Probabilistic Models for One-Day Ahead Solar Irradiance Forecasting in Renewable Energy Applications

Carlos V. A. Silva*, Lipyeow Lim*, Duane Stevens† and Dora Nakafuji‡

* Dept. of Info. & Comp. Sciences
University of Hawaiʻi at Mānoa
Honolulu, HI, USA
Email: {cvas,lipyeow}@hawaii.edu

† Dept. of Atmospheric Sciences
University of Hawaiʻi at Mānoa
Honolulu, HI, USA
Email: dstevens@soest.hawaii.edu

‡ Hawaiian Electric Company
Honolulu, HI, USA
Email:dora.nakafuji@heco.com

***Abstract*—Solar irradiance forecasting is an important problem in renewable energy management where any dips in solar energy generation must be made up for by reserves in order to ensure an uninterrupted energy supply. In this paper, we study several data mining methods for short term solar irradiance forecasting at a given location. In particular, we apply linear regression, probabilistic models, and naive Bayes classifier to forecast solar irradiance one day ahead, i.e., we forecast what tomorrow's solar irradiance will be like at sundown today. We evaluate the forecasting performance of our adaptations of the three models using land-based weather data from several weather stations on the island of Oahu in Hawaiʻi.**

I. Introduction

The increasing adoption of grid-linked photovoltaic energy generation systems at the residential and commercial scale has created the need for accurate forecasts for solar irradiance (measured in Watts per square meter, W/m^2) in order for energy grid operators to manage the variability in renewable energy supply. Essentially energy grid operators need to ramp up conventional energy generation whenever there is a drop in the renewable energy generation in order to meet energy demands. Meteorologists have used physics-based numerical weather prediction (NWP) models such as the Weather Research and Forecasting (WRF) model to make weather (including solar irradiance) forecasts, but such systems often cannot take into account the rich abundance of land-based sensor data that are now available. Computational scientists have also used machine learning techniques such as artificial neural networks for this forecasting problem with reasonable results [7]. The solar irradiance forecasting problem is in fact a suite of related forecasting challenges depending on the type of data available, the time granularity of the data, the lead time, the spatial granularity and many other parameters. It is generally recognized that NWP methods are best for a lead time of a few days, whereas machine learning methods may be more suitable for shorter lead times up to a day or two.

In this paper we address a particular instance of the solar irradiance forecasting problem. Given historical and current time series data of weather variables (most notably, solar irradiance) measured using land-based sensors, we would like to forecast the solar irradiance for the brightest daylight hours (0800-1700) of the next day given the current day's data (up to 1700 hours).

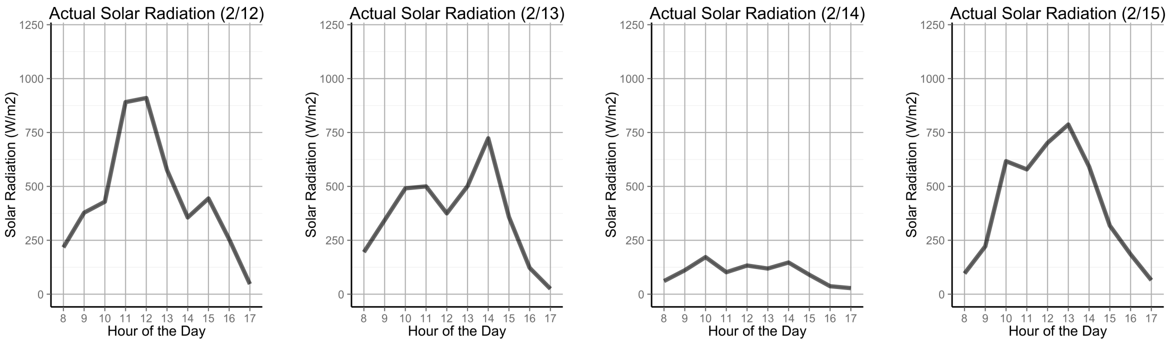
As an example, consider the solar irradiance at Schofield Barracks on the island of Oahu (in Hawaiʻi) between February 12th and 15th, 2014 as shown in Fig. 1. At 1700 on Feb. 13, 2014, we would like to forecast the solar irradiance at 0800, 0900, ..., 1700 on Feb. 14, 2014. In this particular case, an accurate forecast would be very useful for energy grid operators to plan for standard oil-based generation f or the overcast day (Feb. 14).

In this paper we are primarily interested in the use of probabilistic modeling techniques for this problem; however, we also discuss the use of linear regression model as a point of comparison. We have investigated the use of more sophisticated probabilistic models such as Bayesian networks, but have settled on using joint distributions on the solar irradiance time series due to the following reasons. Our analysis of the information-theoretic dependency between the solar irradiance and weather variables at prior times (at the same location) show that solar irradiance is mostly dependent on precipitation and solar irradiance at prior times. Unfortunately, that result is not consistent across different locations on the island. Moreover, there are many locations where we do not have precipitation data. Hence, for this paper we focus on building joint distribution models on the solar irradiance time series data. We also consider the Naive Bayes classifier as an alternative to the joint distribution model.

In order to study how well probabilistic models perform for the solar irradiance forecasting problem, we chose publicly available weather observation data from land-based weather stations on the island of Oahu in the state of Hawaiʻi, USA to build, tune and test our probabilistic models.

The contributions of this paper are as follows:

1) We propose specific adaptations of linear regression models, probabilistic models, and Naive Bayes models for the specific 1-day ahead solar irradiance forecasting problem described above. The importance of this specific problem is validated with the local Hawaiian utility company. We evaluated our models



(c) Overcast Day

Fig. 1. Land-based observations of solar irradiance (W/m^2) at Schofield Barracks from 02/12/2014 to 2/15/2014 between 0800 and 1700.

using real weather station data from the island of Oahu in Hawai'i.

2) We studied the mutual information between daily solar irradiance profiles and (other) weather variables at different prior times for different locations on the island. We found that solar irradiance is consistently dependent on itself at prior times. For some locations, precipitation also yields significant mutual information. The mutual information characterization is different for each location.
3) We found that the size of the training data set has an effect on the forecast error patterns. Naive Bayes classifier seems to perform very well for small training data sets (2 years).
4) Dynamically (for each prediction) choosing a model from an ensemble of probabilistic models using entropy or entropy weighted with support increases the robustness of the forecast.
5) In general, probabilistic models that are dependent on the solar irradiance profile the previous day provides good forecasting performance. If the data set is small, Naive Bayes classifier might be more accurate. Linear regression is not far behind in accuracy.

The rest of the paper is organized as follows. Section II will describe the forecasting models we used with our proposed adaptations. Section III will describe the experiments and the results. Section IV will describe related work and we conclude in Section V.

II. Modeling Solar Irradiance

A. Problem Formulation

We assume that land-based weather stations collect observation data on solar irradiance (W/m^2) and other weather variables such as temperature, relative humidity, etc. at a fixed sampling rate. The data associated with each weather variable is conceptualized as a time series.

> Given historical weather time series data, learn a model that would predict the solar irradiance for the daylight hours of the next day given the observation data up to 1700 of the current day.

Since we are interested in daylight hours with significant amounts of solar insolation, we ignore data points between 1701 to 0759.

We investigate the use of three different forecasting models for this problem: Linear Regression, Naive Bayes, Probabilistic Model.

B. Linear Regression Models

Given that the solar irradiance time series is continuous, an obvious solution is to use linear regression. Let each data point in the solar irradiance time series be denoted by S_t, where t denotes the timestamp at the granularity of hours. To forecast the solar radiation of each hour 1-day ahead, we create a separate *linear regression model* for each hour of the next day using all hours of the previous (consecutive) w days. For example, the model for predicting $S_{20140214.0900}$ with $w = 1$ would be,

$$\begin{aligned} S_{20140214.0900} &= c_1 \cdot S_{20140213.1700} \\ &\quad + c_2 \cdot S_{20140213.1600} + c_3 \cdot S_{20140213.1500} \\ &\quad + \ldots + c_{10} \cdot S_{20140213.0800} + c_{11}. \qquad (1) \end{aligned}$$

Hence, there would be 10 linear regression models for each hour between 0800 and 1700. We use the standard least squares algorithm to obtain the coefficients c_i for each model.

C. Probabilistic Models

An interesting alternative to using continuous techniques (eg. linear regression) is to discretize the data and apply probabilistic techniques. Since we are interested in 1-day ahead prediction, we discretize the data into daily profiles (or weather regimes). For each day in the data set, we construct a vector of solar irradiance values from 0800h to 1700h. The set of daily vectors are then fed to a clustering technique such as k-means. Figure 2 shows the daily profiles (centroids) from k-means clustering [6] with $k = 5$ and the euclidian distance measure. In consultation with domain experts, it was determined that $k = 5$ gives profiles that fit human intuition of the solar characteristics for that location.

Using the five daily profiles from k-means clustering, we transform the original solar irradiance time series (hourly sampling rate or less) into a sequence of daily solar irradiance profiles. We now redefine S_t to be the discrete random variable for the daily profile at a particular date t. Oftentimes, we will use relative dates in the subscript of S_t instead of actual date

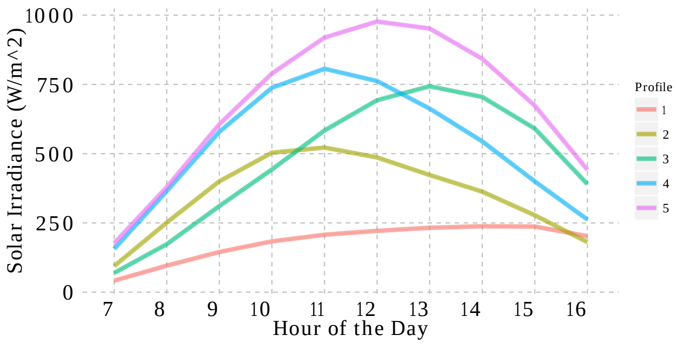


Fig. 2. Solar irradiance (W/m^2) profiles from 01/01/2003 to 12/31/2013 between 8 AM and 5 PM of Schofield Barracks. The centroids are obtained by applying k-means with k = 5 and euclidian distance as similarity function. C0875 profiles are similar and therefore omitted from the work.

such as $t = 20030101$. Given a window size w, we construct a joint probability distribution as

$$P(S_t, S_{t-1}, \ldots, S_{t-w+1}), \tag{2}$$

and use the following prediction function to predict S_t given that the previous $(w-1)$ days' profiles are $\langle s_1, s_2, \ldots, s_{w-1}\rangle$,

$$\begin{aligned}\hat{s} \;\; = \;\; & \arg\max_s P(S_t{=}s|S_{t-1}{=}s_1, S_{t-2}{=}s_2, \ldots, \\ & S_{t-w+1}{=}s_{w-1}).\end{aligned} \tag{3}$$

Note that the joint probability distribution can be easily estimated by counting the number of occurrences of each distinct w-day sequence in the discretized time series data. For the days when the previous $(w-1)$-day sequence $\langle s_1, s_2, \ldots, s_{w-1}\rangle$ does not occur in the training data at all, the conditional distribution does not exist for that sequence and we return the most frequent daily solar irradiance profile (using the prior distribution) as the prediction.

We are still faced with one problem: if we have a set of values for w, we will have a set of different probabilistic models (one for each value of w). How do we choose which model to use (i.e. choose w)? It should be clear that w is associated with the question of how much history is needed to predict the profile for the next day. We investigate three approaches:

Fixed Choose the w that minimizes the prediction errors on the training data and use the same w for all the testing instances,

Entropy Given a testing instance, dynamically choose the w that minimizes the entropy of the posterior distribution (Eqn 3),

Support Similar to entropy, but further weight the entropy with support.

The fixed method is straightforward and requires no further explanation.

For the entropy method, we are given a testing instance which is a profile sequence $\langle s_1, s_2, \ldots, s_{w-1}\rangle$ and need to predict the profile on day i. Let $H(w)$ denote the entropy for the distribution,

$$P(S_t|S_{t-1}{=}s_1, S_{t-2}{=}s_2, \ldots, S_{t-w+1}{=}s_{w-1}). \tag{4}$$

The entropy method would choose w as

$$\hat{w} = \arg\min_w H(w). \tag{5}$$

Minimizing the entropy ensures that the model with the most skewed posterior distribution is chosen. Recall that a uniform distribution cannot distinguish between the five profiles. A skewed distribution, on the other hand, means that the historical data tends to favor a particular profile given the $w-1$ previous days' profiles.

One difficulty with the entropy method is that it is possible for a skewed distribution to be based on very few data points (giving us less confidence), while a less skewed distribution may be supported by a large portion of the data (giving us more confidence). Note as the window size w increases, the number conditional variables $(S_{t-1}{=}s_1, S_{t-2}{=}s_2, \ldots, S_{t-w+1}{=}s_{w-1})$ increases, but the number of occurrences of each distinct sequence of $\langle s_1, s_2, \ldots, s_{w-1}\rangle$ decreases, resulting in lower support for that conditional distribution. To account for the support, we weight the entropy by the number of supporting data points for that distribution. The support method would choose w as

$$\hat{w} = \arg\min_w \frac{H(w)}{N(s_1, s_2, \ldots, s_{w-1})}, \tag{6}$$

where $N(s_1, s_2, \ldots, s_{w-1})$ denote the number of occurrences of the sequence $\langle s_1, s_2, \ldots, s_{w-1}\rangle$ in the data.

D. Naive Bayes Classifier

An alternative to using the joint distribution with different window size w is to apply the Naive Bayes assumption, i.e., the predictor variables $(S_{t-1}, \ldots, S_{t-w+1})$ are independent given the dependent variable. The solar irradiance profile on the date t can then be predicted using

$$\hat{s} = \arg\max_s P(S_t{=}s) \prod_{i=1}^{w-1} P(S_{t-i}|S_t{=}s). \tag{7}$$

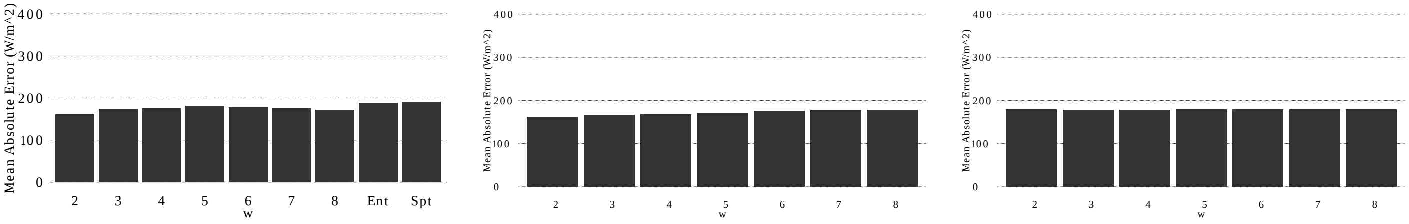
Observe that the Naive Bayes assumption is used to factorize the full joint distribution into a product of lower-order joint distributions ($P(S_{t-i}|S_t{=}s)$'s). For small data sets, estimating the lower-order distributions would often be more accurate than estimating the full joint distribution.

III. Experiments

Data. We used publicly available weather station data from Mesowest Network[1]. While we experimented on all weather stations with solar irradiance data on the island of Oahu, we present results on two specific weather stations. Schofield Barracks (SCBH1) station is located in the center of Oahu and has data starting from 2003 to 2014 (giving a total of 4079 time points). New Jersey Avenue (C0875) is located on the southeast of Oahu and we used the data between 2012 and 2014 when solar irradiance data is available for that station. Whenever the data sampling rate is finer than every hour, we aggregate (usually average) the data to hourly time points.

Preprocessing. Using real sensor data has many challenges. The foremost being missing data. While interpolation

[1] http://mesowest.utah.edu/

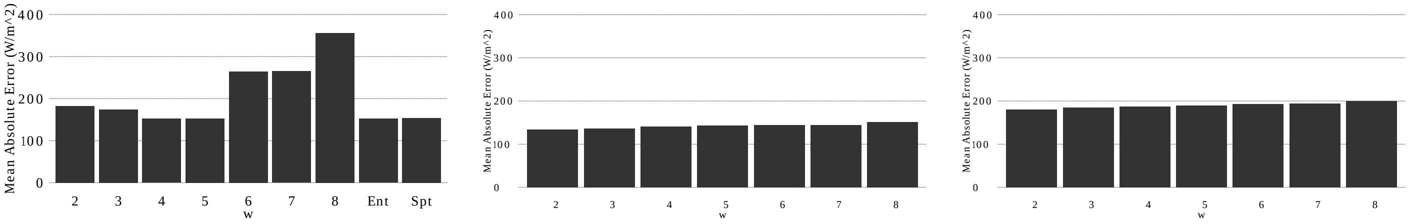


(a) Probabilistic Model

(b) Naive Bayes

(c) Linear Regression

Fig. 3. Mean absolute error of all forecast models for SCBH1 Station.



(a) Probabilistic Model

(b) Naive Bayes

(c) Linear Regression

Fig. 4. Mean absolute error of all forecast models for C0875 Station.

is certainly an option, we chose to discard missing data from our training and testing data set in order to avoid interpolated data from biasing our results.

Performance Measures. We measure the accuracy of our forecasting models by measuring the mean absolute error (MAE) of each daylight hour being predicted (0800-1700). Given a testing time series data set with hourly time points (daylight hours only) indexed from 0 to n, MAE is defined as

$$\text{MAE} = \frac{1}{n}\sum_{t=0}^{n} |s_t - \hat{s}_t|. \tag{8}$$

To evaluate the performance of the models, we split the dataset into a training data set (earlier years) and a testing data set (2014 data). The MAS is measured on the testing data set. We did not use cross validation because the use of earlier years for training and later years for test better mimic operational usage of such forecasts. Since the probabilistic models predicts the daily solar irradiance profile for the next day, the hourly solar irradiance in the profile is used to compute the MAE. Note that the maximum solar irradiance during the sunniest hour on the sunniest day is approximately 1200 W/m^2. In the following plots, "Ent" and "Spt" denote respectiviely the entropy and support method for choosing the best joint distribution for the probabilistic model approach.

A. Overall Performance

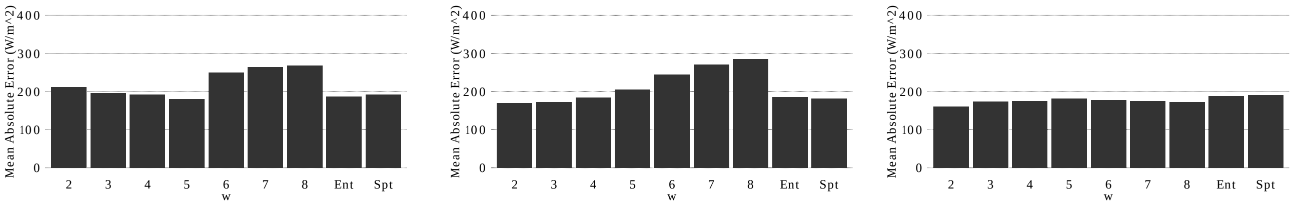
Figures 3 and 4 show the accuracy of the forecasts of the different models over different values of w. For SCBH1, observe that the probabilistic model with a fixed $w = 2$ yielded the most accurate forecast although Naive Bayes and linear regression are very close. The window size w did not seem to make much difference for each of the models. In fact, a bigger window (and hence more history) seems to make the forecasts slightly worse.

More interestingly, for C0875, Naive Bayes is the most accurate and the probabilistic models is significantly worse especially with $w = 6, 7, 8$. However, the entropy and support based methods for dynamically choosing the optimal w (and hence model) seem to provide much needed robustness, since their performance seems very close to the best models ($w = 4, 5$).

There are two possible reasons why the performance characteristics of the two stations could be so different. First, the weather patterns at the two stations could be so different as to favor different forecasting models. Second, the SCBH1 result uses 10 years of training data, while the C0875 result uses only 2 years worth of training data. Could that account of the difference ?

Varying the training set size. To shed light on the question of how training data set size affects forecasting performance, we measured the accuracy of the probabilistic models for the SCBH1 station for different training data set size $(2, 3, \ldots, 10$ years). In the interest of space, Fig. 5 only shows the results for 2, 3 and 10 years and we denote the shape of those 3 error plots as pattern A, B, and C respectively. Our analysis of the rest of the results (not shown) show a repeating error pattern between training data set size of 2-10 years: A,B,B,B,C,B,C,B,C. We are currently still investigating the reason for this repeating pattern. We conjecture that there may be a hidden weather-related variable (El Niño years?) that we have not accounted for. In any case, the error plots for the probabilistic models for SCBH1 with two years of training data does look similar in shape to that for C0875, which leads us to think that it is likely that the difference in the size of training data is likely to account for the difference between the results of the two stations.

A consequence of the above reasoning is that the Naive Bayes classifier is likely to be a better predictor when the training data set is small as compared to the more complex



(a) 2 years of training data (pattern A)

(b) 3 years of training data (pattern B)

(c) 10 years of training data (pattern C)

Fig. 5. Accuracy of probabilistic models using training data with different size.

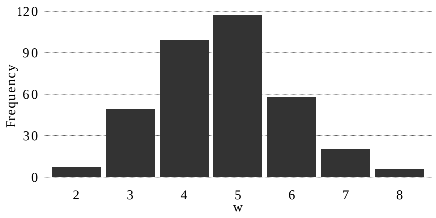


Fig. 6. Number of times the entropy method chose each of the w values/models for SCBH1 using two years of training data (2012 and 2013).

joint distributions. That does fit our mathematical intuition that since the Naive Bayes classifier only uses joint distributions of size $w = 2$, given a small training data set, it is likely to be able to have more support for those distributions than joint distributions where $w > 2$.

Robustness of the Entropy & Support methods. Observe in the errors for the probabilistic models on C0875 Fig. 4(a) and on SCBH1 Fig. 5(a) that while the entropy and support methods for dynamically choosing a w (and hence a joint distribution model) do not give the best accuracy, they consistently provide a forecast that is very close to the best fixed-w model. We plot in Fig. 6 the number of times the entropy method chooses a particular joint distribution (denoted by $w = 2, \ldots, 8$) for SCBH1. The plot shows that the entropy method tends to favor the model associated with $w = 4, 5$ which corresponds to the two most accurate models in Fig. 5(a).

B. Dependency on Other Weather Variables

Will including other (land-based) weather variables in our forecasting models help? We address that question in this section. We analyze the mutual information between the daily solar irradiance profile paired with various weather variables (including itself) and at varying lead times to determine the strength of the dependency. The weather variables under consideration (pressure, temperature, humidity, dew point, wind and precipitation) are chosen based on their widespread availability on most weather stations on Oahu where solar irradiance data are also available. In particular, we found that pressure data is only available with solar data at C0875.

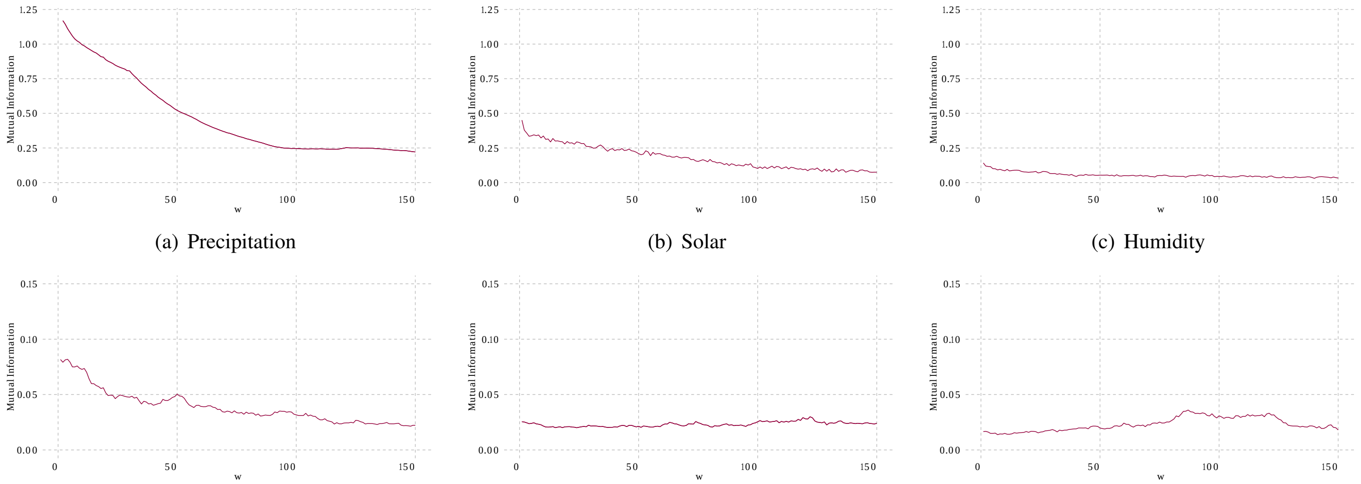
We calculate the mutual information between each weather variable time series and the solar radiation time series by varying the amount of days **w** between both time series. For the solar mutual information we use the same time series with varying values of **w**. Figures 7 shows the calculated mutual information for SCBH1 for 11 years of data (2003 to 2014). Observe that precipitation yields the strongest dependency with solar irradiance profiles across all weather variables, even stronger than its auto-dependence. From a meteorology point of view, it is not surprising since rain (precipitation) requires rain clouds which would block the sun's radiation.

We do need to be careful with generalizing the results to other stations on the island, because the Hawaiian islands are well-known for its wide variety of micro-climates. Even regions in close proximity may exhibit wildly different climate conditions! Fig. 8 plots the mutual information of solar irradiance profile with other variables for the C0875 station using two years of data. Unfortunately precipitation is not measured at C0875. We observe only very weak dependency with the other weather variables for C0875.

IV. Related Work

Current state-of-the-art solar irradiation forecasting models can be divided into two broad categories: *physical models* and *statistical models*. Physical models uses mathematical equations to describe the physics and dynamics of the atmosphere. High performance computing systems and numerical methods are then used to simulate the physical models forward in time for forecasting. Statistical models are typically based on extracting statistics from historical data in order to predict solar irradiance. Examples include neural networks [7], and auto-regressive models [5]. We can further categorize statistical models according to how they address (1) *time granularity* (e.g. hour, day), (2)*resolution and location* (e.g. Martin et al. [3]; Moreno et al. [4] uses global irradiance while Wang et al. [7] included land-based solar irradiation), and (3) *transformations* (e.g. derivative or meteorology based composition of variables related to solar radiation [7]).

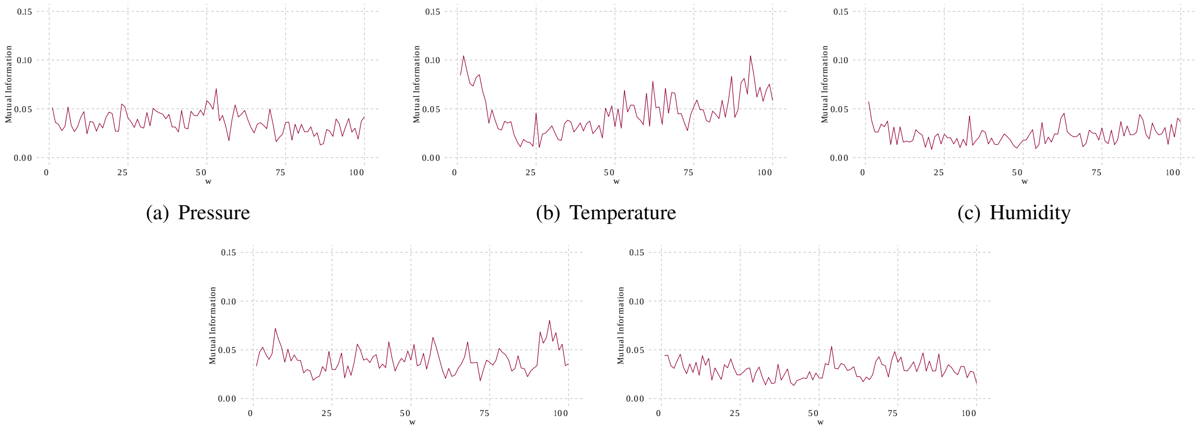
Our work explored one particular forecasting *time granularity*, one-day ahead, using **statistical models**. Previous work on solar forecasting for Oahu island, Hawai'i, focused on the minute level (short-term forecasting):[8] evaluated the performance of the lasso and other linear models for 5-min solar irradiance forecasting using land-based sensors;[1] focused on modeling weather patterns on Oahu island. It is also important to note that different levels of *time granularity* address very different questions and have different applications. For instance, [8] uses very short-term irradiance forecasting to manage the variability within a central PV power plant. Our work is more focused on ensuring uninterrupted energy supply for the *following* days instead of smaller fluctuations during the *same* day. Other concerns include identifying the best locations in terms of average solar irradiance on the island of Lanai in Hawaii [9], and normalizing solar radiation measurements in the island of Maui using clear-sky radiation models [2].



(a) Precipitation (b) Solar (c) Humidity

(d) Dewpoint (e) Wind (f) Temperature

Fig. 7. Mutual information of daily solar irradiance profiles paired with itself and other weather variables over different lead times (in days) for SCBH1.



(a) Pressure (b) Temperature (c) Humidity

(d) Dew Point (e) Wind

Fig. 8. Mutual information of daily solar irradiance profiles paired with itself and other weather variables over different lead times (in days) for C0875.

V. Conclusion

In this work, we evaluated the forecasting performance of our adaptations of the three models using land-based weather data from several weather stations on the island of Oahu in Hawaii. Different from previous work, we investigated solar forecasting at the daily level. We were able to observe consistent error patterns across different models conditioned on amount of training data provided, and verified weather variables can not so quickly generalize across different locations even within the same island distance. We plan to extend this work through a similar analysis done here in order to explore the performance of the models with these and other weather variables.

References

[1] L. M. Hinkelman. Differences between along-wind and cross-wind solar irradiance variability on small spatial scales. *Solar Energy*, 88:192 – 203, 2013.

[2] R. J. Longman, T. W. Giambelluca, and M. A. Nullet. Use of a clear-day solar radiation model to homogenize solar radiation measurements in hawaii. *Solar Energy*, 91:102 – 110, 2013.

[3] L. Martn, L. F. Zarzalejo, J. Polo, A. Navarro, R. Marchante, and M. Cony. Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning. *Solar Energy*, 84(10):1772 – 1781, 2010.

[4] A. Moreno-Munoz, J. De la Rosa, R. Posadillo, and V. Pallars. Short term forecasting of solar radiation. In *Industrial Electronics, 2008. ISIE 2008. IEEE International Symposium on*, pages 1537–1541, June 2008.

[5] G. Reikard. Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, 83(3):342 – 349, 2009.

[6] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson, 2006.

[7] F. Wang, Z. Mi, S. Su, and H. Zhao. Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters. *Energies*, pages 1355–1370, 2012.

[8] D. Yang, Z. Ye, L. H. I. Lim, and Z. Dong. Very short term irradiance forecasting using the lasso. *Solar Energy*, 114:314 – 326, 2015.

[9] A. Zagouras, H. T. Pedro, and C. F. Coimbra. Clustering the solar resource for grid management in island mode. *Solar Energy*, 110:507 – 518, 2014.