# Fine-grained Photovoltaic Output Prediction using a Bayesian Ensemble

**Prithwish Chakraborty**[1,2], **Manish Marwah**[3], **Martin Arlitt**[3], and **Naren Ramakrishnan**[1,2]
[1]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061
[2]Discovery Analytics Center, Virginia Tech, Blacksburg, VA 24061
[3]Sustainable Ecosystems Research Group, HP Labs, Palo Alto, CA 94304

**Abstract**

Local and distributed power generation is increasingly reliant on renewable power sources, e.g., solar (photovoltaic or PV) and wind energy. The integration of such sources into the power grid is challenging, however, due to their variable and intermittent energy output. To effectively use them on a large scale, it is essential to be able to predict power generation at a fine-grained level. We describe a novel Bayesian ensemble methodology involving three diverse predictors. Each predictor estimates mixing coefficients for integrating PV generation output profiles but captures fundamentally different characteristics. Two of them employ classical parameterized (naive Bayes) and non-parametric (nearest neighbor) methods to model the relationship between weather forecasts and PV output. The third predictor captures the sequentiality implicit in PV generation and uses motifs mined from historical data to estimate the most likely mixture weights using a stream prediction methodology. We demonstrate the success and superiority of our methods on real PV data from two locations that exhibit diverse weather conditions. Predictions from our model can be harnessed to optimize scheduling of delay tolerant workloads, e.g., in a data center.

## Introduction

Increasingly, local and distributed power generation e.g., through solar (photovoltaic or PV), wind, fuel cells, etc., is gaining traction. In fact, integration of distributed, renewable power sources into the power grid is an important goal of the smart grid effort. There are several benefits of deploying renewables, e.g., decreased reliance (and thus, demand) on the public electric grid, reduction in carbon emissions, and significantly lower transmission and distribution losses. Finally, there are emerging government mandates on increasing the proportion of energy coming from renewables, e.g., the Senate Bill X1-2 in California, which requires that one-third of the state's electricity come from renewable sources by 2020.

However, renewable power sources such as photovoltaic (PV) arrays and wind are both variable and intermittent in their energy output, which makes integration with the power grid challenging. PV output is affected by temporal factors such as the time of day and day of the year, and environmental factors such as cloud cover, temperature, and air pollution. To effectively use such sources at a large scale, it is essential to be able to predict power generation. As an example, a fine-grained PV prediction model can help improve workload management in data centers. In particular, a data center's workloads may be "shaped" so as to closely match the expected generation profile, thereby maximizing the use of locally generated electricity.

In this paper, we propose a Bayesian ensemble of three heterogeneous models for fine-grained prediction of PV output. Our contributions are:

1. The use of multiple diverse predictors to address fine-grained PV prediction; while two of the predictors employ classical parametrized (naive Bayes) and non-parametric (nearest neighbor) methods, we demonstrate the use of a novel predictor based on motif mining from discretized PV profiles.
2. To accommodate variations in weather profiles, a systematic approach to weight profiles using a Bayesian ensemble; thus accommodating both local and global characteristics in PV prediction.
3. Demonstration of our approach on real data from two locations, and exploring its application to data center workload scheduling.

## Related Work

Comprehensive surveys on time series prediction (Brockwell and Davis 2002; Montgomery, Jennings, and Kulahci 2008) exist that provide overviews of classical methods from ARMA to modeling heteroskedasticity (we implement some of these in this paper for comparison purposes). More related to energy prediction, a range of methods have been explored, e.g., weighted averages of energy received during the same time-of-the-day over few previous days (Cox 1961). Piorno et al. (2009) extended this idea to include current day energy production values as well. However, these works did not explicitly use the associated weather conditions as a basis for modeling. Sharma et al. (2011b) considered the impact of the weather conditions explicitly and used an SVM classifier in conjunction with a RBF kernel to predict solar irradiation. In an earlier work (Sharma et al. 2011a), the same authors showed that irradiation patterns

and Solar PV generation obey a highly linear relationship, and thus conclude that irradiance prediction was in turn predicting the Solar PV generation implicitly. In other works, Lorenz et al. (2009) used a benchmarking approach to estimate the power generation from photovoltaic cells. Bofinger et al. (2006) proposed an algorithm where the forecasts of an European weather prediction center (of midrange weathers) were refined by local statistical models to obtain a fine tuned forecast. Other works on temporal modeling with applications to sustainability focus on motif mining; e.g., Patnaik et al. (2011) proposed a novel approach to convert multivariate time-series data into a stream of symbols and mine frequent episodes in the stream to characterize sustainable regions of operation in a data center. Hao et al. (2011) describe an algorithm for peak preserving time series prediction with application to data centers.

## Problem Formulation

Our goal is to predict photovoltaic (PV) power generation from i) historic PV power generation data, and, ii) available weather forecast data. Without loss of generality, we focus on fine-grained prediction for the next day in one hour intervals. Such predictions are useful in scheduling delay tolerant work load in a data center that "follows" renewable supply (Krioukov et al. 2011). Furthermore, these predictions need to be updated as more accurate weather forecast data and generation data from earlier in the day become available.

Let us denote the actual PV generation for $j^{th}$ hour of $i^{th}$ day by $a_{i,j}$. Let $I$ be the number of days and $J$ be the maximum number of hours per day for which data is available. Then, the actual PV generation values for all the time points can be expressed as a $I \times J$ matrix, which we denote as $A = [a_{i,j}]_{I \times J}$ Corresponding to each entry of PV generation, there is a vector of weather conditions. Assuming $K$ unique weather attributes (such as temperature, humidity, etc.), each time point is associated with vector $\omega_{i,j} = \langle \omega_{i,j}[1], \omega_{i,j}[2], ..\omega_{i,j}[K] \rangle$, where $\omega_{i,j}[k]$ denotes the value for the $k$-th weather condition for the time-slot $(i, j)$. Corresponding to the PV generation matrix, we can define a matrix of weather conditions given by $\omega = [\omega_{i,j}]_{I \times J}$. Finally, for each time slot, weather forecast data is continuously collected. The predicted weather condition for $j^{th}$ hour of the $i^{th}$ day, at an offset of $t$ hours is given by the vector $\rho_{i,j,t} = \langle \rho_{i,j,t}[1], \rho_{i,j,t}[2], ..., \rho_{i,j,t}[K] \rangle$.

Then, with reference to time-slot $(e, f)$, given the values for $a_{i,j}$ and $\omega_{i,j}, \forall (i, j) \leq (e, f)$ and weather forecast $\rho_{e,f+1,1}, \rho_{e,f+2,2}, .., \rho_{e,J,J-f}$; we need to determine the prediction for PV generation for the remaining time slots of the day, i.e., $(e, f), (e, f + 1), ..(e, J)$.

## Methods

We propose a novel Bayesian ensemble method that aggregates diverse predictors to solve the PV prediction problem described above. An architectural overview of the method is shown in Figure 1. As an initial pre-processing step, we determine the common modes of daily PV generation profiles via distance based clustering of historical generation data. Once such profiles are discovered, we represent the unknown PV generation of a day as a mixture of these profiles. The mixture coefficients are independently estimated using three different predictors and finally averaged using a Bayesian approach. These predictors are (i) a modified naive Bayes classifier that derives its features from weather forecast; (ii) a weighted k-nearest neighbor classifier that uses past generation during the same day as features; and, (iii) a predictor based on motifs of day profiles in the recent past.

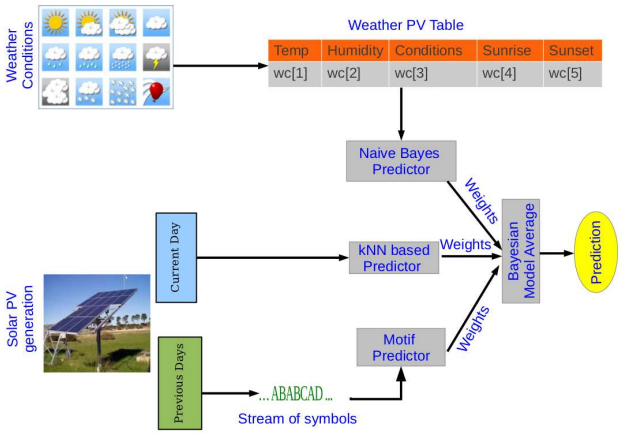



Figure 1: Schematic Diagram of the proposed Bayesian Ensemble method. All the predictors make use of the profiles discovered from historical data.

**Profile Discovery.** The first step is profile discovery which takes as input the available historic PV generation data and outputs characteristic day long profiles. Let us denote the actual PV generation for an entire $i^{th}$ day by the vector $\vec{x_i}$ which can be written as: $\vec{x_i} = \langle a_{i,1}, a_{i,2}, \cdots, a_{i,J} \rangle$. Then we can express the entire PV dataset ($A$) as, $A = (\vec{x_1}, \vec{x_2}, .., \vec{x_I})^T$.

This dataset $A$ is clustered using Euclidean distance between the $J$ dimensional feature vectors (i.e power generation for each day) by k-means (Lloyd 1982) algorithm into $N$ clusters. The value of $N$ is a parameter in the ensemble method and is estimated by minimizing the cross-validation error of the ensemble, keeping other parameters fixed. This yields $N$ day long profiles. Let us denote these profiles by $D = \{D_1, D_2, .., D_N\}$ and the corresponding centroids by $\mu = \{\mu_1, \mu_2, \cdots \mu_N\}$. This step is required to be run only once on the entire dataset.

**Naive Bayesian predictor.** The NB predictor estimates the mixture coefficients given the weather forecast, assuming conditional independence of features (we assume they follow Gaussian distributions). If we denote all the training information obtained from the weather-PV table, such as the likelihood functions and priors of the profiles, by $\gamma$, and the weather forecast by $\rho_{i,j} = \langle \rho_{i,j+1,1}, \rho_{i,j+2,2}, \cdots, \rho_{i,J,J-j} \rangle$, the posterior probability of profile labels, for each remaining time slots, is computed

as:

$$Pr(D_n|\rho_{i,j+t,t},\gamma) \propto \left(\prod_k L(D_n|\rho_{i,j+t,t}[k])\right) Pr(D_n)$$

finally giving

$$Pr(D_n|\rho_{i,j},\gamma,C_1) = \frac{\sum_{t=1}^{J-j} Pr(D_n|\rho_{i,j+t,t},\gamma)}{\sum_{n=1}^{N}\sum_{t=1}^{J-j} Pr(D_n|\rho_{i,j+t,t},\gamma)} \quad (1)$$

where $C_1$ indicates classifier 1.

**k-NN based predictor.** The k-NN (Dudani 1976) based predictor uses prior PV generation during the same day as a feature and assigns mixing coefficients based on Euclidean distance from centroids of discovered daily profiles. In order to make a prediction at the $j^{th}$ hour of the $i^{th}$ day for the rest of that day, we consider the already observed PV output values for the $i^{th}$ day $\vec{x}_i(1:j) = \{a_{i,1}, a_{i,2}, \cdots, a_{i,j}\}$. Next, we find the Euclidean distance of this vector to the truncated centroid vectors (first $j$ dimensions) of the PV profiles and find the probability of the $i^{th}$ day belonging to a cluster as given by the following equation.

$$Pr(\vec{x_i} \in D_n|\vec{x}_i(1:j), C_2) = \frac{1}{\phi\|\vec{x}_i(1:j) - \vec{\mu}_n(1:j)\|_2} \quad (2)$$

where $\phi$ is a normalizing constant found as: $\phi = \sum_n \frac{1}{\|\vec{x}_i(1:j)-\vec{\mu}_n(1:j)\|_2}$, where $C_2$ indicates classifier 2.

**Motif based Predictor.** The final predictor exploits the sequentiality in PV generation between successive days to find motifs and give membership estimates of the profiles based on such motifs. For this step, we consider the entire PV data as a stream of profile labels. We further consider a "window size": the maximum number of past days that can influence the profile and treat the stream as a group of vectors of the form $d_{i-1}, d_{i-2}, \cdots, d_{i-W}$ where $d_j \in D$ $(j < i)$ denotes the profile label of the $j$th data point. Sliding the window we can get different values of such vectors and can mine for motifs.

*Definition 1: For a window $W_i$ ($|W_i| = W$) containing labels $\langle d_{i-W}, \cdots, d_{i-2}, d_{i-1}\rangle$, eligible episodes are defined as all such sequences $ep = \langle d_{p_1}, d_{p_2}, d_{p_3}, \cdots\rangle$, such that $p_1 < p_2 < p_3 < \cdots$.*

Definition 1 formalizes the term eligible. As evident from the definition, we allow episodes to contain gaps. The only criterion is that they must maintain the temporal order. Furthermore, we can define an episode $ep_1$ to be a sub-episode of $ep_2$, denoted by $ep_1 \leq ep_2$, if $ep_1$ is a sub-sequence of $ep_2$. From the definition of sub-episodes it is evident that if $ep_1 \leq ep_2$ then, $ep_1 \notin W_j$ implies $ep_2 \notin W_j$; i.e. the sub-episode property is anti-monotonic. The support of an episode $ep_i$ ,denoted by $sup_{ep_i}$, is equal to the number of windows $W_j$, such that $ep_i \in W_j$. Finally we can now defined a maximal frequent eligible episode as follows:

*Definition 2: An eligible episode $ep_i$ is maximal frequent iff:*

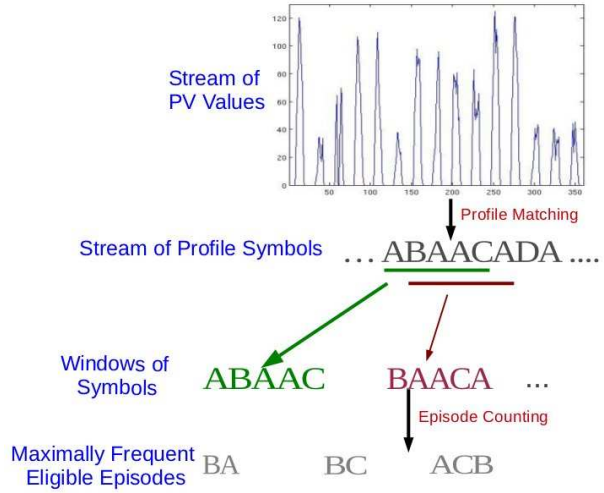



Figure 2: Illustration to show the process of finding motifs

1. $sup_{ep_i} > \tau$
2. $\nexists ep_j$ *such that* $ep_i \leq ep_j$

Thus, we can use an Apriori approach (Agrawal and Srikant 1994) to prune out episodes when trying to find the maximally frequent ones (as relation is anti-monotonic).

From the training data set, we consider such windows of size $W$ and through apriori-counting as in (Patnaik et al. 2011) get the entire corpus of maximally frequent episodes ($FE$) and their corresponding supports. We refer to such maximally frequent episodes as motifs. Here $\tau$, the support threshold is a parameter and as before, we estimate this value through cross-validation keeping the other values constant.

While predicting for the $i$-th day, we need to find motifs which contains the $i$-th label and labels of some of the previous days. For this we chose the immediately preceding window of size $W-1$ (since the label for $i$-th day must already be part of the motifs). To find mixing coefficient of profile $D_n$ for the $i$th day, we consider all those maximally frequent episodes that end with $D_n$ denoted by $ep(D_n)$. Let us denote the set of all such episodes by $\langle ep(D_n)\rangle = \{ep(D_n)_1, ..., ep(D_n)_P\}$ where $P$ denotes the number of such episodes in the set. Then, within a window $W_i$, support of the entire set is given by:

$$sup(\langle ep(D_n)\rangle) = \sum_p sup_{ep(D_n)_p}$$

Then membership of a profile is given by:

$$Pr(\vec{x}_i \in D_n|\vec{x}_{i-1}, \vec{x}_{i-2}, .., \vec{x}_{i-W-1}, C_3) = \frac{sup(\langle ep(D_n)\rangle)}{\sum_{n=1}^{N} sup(\langle ep(D_n)\rangle)} \quad (3)$$

where $C_3$ indicates classifier 3.

This counting step can be potentially very expensive. However, for the current problem the best window sizes are small. Through cross-validation we picked a window of size

5 (from the range 3 to 10). Hence, the counting step, even if naively implemented, is not too expensive.

**Bayesian Model Averaging** Finally, we aggregate the memberships obtained from the three predictors and combine them to arrive at the final prediction as follows.

$$\begin{aligned}
&P\left(\vec{x_i} \in D_n | \mathcal{D}_{i,j} = \langle \vec{x}_i(1:j), \rho_{i,j}, \vec{x}_{i-1}, \vec{x}_{i-2}, .., \vec{x}_{i-W-1} \rangle\right) \\
&= \sum_{l=1}^{3} P\left(\vec{x_i} \in D_n, C_l | \mathcal{D}_{i,j}\right) \\
&= \sum_{l=1}^{3} P\left(\vec{x_i} \in D_n | C_l, \mathcal{D}_{i,j}\right) \times P\left(C_l | \mathcal{D}_{i,j}\right) \\
&= P\left(\vec{x_i} \in D_n | C_1, \rho_{i,j}\right) P\left(C_1 | \mathcal{D}_{i,j}\right) + \\
&\quad P\left(\left(\vec{x_i} \in D_n | C_2, \vec{x}_i(1:j)\right) P\left(C_2 | \mathcal{D}_{i,j}\right) + \right. \\
&\quad \left. P\left(\vec{x_i} \in D_n | C_3, \vec{x}_{i-1}, \vec{x}_{i-2}, .., \vec{x}_{i-W-1}\right) P\left(C_3 | \mathcal{D}_{i,j}\right)\right.
\end{aligned} \tag{4}$$

We use Bayesian Model Averaging (BMA), as outlined in (Raftery et al. 2005), operating on mutually exclusive parts of the data to compute the values of $P(C_l|\mathcal{D}_{i,j})$.

$$P(C_l|\mathcal{D}_{i,j}) \propto P(\mathcal{D}_{i,j}|C_l) \times P(C_l) \tag{5}$$

Assuming a uniform prior on the classifiers, equation 5 can be written as:

$$P(C_l|\mathcal{D}_{i,j}) \propto P(\mathcal{D}_{i,j}|C_l) \tag{6}$$

The values of $P(\mathcal{D}_{i,j}|C_l)$ can be viewed as the proportion of data explained (truly predicted) when using classifier $C_i$. This can be estimated by constructing a confusion matrix and taking the relative frequency of true positives as an estimate.

Then the predicted solar PV values can be estimated as :

$$\begin{aligned}
&\mathbb{E}\left(\vec{x}_i(j+1:J)|\mathcal{D}_{i,j}\right) \\
&= \sum_{\vec{x}_i(j+1:J)} \vec{x}_i(j+1:J) P\left(\vec{x}_i(j+1:J)|\mathcal{D}_{i,j}\right) \\
&= \sum_{\vec{x}_i(j+1:J)} \vec{x}_i(j+1:J) \sum_{n=1}^{N} P\left(\vec{x}_i(j+1:J), \vec{x_i} \in D_n | \mathcal{D}_{i,j}\right) \\
&= \sum_{\vec{x}_i(j+1:J)} \vec{x}_i(j+1:J) \sum_{n=1}^{N} P\left(\vec{x}_i(j+1:J) | \vec{x_i} \in D_n\right) \\
&\qquad \times P\left(\vec{x_i} \in D_n | \mathcal{D}_{i,j}\right) \\
&= \sum_{n} \sum_{\vec{x}_i(j+1:J)} \left(\vec{x}_i(j+1:J) P\left(\vec{x}_i(j+1:J) | \vec{x_i} \in D_n\right)\right) \\
&\qquad \times P\left(\vec{x_i} \in D_n | \mathcal{D}_{i,j}\right) \\
&= \sum_{n} \vec{\mu}_n(j+1:J) P\left(\vec{x_i} \in D_n | \mathcal{D}_{i,j}\right)
\end{aligned} \tag{7}$$

## Baseline Models

**Previous Day as prediction** This is the simplest baseline model, where prediction for a particular hour during the day is simply the PV generation during the same hour on the previous day.

**Autoregression with weather** In this model, autoregression outputs as in (Piorno et al. 2009) are explicitly modified by the influence of weather information. All weather attributes except sunrise and sunset times are clustered into $NC$ groups using k-means clustering (Lloyd 1982) based on actual solar PV generations corresponding to these conditions in the training set. A set of such weather attributes is represented by the mean Solar PV value of the corresponding cluster, denoted by $l(i,j)$. The model can then be given separately for $J-1$ hours (based on offset of prediction) as:

$$\begin{aligned}
A_{i,j} \quad = \quad & \beta_1^t * 1 + \beta_2^t * l(i,j) + \beta_3^t * [i - SR(j)] \\
& + \beta_4^t * [ST(j)i] + \beta_5^t * P_{i,j,t}^a + \epsilon
\end{aligned} \tag{8}$$

where $P_{i,j,t}^a$ is the auto-regressed prediction from (Piorno et al. 2009), $\epsilon$ is the error term we try to minimize and the weights $\{\beta\}$ are estimated using linear least square regression.

**Stagewise Modelling** Another baseline model tried is a stagewise model. This draws inspiration from (Hocking 1976), where the prediction is done in correlated stages : improving the result at every additional stage.

Here we consider an average model to the actual data as the first stage, auto-regression as the next and, weather regression as the final one, where only the error values from a preceding stage is passed onto the next one.

## Experimental Results

**Datasets.** Data was collected from a 154 kW PV installation at a commercial building in Palo Alto, CA. PV output (in kWh) in 5 min intervals was collected from March 2011 to November 2011 (267 days) and aggregated hourly. Hourly weather data was collected from a nearby weather station and included temperature, humidity, visibility and weather conditions, which mainly relate to cloud cover and precipitation. We also needed weather forecast data, which is typically not available after the fact. We ran a script to collect weather forecast data every hour. We also collected solar and weather data from a site in Amherst, MA.

**Data preparation.** The data for this experiment contains a total of 3,747 data points with 69 missing values. While the PV generation values are numeric, the corresponding weather data contains both numeric and categorical variables. The numeric values were imputed using linear interpolation while a majority mechanism over a window was used for the categorical ones. We use 6-fold cross validation for evaluating predictor performance and for estimating some parameters. To account for seasonal dependence, each month is uniformly split among all folds.

**Parameters.** Some heuristics were applied to compute the probabilities of the Bayesian ensemble method. As mentioned, the likelihood needs to be estimated for the classifiers ($C_l$). We assume the values $P(\mathcal{D}_{i,j}|C_l)$ to be dependant only on the hour of the day i.e. we neglect the effect of seasons on classifier beliefs. Under this assumption, ideally the values need to be estimated for each hour. Instead here we apply a simple heuristic. We assume that the data is progressively explained better by the k-NN estimator ($C_2$) while the motif estimator, which estimates in a global sense i.e., without looking at the data, explains the data in a consistent manner irrespective of the hour of the day. These heuristics are given below:

$$\begin{aligned}
&P(\mathcal{D}_{i,j}|C_3) = \theta \\
&P(\mathcal{D}_{i,j}|C_2) = min(1-\theta, \alpha \times j + \beta) \\
&P(\mathcal{D}_{i,j}|C_1) = 1 - \theta - P(\mathcal{D}_{i,j}|C_2)
\end{aligned} \tag{9}$$

where all the values in the left hand side of the equations are bounded between 0 and 1 during computation.

**Predictor training.** For all methods, parameters (including the heuristic combination weights) were selected by training over a range of values and picking the one with least cross-validation error. The basic daily generation profiles were extracted by k-means clustering over the entire dataset. The number of clusters were set at 10 based on cross-validation over 5 to 15 clusters. Some of the daily profiles obtained are shown in Figure 3. Each plot represents a cluster, and the intra-cluster variance is shown by boxplots. Cluster 2 shows a high level of generation and a low variance, likely corresponding to clear summer days, while Cluster 6 is likely related to cloudy conditions, winter days when the angle of sun is low, or when some of PV panels are not operating due to an anomaly.

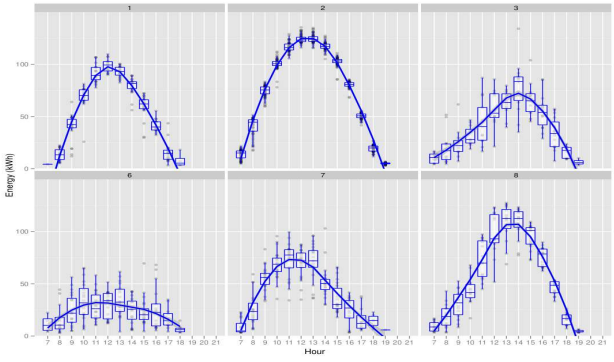



Figure 3: Six of the ten daily profiles obtained via clustering.

For the motif-based predictor, a threshold support parameter of $\tau = 70$ and a window size of 5 were used. A total of 49 maximal frequent episodes were found. Since this number is not large, during prediction for finding motifs that end in a particular symbol, even a naive strategy for counting support will work. Also, since the motifs are day long, this only needs to be done once per day.

**Error metrics.** For comparing the performance of the models, three distinct error metrics were considered. Using $A$ to denote the actual output and $P$ for the predicted, these are defined as — (1) Percentage absolute error: $\sum_{A_{i,j}>3} \left|\frac{A_{i,j}-P_{i,j}}{A_{i,j}}\right| * 100\%$; (2) Percentage root mean square error: $\sqrt{\sum_{A_{i,j}>3} \left(\frac{A_{i,j}-P_{i,j}}{A_{i,j}}\right)^2} * 100\%$; (3) Relative absolute error: $error = \frac{\sum |P_{i,j}-A_{i,j}|}{\sum |A_{i,j}-\bar{A}_j|} * 100$. For (1) and (2), errors for only the values $A_{i,j} > 3$ are considered as errors corresponding to smaller values in the denominator may dominate the overall error. Also, for the system concerned any PV generation less than 3kwH is unusable.

**Results** Based on the cross-validation error, we evaluate two variants of the proposed methods. These are Ensemble2, which is a Bayesian combination of two predictors (NB and k-NN) and Ensemble3, which includes the motif-based predictor as well. In addition, three baseline methods are also included: previous day as prediction (PreviousDay), autoregression with weather (ARWeather), and stagewise regression (Stagewise). The results of these methods are summarized in Table 1.

| Method | Testing Error | | |
|---|---|---|---|
| | Per. Abs. Error | Per. RMS Error | Rel. Abs. Error |
| PreviousDay | 20.54 | 20.65 | 20.81 |
| ARWeather | 18.54 | 18.31 | 19.73 |
| Stagewise | 12.77 | 12.68 | 15.66 |
| Ensemble2 | 10.04 | 10.01 | 10.01 |
| Ensemble3 | 8.13 | 8.21 | 8.34 |

Table 1: Performance at 1-hour offset.

Our proposed methods perform better than the three baseline methods. Ensemble3 performs about 4% better than the best baseline method (Stagewise), and about 1% better that Ensemble2. We also present an unpaired t-test of the competing methods against Ensemble3 in Table 2 and we find that our results are statistically significant.The box plot of percentage errors together with the raw values are shown in Figure 4. The red dot indicates the average value. The average error and variance is least for Ensemble3. As expected, PreviousDay fares the worst. Figure 5 shows the residual errors for the five methods. Again, the superior performance of Ensemble2 and Ensemble3 (both in terms of average and variance) is apparent.

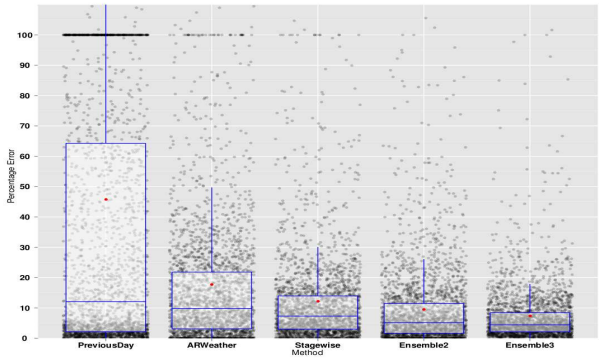



Figure 4: Comparison of the error (%) of different methods.

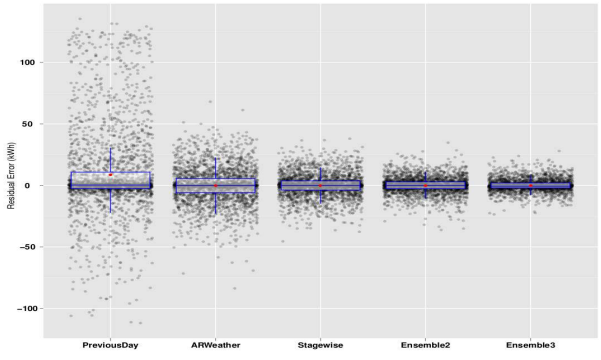



Figure 5: Residual error of the methods.

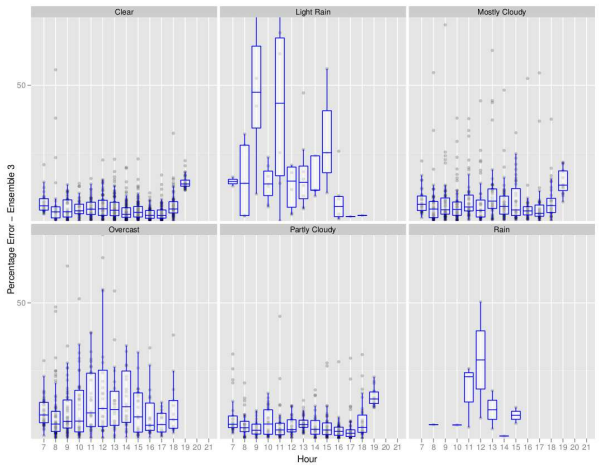



Figure 6: Error conditioned on weather and hour of day (Ensemble3 method).

| Method | Unpaired t-test with respect to Ensemble3 | | | | |
|---|---|---|---|---|---|
| | Std. Error | t | 95% Conf. Interval | Two-tailed P | Significance |
| PreviousDay | 1.113 | 11.1474 | 9.9295 to 14.8905 | $< 0.0001$ | Extremely Significant |
| ARWeather | 0.882 | 11.8070 | 8.4455 to 12.3745 | $< 0.0001$ | Extremely Significant |
| Stagewise | 0.667 | 6.9570 | 3.1539 to 6.1261 | $< 0.0001$ | Extremely Significant |
| Ensemble2 | 0.514 | 3.7167 | 0.7650 to 3.0550 | 0.0040 | Very Significant |

Table 2: Unpaired t-tests on 1-hr prediction data compared against Ensemble3. $N = 6$ and degrees of freedom$= 10$

Figure 6 shows the percentage error distribution for Ensemble3 conditioned on weather and hour of day. The errors are lowest for clear days, and worst for rain conditions. Cloudy conditions increase both average error and variance.

We observed that some of the predictions were higher than the capacity of the PV system, so for all methods we capped the maximum predicted output. In fact, we realized we could do better. From almost an year of data, we found the maximum generation for each hour and bounded the output by that amount plus a small margin. This can be further improved to add the month (season) as well. This optimization was applied to all methods and the gain in performance was in the range of 0.6% to 1%.

**Updating predictions.** As we get closer to the hour for which a prediction is made, we update the prediction based on better weather forecast and the PV generation already seen that day. Figure 7 shows the progressive improvement in average accuracy of PV output prediction for 12pm. The plot shows the cross-validation percentage absolute errors with standard deviation marked. The Bayesian ensemble method (Ensemble3) performs the best. One likely reason is the fact that in the ensemble method we predict memberships of the characteristic daily profiles and thus consider a more global view. On the other hand, the regression models are tailored to one hour prediction and the future hour predictions are based on assuming the predictions as the true value for the unknown hours (for more than one hour offset in prediction). Furthermore, the standard deviation for the ensemble method is also lower than the other models (except the average one), mainly because we are selecting among some known patterns and standard deviation results from difference in the mixing coefficients rather than completely new predictions as is the case in regression models.

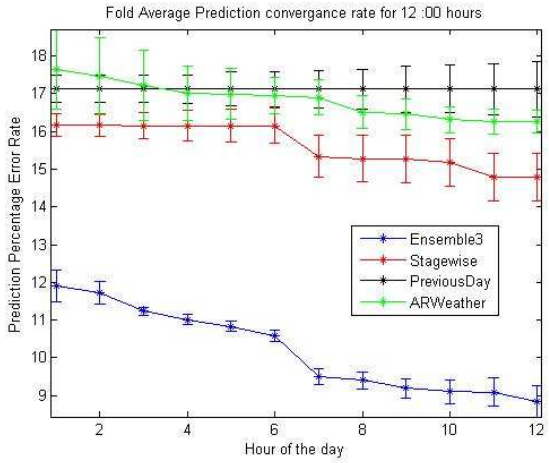
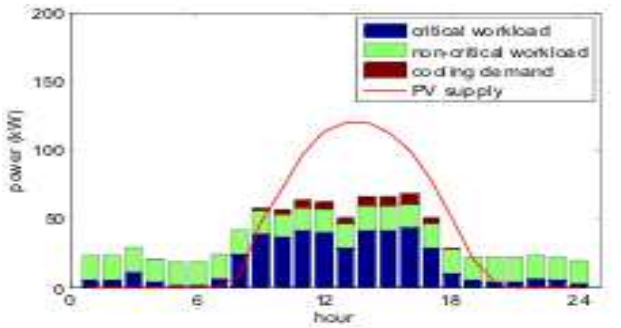



Figure 7: Change in error(%) as the prediction is updated.

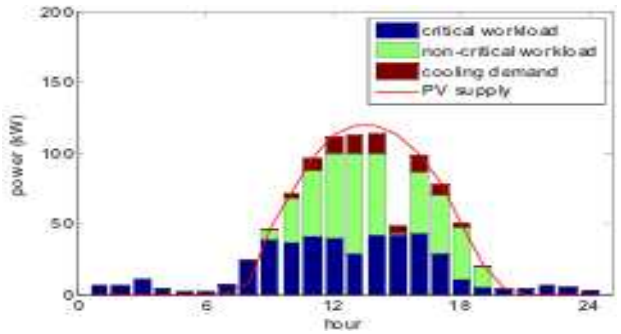
**Applying fine-grained PV prediction to data centers.** The accurate prediction of PV generation is an important functionality required by proposed net-zero energy data centers (Arlitt et al. 2012). Data centers use a lot of electricity, and their operators are looking for ways to reduce their dependence on public electric grids. Within data centers, an opportunity exists to reschedule non-critical (i.e., delay tolerant) workloads such that they run when PV generated electricity is available, so operators have the option of turning off IT and cooling infrastructure at other times of the day. Having accurate predictions of the available PV output enables this renewable source of energy to be utilized much more effectively. For example, Figure 8(a) shows a typical data center workload where non-critical jobs are uniformly scheduled throughout the day. If accurate, fine-grained PV predictions are available, then the non-critical jobs can be scheduled such that they run only when PV power is available (Figure 8(b)). In this specific case, the optimized workload schedule results in 65% less grid energy being consumed than with the original schedule, as the data center is able to consume most of the renewable energy directly (Arlitt et al. 2012).

## Discussion

We have demonstrated a systematic approach to integrate multiple predictors for PV output prediction. Our encouraging results sets the foundation for more detailed modeling. In particular, it is known (Sharma et al. 2011b) that solar irradiance and PV generation generally follow a highly linear relationship. Thus, our models here can easily be adapted to predict for solar irradiance. We have analyzed historical data for the site mentioned in (Sharma et al. 2011b) for the period March 2006 to May 2007. The cross-fold average rms error for the Bayesian ensemble method for the




(a)




(b)

Figure 8: (a) Data center workload schedule; (b) Optimized workload schedule based on predicted PV energy supply (Arlitt et al. 2012)

entire period was found to 101 watts/$m^2$. This was significantly lower than that for the SVM-RBF model proposed in (Sharma et al. 2011b) where for a period of observation Jan 2010 to Oct 2010 the reported prediction rms error was 128 watts/$m^2$. (However, due to unavailability of predicted values of weather conditions for the site, we used actual weather conditions to make our prediction.) Ongoing work is focused on validating the efficacy of predicting the irradiance as a precursor to PV generation.

## References


Agrawal, R., and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*.

Arlitt, M.; Bash, C.; Blagodurov, S.; Chen, Y.; Christian, T.; Gmach, D.; Hyser, C.; Kumari, N.; Liu, Z.; Marwah, M.; McReynolds, A.; Patel, C.; Shah, A.; Wang, Z.; and Zhou, R. 2012. Towards the design and operation of net-zero energy data centers. In *13th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*.

Bofinger, S., and Heilscher, G. 2006. Solar electricity forecast - approach and first results. In *20th European PV conference*.

Brockwell, P. J., and Davis, R. A. 2002. *Introduction to Time Series and Forecasting*. New York: Springer-Verlag, 2 edition.

Cox, D. R. 1961. Prediction by exponentially weighted moving averages and related methods. *Royal Statistical Society* 23(2):pp. 414–422.

Dudani, S. A. 1976. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man and Cybernetics* SMC-6(4):325–327.

Hao, M.; Janetzko, H.; Mittelstdt, S.; Hill, W.; Dayal, U.; Keim, D.; Marwah, M.; and Shama, R. 2011. A visual analytics approach for peak-preserving prediction of large seasonal time Series. In *EuroVus*.

Hocking, R. R. 1976. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics* 32(1):pp. 1–49.

Krioukov, A.; Goebel, C.; Alspaugh, S.; Chen, Y.; Culler, D.; and Katz, R. 2011. Integrating renewable energy using data analytics systems: Challenges and oppurtunities. In *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

Lloyd, S. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory* 28(2):129–137.

Lorenz, E.; Remund, J.; Mller, S. C.; Traunmller, W.; Steinmaurer, G.; Pozo, D.; Ruiz-Arias, J. A.; Fanego, V. L.; Ramirez, L.; Romeo, M. G.; Kurz, C.; Pomares, L. M.; ; and Guerrero, C. G. 2009. Benchmarking of different approaches to forecast solar irradiance. In *24th European Photovoltaic Solar Energy Conference*.

Montgomery, D. C.; Jennings, C. L.; and Kulahci, M. 2008. *Introduction to Time Series Analysis and Forecasting*. Wiley.

Patnaik, D.; Marwah, M.; Sharma, R. K.; and Ramakrishnan, N. 2011. Temporal data mining approaches for sustainable chiller management in data centers. *ACM Transactions on Intelligent Systems and Technology* 2(4).

Piorno, J. R.; Bergonzini, C.; Atienza, D.; and Rosing, T. S. 2009. Prediction and Management in Energy Harvested Wireless Sensor Nodes. In *Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on*.

Raftery, A. E.; Gneiting, T.; Balabdaoui, F.; and Polakowski, M. 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review* 133:1155–1174.

Sharma, N.; Gummeson, J.; Irwin, D.; and Shenoy, P. 2011a. Leveraging weather forecasts in energy harvesting systems. Technical report, University of Massachusetts Amherst, Tech. Rep.

Sharma, N.; Sharma, P.; Irwin, D.; and Shenoy, P. 2011b. Predicting solar generation from weather forecasts using machine learning. In *Proceedings of Second IEEE International Conference on Smart Grid Communications(SmartGridComm)*.