Hour-Ahead Solar PV Power Forecasting using SVR Based Approach

Abdullah Alfadda, Rajendra Adhikari, Murat Kuzlu, Saifur Rahman
Virginia Tech – Advanced Research Institute, Arlington, VA, 22203

***Abstract*—The use of solar photovoltaic (PV) in power generation has grown in the last decade. Unlike the traditional power generation methods (i.e. oil and gas), the solar output power is fluctuating and uncertain, mainly due to clouds movement and other weather factors. Therefore, in order to have a stable power grid, the electricity utilities need to forecast the solar output power, so they can prepare ahead adequately. In this work, hour-ahead solar PV power forecasting is performed using Support Vector Regression (SVR), Polynomial Regression and Lasso. The implemented regression models were tested under different feature selection schemes. These features include weather conditions (i.e. sky condition, temperature, etc.), power generated in the last few hours, day and time information. Based on the comparative results obtained, the SVR forecasting model outperforms the other two models in terms of accuracy.**

***Index Terms*—Solar Forecasting, Machine Learning, Support Vector Regression, Lasso.**

I. Introduction

According to Energy Information Administration (EIA) report released in 2014, renewable energy resources represent 11% of the total power generated in the U.S. [1]. The annual photovoltaic (PV) installations jumped from 4 MW per year in 2004 to 6201MW in 2014. Also, the share of the newly installed solar power generators increased from 10% in 2010 to 32% in 2014, when compared to the other energy resources, such as oil and gas. The use of solar PV in electric companies increased as well. One main drawback when using solar PV, is the fluctuation in its power output, when compared to the other conventional methods (i.e. gas and oil), this can be problematic for the operation of the power grid. This fluctuation in the output power is caused by many weather factors, but mainly due to the change in solar irradiance caused by clouds and due to seasonal variation owing to different sun positions. In order to plan for the avoidance of power outages and blackouts, electric utilities and transmission companies need to predict the variability in solar PV output power, so they can plan ahead adequately. Solar PV power forecasting is very beneficial to these organizations, since it will allow them to adjust the plant output power ahead of time, so that they can optimize the overall performance of grid. Solar PV power prediction methods in the literature can be classified into different categories. In terms of the time horizon, it could be either a short-term forecasting domain or a long-term forecasting domain. The short-term forecasting domain includes predictions for day-ahead, hour-ahead or intra-hour, while the long-term forecasting domain includes predictions for the next several days, weeks, months or years. In terms of predicted forecast output, the forecasting can be categorized into deterministic or probabilistic forecasting methods. In deterministic forecasting a fixed number is provided for the output power, while in probabilistic forecasting methods the prediction is produced in terms of probability distribution of expected output power. In terms of geographic locations used, it can be categorized into point forecasting techniques, which are used for predicting output power for a particular geographic location of a single solar PV power plant, while area forecasts are concerned with the aggregated power output of a large geographically dispersed set of solar PV power plants [2, 3]. In this paper our focus is deterministic, short-term hour-ahead, point forecasting of solar PV panels installed in our lab facility at Arlington, Virginia, USA.

Solar PV power forecasting has become an active area of research in the last decade, many work have been conducted in this field recently. In [4] the relation of the solar irradiance variability between the neighboring sites as a function of their distance has been studied. Solar irradiance forecasting based on cloud motion prediction has been shown in [5]. A new probabilistic model approach implemented to forecast the solar irradiance is shown here [6], where the expected output power will fall into a range with confidence interval applied in a very short-term. In [7] authors used the irradiance, cloud cover and ambient temperature data and applied k-Nearest Neighbors (k-NN) and SVM to perform the hour-ahead solar power forecasting. In [8] a method to predict day ahead solar power using only historical data is presented, that is based on k-NN and neural networks approaches. In [9] Weighted Support Vector Machine (WSVM) is implemented by categorizing similar days (i.e., days that have the similar features) into different classes on a short term scenario. Comprehensive reviews and comparisons between different solar forecasting techniques such as Artificial Neural Networks (ANN), Fuzzy Logic and Expert Systems could be found in [10-13]. The contribution of this paper is that we propose a new set of features and their extraction methods, which will be used as an input to the prediction model, such that it will improve the hour-head solar power prediction. Moreover, the performance of different prediction methods are compared and based on the obtained results we see that the use of Support Vector Regression (SVR) outperforms the other implemented methods, when these new set of features are applied.

Our SVR model is compared with three different regression algorithms: Linear Regression, Quadratic regression and Lasso. The results have been verified by comparing the set of features applied in this work with the most popular set of features used in solar PV forecasting. It is found that some of the features just act as noise and will over-fit the SVR prediction model, so excluding these features will improve the overall accuracy of the model.

II. Background

In this section, a brief overview about the implemented regression models will be introduced. In the first section, it will be shown that how the SVR can be applied to solve the proposed regression problem. The second section discusses the application of Lasso in this regression problem. Last section, introduces the linear regression model.

A. *Support Vector Regression*

The basic structure of Support Vector Machines (SVM) can be slightly modified to directly solve the regression problem. For the given data points (x, y), a linear decision function $f(x) = wx + b$ can be trained using a linear regression technique, where w is a weight vector and b is a bias term. The objective here is to find the flattest fit, which will lead to the following optimization problem:

$$\begin{aligned}&\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}(\xi_i + \xi_i^*)\\ &\text{s.t.}\\ &y_i - f(x_i, w) \leqslant \epsilon + \xi_i^*\\ &f(x_i, w) - y_i \leqslant \epsilon + \xi_i\\ &\xi_i, \xi_i^* \geq 0, i = 1, ..., L\end{aligned} \tag{1}$$

Where user selectable ϵ represents the tolerance boundary for the linear fit. After the optimum solution for w and b is found, the function f can be used to predict output y for any future input x. Here, the constrained has been relaxed to allow data points to lie outside of the tolerance range ϵ, but by including the losses associated with data points lying outside the tolerance range into the objective function as $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_i^*$. L is the total number of data points, C is a tradeoff between outlying data amount and flatness of f, we effectively try here to minimize such loss to have the optimal fit.

B. *Lasso Regression*

The other implemented model is the Lasso, which is a method of excluding unnecessary parameters, hence avoid overfitting. Mathematically, it is just a Lasso constraint applied to the ordinary least squares linear regression, as shown below:

$$\begin{aligned}&\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\\ &\text{s.t.}\\ &\sum_{j=1}^{p}|\beta_j| \leq t\end{aligned} \tag{2}$$

This constraint can be removed by rewriting the equation in the following equivalent form:

$$\min_{\beta}\left(\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right) \tag{3}$$

We will be solving this equation to find the model parameters β. Where λ is a regularization parameter, p is the vector length of x, and n is the number of training data points.

C. *Linear Regression*

Linear Regression is a widely used technique in time series prediction, so it was our first attempt when trying to forecast the output power. The linear regression equation is shown below:

$$\hat{Y}_t = b_0 + b_1X_{1t} + b_2X_{2t} + ... + b_kX_{kt} \tag{4}$$

Where $\hat{Y}_t$ is the predicted output power at time t, X_{kt} is the k input value (i.e. features) at time t and k is the total number of features implemented.

III. Dataset

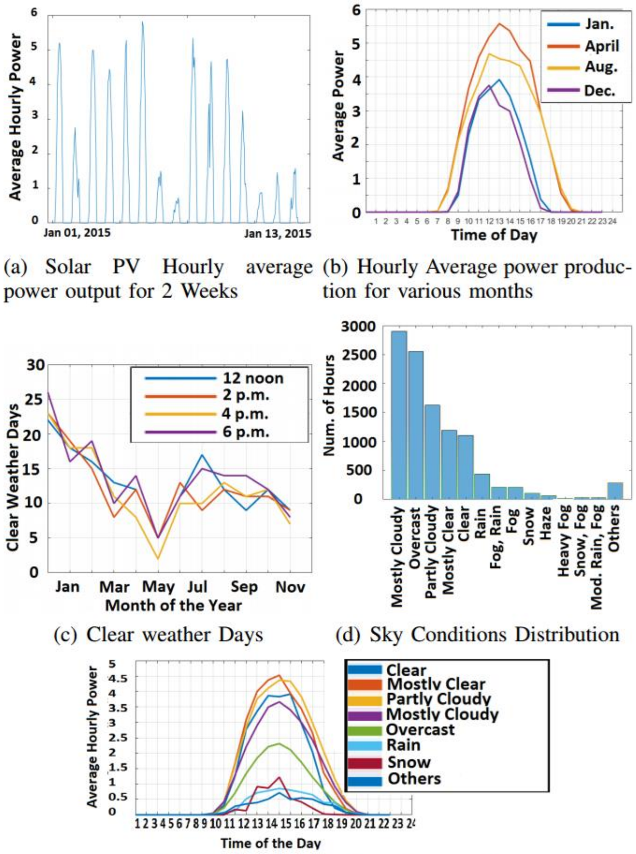
The dataset used in this work was collected from solar PV power station, located at the roof of Virginia Tech Research Center (VTRC), in Arlington, VA. The data were collected over almost 15 months, starting from January 1, 2015 to March 26, 2016. The data are collected every 15 seconds, but it have been averaged to every hour. The Data have the following information:

- Weather Temperature.
- Solar Irradiance.
- Module Temperature.
- Wind Velocity.
- Output Power.

Besides this dataset, we have collected the weather data for the Arlington area, from the MesoWest archive using the weather station in the nearby Ronald Reagan Washington National Airport (IOCA Code: KDCA). The data are available every hour, and consist of Temperature, Dew Point, Relative Humidity, Visibility, Wind Speed, Wind Direction and Cloud Cover. We have combined the two datasets (solar PV power output and weather data) together, to have our raw dataset as shown in Table I.

IV. Data Analysis

To have a general idea about the nature of the dataset, and to visually inspect if the dataset looks reasonable, we have analyzed the power generated by the solar PV for different hours of the day, different months and different cloud cover conditions. Fig. 1(a) shows the hourly average solar power production in the first two weeks of January 2015. We can see that the output power has a high variation, this is because the different weather conditions (i.e., sunny, cloudy, etc.). Also, to verify that our collected data match the solar output power, we plotted that output power for selected months, and as shown in Fig. 1(b) the generated power peaks around April, which makes sense for our solar panel located at 38 degree north latitude and oriented to south (22-degree-southeast), that is justified, since the sun position in April will have a maximum irradiance at this location, when compared to the other months shown in the Fig. 1(b).



(a) Solar PV Hourly average power output for 2 Weeks

(b) Hourly Average power production for various months

(c) Clear weather Days

(d) Sky Conditions Distribution

(e) Hourly power production for various cloud cover conditions

Figure 1. Data analysis

Fig. 1(c) and Fig. 1(d) give sense to the distribution of various cloud cover conditions. Fig. 1(c) shows the count distribution for clear weather days in different months and different hours of the day. Fig. 1(d) shows that there was almost 3000 hours of mostly cloudy sky and only around 1000 hours of clear sky condition during the study period. It is clear that there is prominent variation in clear weather hours from month to month (and less so between hours of the day). It is worth noting that clear weather hours are much less in the summer months. This finally leads us to Fig. 1(e), which shows the variation in power production based on different cloud cover types. It should be understood that the power production is averaged over all months. Therefore, the monthly variation in cloud cover type Fig. 1(b) and the monthly variation in 'Clear' weather power Fig. 1(c) will have significant impact on the nature of this graph. Quite notably, the average power produced during 'mostly clear' and 'partially cloudy' sky conditions, ironically, looks better than 'Clear' sky condition. However, this can be explained based on the fact that 'Clear' sky condition occurred mostly during lower power production months, while 'Mostly Clear' and 'Partially Cloudy' conditions occurred more during high power production months.

Table I. RAW DATASET DETAILS

Column 1	**Column 2**	**Column 3**	**Column 4**	**Column 5**
Month and Hour of Day	Temperature	Relative Humidity	Visibility	Wind Speed
Column 6	**Column 7**	**Column 8**		**Column 9**
Wind Direction	Cloud Cover	Past n-Hours Power		Output Power

V. APPROACH

A. Feature Transformation

Cloud cover is the text description of the sky condition (Cloudy, Clear, Overcast, etc.). Since the cloud cover data are in text form, and in order to use it for regression, we need to convert it into a numerical form. Each sky condition type has been given a numerical value based on the historical data; for each hour of the day for each month, a score has been assigned to each cloud cover type, which is equal to the average power produced at that hour in that month during that cloud cover condition. If no such cloud cover condition exists for a given month or a given hour, no score is assigned for that month or hour. Finally, all the scores for various hours of the day and various months were averaged out to generate a single score for a given cloud cover type, the score assignment is shown below:

$$\sigma(i) = \frac{1}{N_i}\sum_{d=1}^{N_i}\sum_{h=6}^{21} Power(d,h) * I(d,h,i) \tag{5}$$

Where

$$N_i = \sum_{d=1}^{365}\sum_{h=6}^{21} I(d,h,i) \tag{6}$$

$I(d,h,i)$ is an indicator function, and will have score of one if the sky condition is i for a the day d and hour h, and zero otherwise. N_i is a counter for these occurrences for the cloud condition i.

It is needed now to transform the time of day, and time of year information contained in the date-time stamp, such that it will incorporate the knowledge that hours wrap around at 24 and goes back to 0, and the day of year wraps around at 365 (or 366) and goes back to 1. The transformation must be such that, day 1 and day 365 lie close each other. Similarly, hour 0 and hour 23 should lie close to each other. For this, we performed a polar coordinate transformation of the hour of day and day of year data, so that they lie on the edge of circle. This transformation correctly incorporates the periodicity of the data, this transformation is shown below:

$$day\ of\ year\ x = cos(day\ of\ year * 360/366) \quad (7)$$

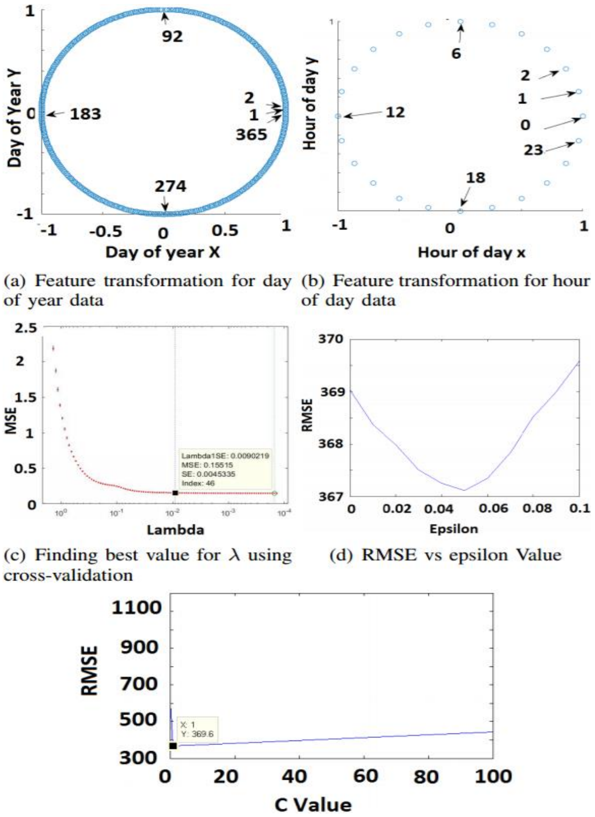
$$day\ of\ year\ y = sin(day\ of\ year * 360/366) \quad (8)$$

$$hour\ of\ day\ x = cos(hour\ of\ day * 360/24) \quad (9)$$

$$hour\ of\ day\ y = sin(hour\ of\ day * 360/24) \quad (10)$$

After this transformation, the day of year will lie in a circle as shown in Fig. 2(a), and the hours of the day will also come to lie in a circle as shown in Fig. 2(b).

The other features are directly used without modification. Lastly we append the power produced on three previous hours, and three previous days at the same hour as a feature variable. Hence, our final dataset looks like as shown in Table II, where column 1 to 17 constitute the data matrix X, and the column 18 represents the data matrix Y.



(a) Feature transformation for day of year data (b) Feature transformation for hour of day data

(c) Finding best value for λ using cross-validation (d) RMSE vs epsilon Value

Figure 2. Feature transformation and variables optimization

B. Validation

In order to select best parameters for the regression model and in order to prevent over fitting, the dataset was split into two primary sets: training and testing set. We extracted out 20% of our data as a test data, while the remaining 80% was used as a training data. The test data were not used anytime during the training phase, except for finally evaluating the performance (error) of the applied models. For SVR and Lasso, the training data set were used to cross-validate the model and tune the different models parameters. For this, a 10-fold cross-validation was implemented, where the training set were split into 10 equal folds, and in each turn a fold is taken out as validation set, while the rest 9 folds are used as a training set. After the training is done on the 9 folds, the taken out fold is used for prediction and calculating the prediction accuracy based on Root Mean Square Error (RMSE). This process is repeated for all of the 10-fold, and average accuracy is calculated, which is called the 10-fold validation accuracy. Parameters can be tuned until the validation accuracy is improved to the optimum level. Once the parameters are optimized, a final testing can be performed on the testing data set to evaluate the overall prediction accuracy. The error measure for regression can be performed using many ways, such as RMSE, Mean Squared Error (MSE) or Standard Error of Estimate (SEE), etc. In this work, calculating the error in all the models, including the polynomial regression, SVR and Lasso, was performed using the same error measure, the RMSE as shown in (11):

$$RMSE = \sqrt{\frac{\sum(\hat{Y}_t - Y_t)^2}{n}} \quad (11)$$

Where Y_t is the actual data, and $\hat{Y}_t$ is the prediction, and n is the total number of testing data points.

C. Implementation

A QR decomposition method was applied on the least squares model to calculate the regression weights for the linear and quadratic regression models. The Lasso regression is implemented with a coordinate descent algorithm as a solver, and in order to pick the correct value of the parameter λ, 10 fold cross-validation was implemented, as shown in Fig. 2(c). Finally for the SVR the parameters C and ϵ were tuned by cross-validating the training set with respect to ϵ value, then selecting the ϵ which minimizes the RMSE. This can be seen in Fig. 2(d). The C value for the SVR was chosen using similar cross-validation method, as shown in Fig. 2(e). The C value that minimizes the RMSE is equal to 1.

Table II. DATASET

Column 1	Column 2	Column 3	Column 4	Column 6
Day of Year x	Day of Year y	Hour of Day x	Hour of Day y	Temperature
Column 6	Column 7	Column 8	Column 9	Column 10
Dew Point	Humidity	Visibility	Wind Speed	Wind Direction
Column 11	Column 12	Column 13	Column 14	Column 15
Sky Condition	Power 1 Hour earlier	Power 2 Hours earlier	Power 3 Hours earlier	Power Generated 1 Day before at the Same Hour
Column 16		Column 17		Column 18
Power Generated 2 Days Before at the Same Hour		Power Generated 3 Days Before at the Same Hour		Current Power

The goal of this work is to compare multiple regression models under different feature selection schemes. In order to compare the results of these different feature selection schemes, an incremental study is performed by progressively increasing the number of features used to feed the prediction models. In this regard, the feature selection scheme starts by training the model under two simple features, the sky-condition and the power generated one-hour before, after that the features are being added on a cumulative approach, while the behavior of the RMSE is being measured under these different models. This can be seen in Table III, where the first column refers to the features used for the model in cumulative order. In all models, the cloud cover data was always included. Therefore, for example, the third row consists of models using the following features: cloud cover data, power one-hour before, power two-hours before, power three-hours before.

VI. Results

The RMSE results on the test data for various regression models are shown in Table III. As can be seen, different models have different RMSE for different feature selection schemes. Out of all the models, the SVR model, using only a subset of features (cloud cover, 3 hours earlier and 3 days before power, hour of the day and day of the year) outperforms all the other implemented models and feature selection schemes. The other models have their best error performance when all the features are included. Fig. 3 shows the plot of actual test data points power versus the power predicted when applied under the SVR model.

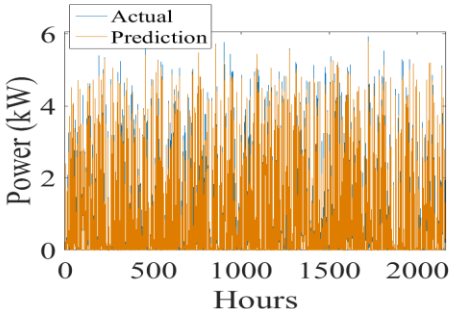


Figure 3. SVR regression results

Our best model performance has an RMSE of 318.4 watts, which equates to 5.3% in error for our 6000 watts capacity solar PV. By comparison, the best model presented in [14] achieved 7.9% in error when applied on the 1-hour time horizon. An improvement of 2.6% in the overall model accuracy has been achieved using the SVR model with the feature selection methods proposed in this paper.

VII. Conclusion

In this study a comparison has been conducted between different regression models to assess their performance under different feature selection schemes. Linear and nonlinear regression, SVR and the Lasso regression all have been implemented and tested. The SVR model outperforms the other regression models when correct set of features are selected. However, if more features are used than necessary, which includes the temperature, humidity, visibility, wind-speed, wind-direction, SVR accuracy tends to degrade due to over-fitting the model with the noise-like data. Interestingly, the hour of the day and day of the year had minor effect on the model performance, while as seen in the results, the power produced few hours ago and few days ago play a crucial role in model accuracy when used alongside the sky condition data.

Table III. RMSE UNDER DIFFERENT REGRESSION MODELS AND FEATURES

Model	Linear Regression	Quadratic Regression	SVR	Lasso
1-Hour Before	567.6	566.2	595.9	578.3
2-Hours Before	448.2	431.5	399.5	458.6
3-Hours Before	433.4	414.8	371.7	436.7
1-Day Before	415.9	381.4	358.7	429.8
2-Days Before	410.8	379.2	347.5	433.1
3-Days Before	409.2	377.1	353.6	435.9
Hour of Day	395	334.4	330.2	411.6
Day of the Year	393.2	326.4	318.4	403.4
Temperature	391.1	325	369.6	428.6
All Features	385.1	319.9	1272.8	393.1

References

[1] Energy Information Administration, “Annual energu Outlook 2014”, 2014.

[2] S. Letendre, M. Makhyoun, and M. Taylor, “Predicting solar power production: irradiance forecasting models, applications and future prospects,” Solar Electric Power Association, Tech. Rep, 2014.

[3] Pelland, Sophie, et al. "Photovoltaic and solar forecasting: state of the art." *IEA PVPS, Task* 14 (2013): 1-36.

[4] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, and T. E. Hoff, “Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance,” Solar Energy, vol. 86, no. 8, pp. 2170–2176, 2012.

[5] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, and B. Washom, “Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed,” Solar Energy, vol. 85, no. 11, pp. 2881–2893, 2011.

[6] F. Golestaneh, P. Pinson, and H. Gooi, “Very short-term nonparametric probabilistic forecasting of renewable energy generation - with application to solar energy,” IEEE Transactions on Power Systems, 2015.

[7] B. Wolff, E. Lorenz, and O. Kramer, “Statistical learning for short-term photovoltaic power predictions,” in Proceedings of DARE 2013 Workshop on Data Analytics for Renewable Energy Integration, September, vol. 23, 2013.

[8] Y. Wang, B. Wang, R. Huang, C.-C. Chu, H. R. Pota, and R. Gadh, “Two-tier prediction of solar power generation with limited sensing resource,” arXiv preprint arXiv:1508.02669, 2015.

[9] R. Xu, H. Chen, and X. Sun, “Short-term photovoltaic power forecasting with weighted support vector machine,” in Automation and Logistics (ICAL), 2012 IEEE International Conference on. IEEE, 2012, pp. 248–253

[10] A. Mellit and S. A. Kalogirou, “Artificial intelligence techniques for photovoltaic applications: A review,” Progress in energy and combustion science, vol. 34, no. 5, pp. 574–632, 2008.

[11] K. D. Orwig, M. L. Ahlstrom, V. Banunarayanan, J. Sharp, J. M. Wilczak, J. Freedman, S. E. Haupt, J. Cline, O. Bartholomy, H. F. Hamann et al., “Recent trends in variable generation forecasting and its value to the power system,” IEEE Transactions on Sustainable Energy, vol. 6, no. 3, pp. 924–933, 2015.

[12] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. Martinez-de Pison, and F. Antonanzas-Torres, “Review of photovoltaic power forecasting,” Solar Energy, vol. 136, pp. 78–111, 2016.

[13] M. Q. Raza, M. Nadarajah, and C. Ekanayake, “On recent advances in pv output power forecast,” Solar Energy, vol. 136, pp. 125–144, 2016.

[14] V. P. A. Lonij, et al. "Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors." *Solar energy* 97 (2013): 58-66.