of the less relevant variables. The noise variable (indicated by the large purple dot) is usually the lowest ranked variable. These plots also indicate



(a) (b)

Fig. 3. Error rates for a decision tree created using the top k variables for 30 min ramps using (a) 10% and (b) 12% thresholds for the Columbia Basin data. The purple dot is the noise variable.



(a) (b)

Fig. 4. Error rates for a decision tree created using the top k variables for 60 min ramps using (a) 15% and (b) 20% thresholds for the Columbia Basin data. The purple dot is the noise variable.

that the lowest error rate is usually obtained with far fewer variables than available, implying that control room operators need to monitor only these variables. Further, a decision tree can result in lower error rate in predicting ramp events than random guessing (50% error rate) or assigning the majority label based on the training set. For example, 34% of the days have 30 minute ramps at 12% threshold in Columbia Basin. So, a prediction that every day is a non-ramp day would be wrong 34% of the time. However, a decision tree using 5 of the 21 variables would give an error rate of around 26.5%. This error can be reduced further through the use of more sophisticated models, such as ensembles of trees, or by using better quality weather data at more appropriate locations.

C. Results for Tehachapi Pass

Tables V and VI list the top seven variables identified by each of the three methods for the Tehachapi Pass region for 30 and 60 minute ramps, respectively. The variables associated with the three weather sites of Bearvalley, Jawbone, and Piutes, are represented using the prefixes B_, J_, and P_, respectively.

TABLE V

SEVEN OF THE TOP-RANKED VARIABLES FOR 30 MIN RAMPS USING (TOP) 75MW AND (BOTTOM) 90MW THRESHOLDS FOR TEHACHAPI PASS. VARIABLES IN BOLD ARE COMMON ACROSS ALL THREE METHODS.

Distance filter	Chi-squared filter	Stump filter
J_speed_g	**B_rhumidity_avg**	**J_rhumidity_avg**
B_rhumidity_avg	**J_rhumidity_avg**	**B_rhumidity_avg**
J_rhumidity_avg	**J_speed_g**	**J_speed_g**
B_atemp_avg	**B_atemp_avg**	**B_atemp_avg**
J_precip	**J_atemp_avg**	**J_atemp_avg**
J_atemp_avg	**P_speed_g**	B_precip
P_speed_g	P_dir	**P_speed_g**

Distance filter	Chi-squared filter	Stump filter
J_speed_g	**B_rhumidity_avg**	**J_rhumidity_avg**
J_speed_avg	**J_rhumidity_avg**	**B_rhumidity_avg**
B_rhumidity_avg	**J_speed_g**	**J_speed_g**
J_rhumidity_avg	**B_atemp_avg**	**B_atemp_avg**
B_atemp_avg	**J_atemp_avg**	**J_atemp_avg**
J_dir	P_dir	B_precip
J_atemp_avg	P_speed_avg	P_speed_avg

As in the case of the Columbia Basin data, we observe that all three methods find certain common variables to be important (these are indicated by bold letters in the tables). The number of these variables can range from 5 to 6, depending on the duration and strength of the ramp event. Of the remaining variables, often two of the methods rank them in the top seven.

TABLE VI
SEVEN OF THE TOP-RANKED VARIABLES FOR 60 MIN RAMPS USING (TOP) 115 MW AND (BOTTOM) 150 MW THRESHOLDS FOR TEHACHAPI PASS. VARIABLES IN BOLD ARE COMMON ACROSS ALL THREE METHODS.

Distance filter	Chi-squared filter	Stump filter
J_speed_g	**B_rhumidity_avg**	**B_rhumidity_avg**
B_rhumidity_avg	**J_rhumidity_avg**	**J_rhumidity_avg**
J_rhumidity_avg	**J_speed_g**	**J_speed_g**
J_dir	**J_dir**	**J_speed_avg**
J_speed_avg	**B_atemp_avg**	**J_dir**
B_dir	B_dir	**B_atemp_avg**
B_atemp_avg	**J_speed_avg**	J_atemp_avg

Distance filter	Chi-squared filter	Stump filter
J_speed_g	**B_rhumidity_avg**	**B_rhumidity_avg**
B_atemp_avg	**J_rhumidity_avg**	**J_rhumidity_avg**
B_rhumidity_avg	**B_atemp_avg**	**B_atemp_avg**
J_rhumidity_avg	**J_speed_g**	**J_atemp_avg**
J_atemp_avg	**J_atemp_avg**	**J_speed_g**
J_speed_avg	**P_atemp_avg**	**P_atemp_avg**
P_atemp_avg	P_dir	B_precip

However, unlike the data from the Columbia Basin, not all the top ranked variables are related to wind speed. For example, the average relative humidity at the Jawbone and Bearvalley sites are considered important variables, as are the average air temperatures at the three sites.

In addition, we observed that wind speed variables at the Jawbone site are considered important, though the same variables at the Bearvalley site do not occur in the top-ranked variables. A further investigation indicated that the speed gust at Bearvalley was usually the lowest ranked variable, often even below the noise variable. We found that this variable had many erroneous values - for example, of the total of 731 days in the study, 112 days had speed gusts in Bearvalley of 44.70 m/s, indicating perhaps an inoperative sensor. And finally, as in the case of Columbia Basin, solar radiation was typically a low ranked variable.

Figures 5 and 6 show the error rates for a decision tree created using the top-ranked variables. The observations made for the Columbia Basin data are valid in this case as well.

V. CONCLUSIONS AND FUTURE WORK

In this paper we used feature selection techniques from data mining to identify important weather variables associated with ramp events in wind generation in two regions - Tehachapi Pass and Columbia Basin. We showed that certain variables were identified by the three methods as being important indicators of days with ramp events. In addition, using a simple decision tree model, we showed that we could use these important variables to predict days with ramp events.

Our future work involves a more careful analysis of the weather data to understand why specific variables at specific sites are considered important as well as the use of better quality weather data from more appropriate locations. We will also investigate better models to determine if they can make more accurate predictions of days with ramp events.

APPENDIX A
LIST OF WEATHER VARIABLES FROM EACH METEOROLOGICAL STATION

The WRCC weather stations provide summary information on the following twenty-eight variables (along with units) for each day:

```
 1   Date
 2   Year
 3   Day of year
 4   Day of run
 5   Solar Rad. total kW-hr/m2
 6   Speed average m/s
 7   Wind dir vector deg
 8   Speed Gust m/s
 9   Air temp Average deg C
10   Air temp Maximum deg C
11   Air temp Minimum deg C
12   Fuel Temp Average deg C
13   Fuel Temp Maximum deg C
14   Fuel Temp Minimum deg C
15   Soil temp Average deg C
16   Soil temp Maximum deg C
17   Soil temp Minimum deg C
18   Relative humidity Average percent
19   Relative humidity Maximum percent
20   Relative humidity Minimum percent
21   Barometric Pressure Average mbar
22   ASCE Et. total mm
23   Penman Et. total mm
24   Heating Degree Days
25   Cooling Degree Days
26   Growing Degree Days Base 40
27   Growing Degree Days Base 50
28   Precipitation Total mm
```

APPENDIX B
UNCORRELATED WEATHER VARIABLES FROM EACH METEOROLOGICAL STATION

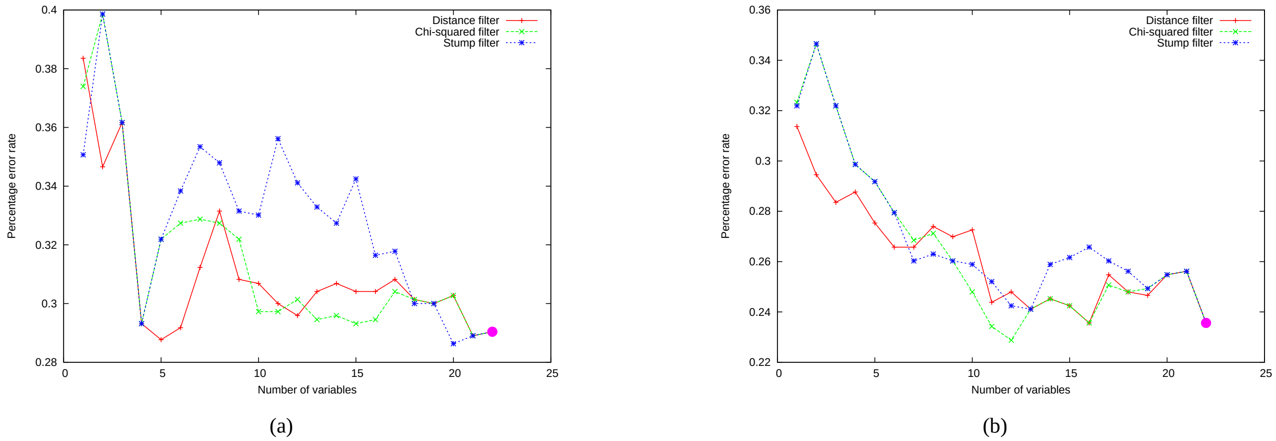
The following seven weather variables were used for each meteorological site:

```
1   Solar Rad. total kW-hr/m2
2   Speed average m/s
3   Wind dir vector deg
4   Speed Gust m/s
5   Air temp Average deg C
6   Relative humidity Average percent
7   Precipitation Total mm
```

ACKNOWLEDGMENT

This work was funded through the Renewable System Interconnect Program in the DOE Office of Energy Efficiency and Renewable Energy. We thank Min-Lin Cheng (SCE), Robert Farber (SCE), John Pease (BPA), and the Western Regional Climate Center for access to the data used in this study.

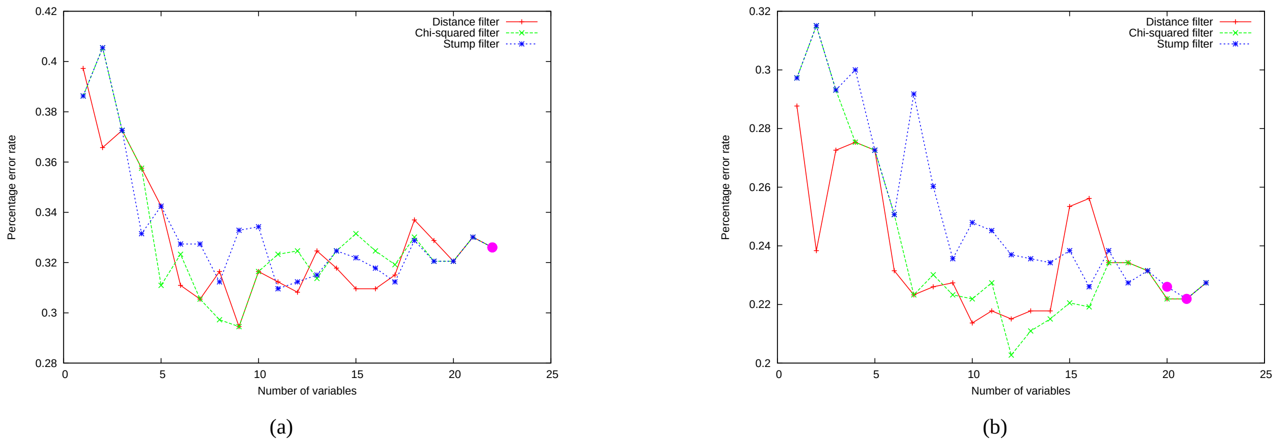
LLNL-CONF-457176: This work performed under the auspices of the U.S. Department of Energy by Lawrence



(a)

(b)

Fig. 5. Error rates for a decision tree created using the top k variables for 30 min ramps using (a) 75 MW and (b) 90 MW thresholds for the Tehachapi Pass data. The purple dot is the noise variable.



(a)

(b)

Fig. 6. Error rates for a decision tree created using the top k variables for 60 min ramps using (a) 115 MW and (b) 150 MW thresholds for the Tehachapi Pass data. The purple dot is the noise variable.

Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

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

[2] 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.

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

[4] ——, “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.

[5] “Balancing act: BPA grid responds to huge influx of wind power,” Bonneville Power Administration Fact Sheet. http://www.bpa.gov/corporate/pubs/fact_sheets/08fs/Wind-Balancing-act-Nov2008.pdf.

[6] “How BPA supports wind power in the Pacific Northwest,” Bonneville Power Administration Fact Sheet. http://www.bpa.gov/corporate/pubs/fact_sheets/09fs/BPA_supports_wind_power_for_the_Pacific_Northwest_-_Mar_2009.pdf.

[7] “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.

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

[9] 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.

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

Chandrika Kamath Chandrika Kamath is a researcher at Lawrence Livermore National Laboratory, where she is involved in the analysis of data from scientific simulations, observations, and experiments. She received her Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign. Her research interests include signal and image processing, machine learning, pattern recognition, and statistics, as well as the application of data mining techniques to the solution of practical problems.