Short-Term Power Forecasting of Solar PV Systems Using Machine Learning Techniques

Mayukh Samanta
Bharath K. Srikanth
Jayesh B. Yerrapragada

Abstract

Roof-top mounted solar photovoltaic (PV) systems are becoming an increasingly popular means of incorporating clean energy into the consumption profile of residential users. Electric utilities often allow the inter-connection of such systems to the grid, compensating system owners for electricity production. As the systems grow in number and their contribution to the overall load profile becomes increasingly significant, it becomes imperative for utilities to accurately account for them while planning and forecasting generation. Additionally, this information is useful for system-owners who want to optimize their production schedule. We use various machine learning and statistical techniques to train models on solar irradiance data and different meteorological parameters to forecast solar irradiance, and therefore power, for different forecasting horizons in the short-term future. We begin with a series of naïve models – linear regression, locally-weighted linear regression and support-vector regression (SVR) – where we only use meteorological information in the present to make future predictions, and then progress to time-series modeling. We find that both linear and locally-weighted linear regression perform rather poorly in the naïve case; we consider only SVR (both the regular and least-squares variation) along with conventional statistical models such as the seasonal auto-regression integrated moving-average (ARIMA) model in the time-series implementation. Our best results, with an RMSE of 40.16 W/m^2, are obtained from using least-squares SVR (LS-SVR) with an RBF kernel, trained on solar irradiance data and meteorological features in the 7 hours prior to the present time t, and from 24 and 48 hours prior to the forecasting time $t + t_f$, where t_f is the forecasting horizon. This model performs better than existing models in literature which use the same dataset.

I. Introduction

Renewable energy accounted for 23% of all electricity generation worldwide as of 2013, with solar PV representing only about 0.85% of the global electricity demand [1]. However, in terms of the proportion of newly installed capacity, solar PV ranked the highest in the world overall among all forms of electricity generation in 2013. In the US, for example, the installed solar PV capacity is forecasted to be 6.6 GW in 2014, more than three times the size it was three years ago [2]. Thus, with a rapid decrease in manufacturing and production costs, solar PV is expected to make a significant contribution to meeting the world's energy demands, and the need for systems-level planning and design is increasingly becoming more critical.

Inter-connection is the process by which PV systems are linked to the electricity grid, and system owners are effectively compensated by utilities for sending power back into the grid. Utilities have complex planning and scheduling procedures in place for electricity generation and pricing from their forecasted estimates of load curves; however, they do not take into account the contribution of distributed generation sources such as roof-top solar PV. This makes accurate forecasts for these systems particularly important, as electricity generation at the utility-scale involves planning at least a few days in advance. Since electricity exhibits variable pricing and utilities typically allow only a capped amount of power to be sent back into the grid, these forecasts would be helpful to a system owner that wanted to maximize their compensation as well. A simplified economic interpretation is that the availability of more information on both the supply and demand sides leads to a reduction in incomplete information, and therefore increased economic efficiency. The importance of accurate forecasting is clear, and this serves as our primary motivation in choosing this problem.

There is a significant amount of literature on short-term power forecasting using various statistical and machine learning methods. Sharma et al. [3] argue that the generation profile of PV systems is heavily dependent on local, site-specific conditions. The use of analytical models to describe the system becomes difficult because the factors that determine solar irradiance, and consequently electric power, vary greatly from one location to another. For instance, incorporating the effect of the idiosyncracies of a particular PV system, such as shading and local weather conditions becomes much harder in an analytical model. This is our motivation for using machine-learning – an automated approach, if implemented correctly, accounts for these local conditions and makes predictions with better accuracy.

Diagne et al. [4] succintly review the various classes of methods that have been employed for irradiance forecasting. The most basic model is the persistence forecast, which is a naive predictor that assumes that the best predictor of solar irradiance at time $t + 1$, X_{t+1}, is the irradiance at time t, X_t. This is the baseline comparison used in the work of Perez et al. [5]. Sharma et. al report results using simple linear regression and SVM models [3], and our work first attempts to primarily build on this. Other interesting ML/AI approaches include the use of neural networks, as in the work of Lauret et al.[6], who use a multi-layer peceptron structure model. More detailed neural network models include the work of Kemmoku et al. [7], Mihalakakou et al. [8], Sfetsos and Coonick [9], Fatih et al. [10] among others, incorporating aspects such as multi-stage to time-delay neural network models. Some approaches attempt to use cloud-cover and satellite imagery data as well, such as the work of Chow et al. [11], which attempts to predict intra-hour, sub-kilometer forecasting of cloud cover using ground-based sky images. Hybrid approaches which attempt to utilize the best aspects of different machine learning approaches have been developed as well [12]. We adopt some of the insights from prior work into our own modeling approach, which is described in the next section.

II. Modeling the Problem

A. Data Set

The dataset used in this work is historical weather data from Amherst, MA, and is maintained by the University of Massachusetts, Amherst – Computer Science Weather Station. This data is from the same source used by Sharma et al., except that we use more a recent dataset.

The data contains solar irradiance values and meteorological observations obtained from a period of January to July 2013 at five minute intervals. We have averaged over this dataset to obtain hourly

data- the resulting dataset contains 5070 rows. The data collected has different parameters, such as timestamp, temperature, wind chill, heat index, humidity, dew point, wind speed, maximum wind speed, wind direction, rain, barometer pressure, extra-terrestrial irradiance and UV irradiance. From this, we have eliminated features that have large amounts of missing data or seemed impractical to use, such as UV irradiance, extraterrestrial irradiance and wind direction.

B. Models Used

In this work, we train a variety of models on our data and determine which features and models enable us to predict solar irradiance better. Our models are divided into two broad categories: naive methods and time-series modeling methods. The key difference between the two classes of methods was the incorporation of historical solar irradiance values as features or parameters in the time-series methods, while these values were absent in the naive methods. In this work, the naive methods are discussed further in Section 3, and the time-series methods in Section 4.

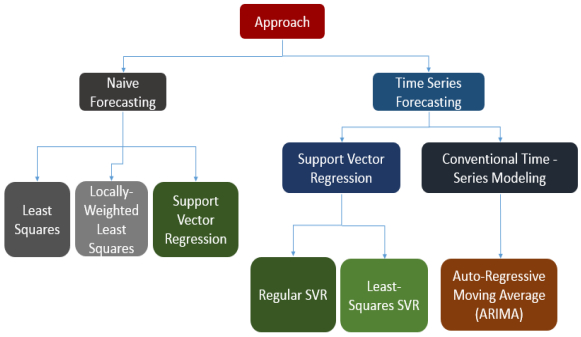

Figure 1. Different models and techniques used

C. Nature of Training and Test Sets

Our training set, unless specified, contains 3600 observations from the dataset (5 months worth of data). Our test sets contain 720 observations (one month worth of data). These training and test sets were selected from a random permutation of the dataset. All error values i.e. root mean squared error (RMSE) and mean absolute error (MAE), have been reported on these random training and test sets. For all of our models, in addition to the training and test sets, we also evaluated their performance on a test set of 30 consecutive days (720 hours). This was interesting as it was a good respresentation of an actual use case for this problem and it allowed us to compare the performance of the naive and time series models that we used. It also allowed us to see how closely the model tracked hourly solar values across a fixed period of time.

III. Naive Methods

Before using any of our models, we experimented with different dimensionality reduction techniques, such as manually removing dataset features that had minimal correlation with solar irradiance, and using Principal Component Analysis [13] to reduce the dimensionality of the feature set. We found that PCA improved the performance of our models significantly. Prior to applying PCA, we preprocessed the feature set so that it had zero mean and unit variance. This result is discussed in more detail in Section III C.

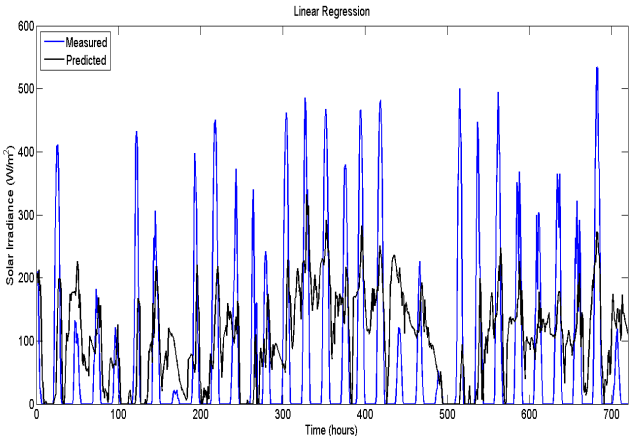

Figure 2. Linear Regression prediction on a test set of 30 consecutive days

A. Linear Least Squares Regression

Linear least-squares regression is among the simplest algorithms that exist to verify the relationship between a dependent variable and a set of independent variables. Given a training feature set X and a corresponding solar irradiance vector y, the trained model θ is obtained from the equation:

$$\theta = (X^T X)^{-1} X^T Y \tag{1}$$

Linear regression performed poorly on the randomly generated sets we used for testing. We found that the model's performance improved by pre-processing the dataset, but it still had a large error in its predictions with a training and test set RMSE of 194.08 W/m^2 and 192.79 W/m^2 respectively. On the test set of 30 consecutive days too, we found that linear regression did not perform too well as can be seen in Figure 2.

B. Locally weighted Linear Regression

Non-parameterized algorithms are an interesting class of algorithms to use on our prediction problem, given that we have a sufficiently large training set. Locally weighted linear regression is a non-parameterized algorithm that is a more specialized form of linear regression, in which the linear coefficients of the model at a particular point are affected to a greater extent by neighboring training examples, and to a far less extent by distant training examples. We used a fairly standard equation to compute the weights of a sample, given by

$$w^i = exp(-(x^i - x)^T (x^i - x)/2\tau^2) \tag{2}$$

and computed the training model using the normal equation

$$\theta = (X^T W X)^{-1} X^T W Y \tag{3}$$

For this model, we had to vary the bandwidth parameter τ, and we found empirically that a value of 20 gave us good performance. Interestingly, for a randomly permuted training and test set, using locally weighted linear regression did not improve the model's performance significantly. With a training and test set RMSE of 193.63 W/m^2 and 192.34 W/m^2 , we felt that this model was not much better than the linear regression model, especially when its intensive computational nature is taken into account. On the test set of 30 consecutive days too, the locally weighted regression model was very similar in performance to the linear regression model.

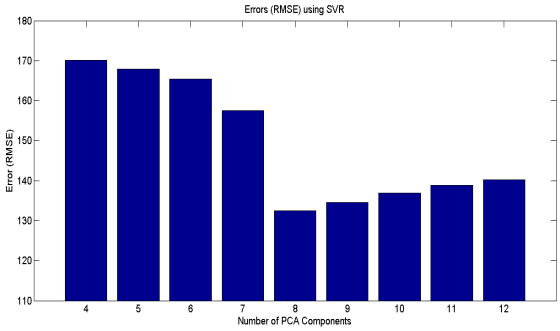

Figure 3. Performance of the SVR Model with varying number of principal components

C. Support Vector Regression

Support Vector Regression (SVR) constructs a hyperplane (or a set of hyperplanes) to perform regression on high-dimensional data. The objective of the SVR model is

$$minimize \frac{1}{2} \|w\|^2 + C \sum (\xi_i - \xi_i^*) \tag{4}$$

subject to

$$y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i$$

$$\langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

The SVR tool we used for this purpose comes from the LibSVM library. The library contains a variety of tools for both classification and regression purposes, and the particular tool we selected for SVR was the epsilon-SVR tool. This tool makes use of two free parameters: C, which represents the penalty parameter of the error term (default = 1); and ϵ (default = 0.1), which represents the value for which no penalty is imposed to the training loss function as long as the predicted values are within a distance ϵ of the actual value of the training examples.

We used three different kernels to test this model - the linear, polynomial and Radial Basis Function (RBF) kernels. We found that the RBF Kernel performed the best, with a train and test RMSE of 130.47 W/m^2 and 140.24 W/m^2. These results were obtained using the original dataset with basic pre-processing. We believe the good performance of the RBF Kernel can be attributed to the way it maps the input dataset to a higher dimensional space, thus capturing non-linear relationships between the feature set and the solar irradiance values. Also, to achieve the performance that we mentioned, we used the value $C = 2000$ and the default setting for ϵ from LibSVM, as we found that the model gave us fairly poor performance with the default value for the parameter of C.

We found PCA to be very effective in improving the perfomance of the SVR model. On varying the number of principal components, we found the model performed best with 8 principal components, with a training and test RMSE of 118.31 W/m^2 and 132.53 W/m^2. The performance of the model with different numbers of principal components can be seen in Figure 2.

We found that the SVR model we described performed reasonably well on the test set of 30 consecutive days, with an RMSE of 82.66 W/m^2.

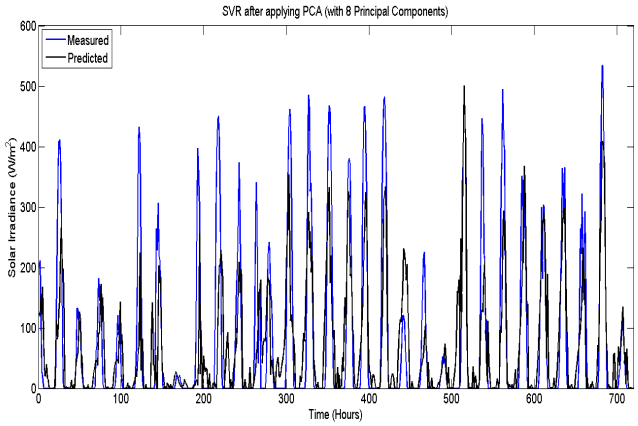

Figure 4. SVR predictions over a test set of 30 consecutive days

IV. Time-Series Modeling

From the implementation of our previous models and the variation of solar irradiance across consecutive hours, we observed that the data series has a fairly salient time-series aspect. That is, the solar irradiance peaks at a particular time in the day, and hence has a fairly seasonal trend. Thus, it makes sense to model solar irradiance as not just a causal relationship with the meteorological feature set, but also as a time-series. We experimented with different models, namely the seasonal auto-regression integrated moving-average (ARIMA) model and the least-squares Support Vector Regression (LS-SVR) to model the time-series aspect of this dataset.

A. ARIMA

ARIMA models are the most general class of models for time-series forecasting [14]. They are composed of autoregressive terms, non-seasonal differences and lagged forecast errors. Basic ARIMA models are non-seasonal and work best with stationary time series. The model is generally referred to as an ARIMA(p, d, q) model where parameters p, d, and q are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. An alternative to the non-seasonal ARIMA model is to use the seasonal ARIMA model, which is denoted as ARIMA(p, d, q) (P, D, Q). A seasonal ARIMA model could consist of seasonal autoregressive (AR) and moving average (MA) terms, and this model can be represented as

$$(1 - L)^{\xi}(1 - L^f)^{\zeta} ln Y_t \left[\frac{\theta_t(L)\Theta_t(L)}{\phi_t(L)\Phi_t(L)} \right] \epsilon_t \tag{5}$$

where $\phi(L)$ is the auto-regressive operator

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p \tag{6}$$

and $\Phi(L)$ is the seasonal auto-regressive operator. $\theta(L)$ is the moving average operator, given as

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + ... + \theta_p L^p \tag{7}$$

and $\theta(L)$ is the seasonal auto-regressive operator. In addition to these, ξ denotes differencing, ζ denotes seasonal differencing and f denotes the cyclical frequency of the model. Previous studies have used an ARIMA$(1, 0, 0)(1, 1, 0)$ model to build a simple forecasting model [15]. To use this technique for solar irradiance data, we used the same model specifications with $f = 24$, to predict hourly data

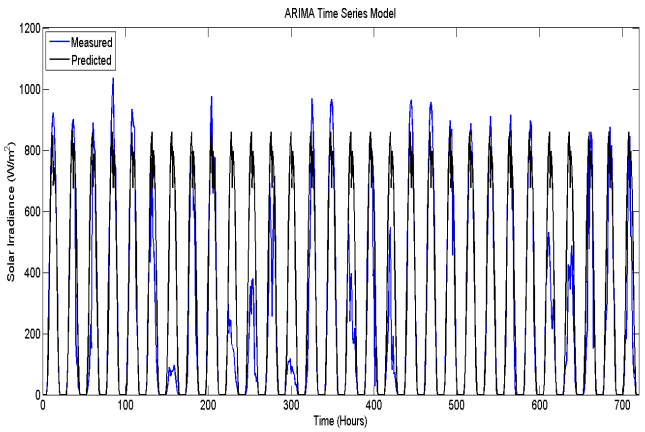

Figure 5. ARIMA predictions over a test set of 30 consecutive days

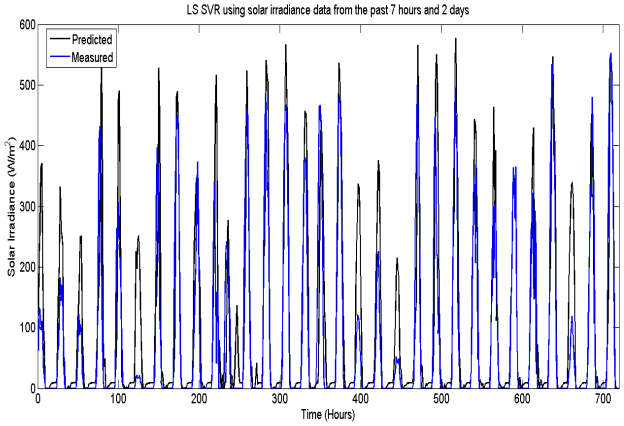

Figure 6. Time Series LS-SVR predictions on a test set of 30 consecutive days

that was seasonal at the 24 hour horizon. We found that this model was excellent at capturing seasonal trends of peaks in the data, i.e. the model was able to predict the occurrence of peaks in solar irradiance in a day very well. However, it was unable to capture day-specific intensity values, as it had no information regarding any of the meteorological factors that are correlated to the magnitude of irradiance peaks a day. Thus the ARIMA model performed poorly as compared to the basic SVR model in terms of error, with an RMSE of 188.32 W/m^2. Subsequently, as an extension to the ARIMA model, we also experimented with providing the model with causal inputs, i.e. the meteorological feature dataset. However, this performed poorly too, as we found that providing causal inputs did not damp the original model's seasonal prediction enough to significantly improve performance.

B. Least Squares Support Vector Regression

Least Squares SVR is a support vector model that is similar in nature to Vapnik's formulation of SVR (from Section III C). One key difference between the Least Squares SVR model and the conventional SVR is that the LS-SVR solves a system of linear equations as opposed to the optimization problem described in Section III C. The least-squares SVR is formulated to minimize the following cost function [16]

$$minimize\, J(w, b, \epsilon) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{2}\sum \epsilon_i^2 \tag{8}$$

subject to

$$y_i - \langle w, x_i \rangle - b = \epsilon_i$$

where x_i can be replaced by a kernel function $\varphi(x_i)$ as in the case of the regular SVR, to map the feature set to higher dimensional space. w and b take the same meaning as before, i.e. they are the weights and the bias of the prediction model respectively. The Lagrangian for this problem can then be expressed as

$$L(w, b, \epsilon; \alpha) = J(w, b, \epsilon) - \sum \alpha_k \{y_i - \langle w, x_i \rangle - b - \epsilon_i\} \tag{9}$$

LS-SVR has a few advantages over conventional SVR in terms of the speed of solving and lower computational intensity, while retaining the advantages of SVRs with respect to solving small sample size data and nonlinear problems [17].

1) Time-Series LS-SVR: We first implemented an LS-SVR model attempting to capture the time-series nature of the dataset and also compared its performance to a time-series implementation using a regular SVR model. As expected, we found that the LS-SVR gave us better accuracy in predictions. We used the RBF kernel, as it gave us the best performance (as described in Section III C). Our feature set in this case consisted of irradiance values obtained at different points in time between 3 and 48 hours before our targeted prediction time (also referred to as the target hour). A typical feature set consisted of a few hours of data earlier on the same day as that of the target hour, as well as data from the same hour on the previous day, or the previous two days. We found that by using irradiance values from 24 and 48 hours previous to the target hour, along with 7 hours of irradiance from the same day we obtained a model with accurate predictions. The error on the test set of this model was an RMSE of 99.54 W/m^2. This result indicated that the LS-SVR model was better suited to our application than any other model we had experimented with thus far. We hypothesize that, this is because LS-SVR works to minimize the error obtained by including the squared error term in the cost function, as opposed to the regular SVR which tolerates errors below a threshold ϵ.

2) LS-SVR Hybrid Model: The increased accuracy of the LS-SVR model on time-series predictions encouraged us to further enhance our model by implementing a hybrid prediction using a feature set that included past irradiance values, as well as the meteorological feature set. Our hypothesis was that the LS-SVR model would be able to better capture the relationship between solar irradiance and the causal inputs, as well as its nature as a time-series, as compared to the ARIMA model with causal inputs. We found that this model did remarkably well and gave extremely accurate predictions as compared to the other models, with a test set RMSE of 40.16 W/m^2. This suggests that LS-SVR with the RBF kernel is among the best-suited models for this regression problem, because it is capable of taking both time-series data and meteorological features into account, without unduly weighting one of these aspects more than the other, thereby providing a result that is far more accurate than other models. As can be seen from Figure 6, this model does a very good job of predicting solar irradiance.

C. ARIMA and LS-SVR Hybrid Model

Finally, we experimented with multiple hybrid techniques using ARIMA (to capture the seasonal nature of the data) and LS-SVR (to

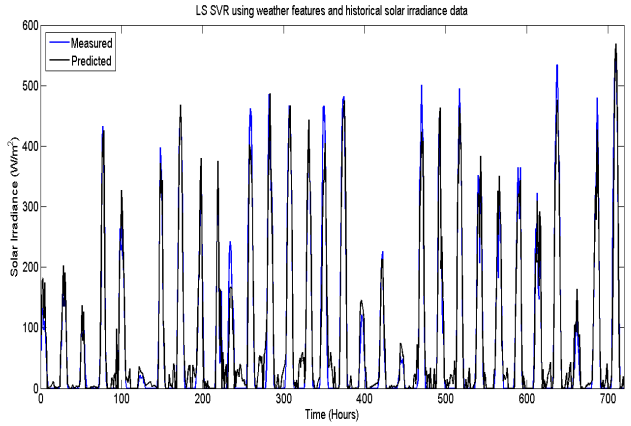

Figure 7. Hybrid LS-SVR predictions on a test set of 30 consecutive days

capture the causal nature). One technique was to model the errors of the LS-SVR prediction as a time-series using ARIMA, and another was to use both model independently to predict solar irradiance; then, subsequently regress over both of these predictions to obtain a prediction that is essentially an average of the two independent predictions. In both of these cases, the hybrid models did not lead to an improved performance, when compared to our previous models. However, we believe that with further investigation, we can develop such a hybrid model incorporating both ARIMA and LS-SVR, to help predict irradiance values to a very high degree of accuracy.

V. Summary

We have implemented a variety of models which model solar irradiance using only meteorological data, such as linear regression, locally weighted regression, and SVR, as well as models that capture the time-series nature of the data, such as ARIMA and LS-SVR. We found that the LS-SVR model performed the best; however, while these time-series models were effectively able to capture the time dependence of our target variable, they were unable to model the causal relationship between meteorological data and solar irradiance. To overcome this, we tried different hybrid models using LS-SVR and ARIMA, which did not give us optimal results. Instead, we found that by incorporating meteorological data into the feature set for the LS-SVR model, we obtained far more accurate predictions. We feel that this is because, unlike other hybrid models, the seasonal component and the weather data are more appropriately weighted by the LS-SVR model, both in terms of the features chosen by us (e.g., the extent of historical information provided) and the inherent nature of the model itself . Table 1 summarizes the performance of the models we have used in this work.

VI. Future Work

We believe that this work can be improved by using a better constructed hybrid model, as we found that our models often predict a non-zero solar irradiance value during periods of the day when there is relatively low (or zero) sunlight. By using additional features such as cloud cover data, and models that capture seasonality well, such as the ARIMA model, the accuracy of our predictions can be further improved. Another issue that we faced with models that used the ARIMA framework was the requirement that we use sequential data points in our training as well as test sets, which leads to non-optimal results. In our future work, we would like to explore methods through which we could compensate for this as well.

Table I
Errors in the predictions of various models

Model	**Samples**		**RMSE** (W/m^2)		**MAE** (W/m^2)	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Linear Reg	3600	720	194.09	192.80	145.53	141.35
LWLR	3600	720	193.63	192.33	145.11	140.94
SVR (w/o PCA)	3600	720	130.47	140.25	74.76	85.73
SVR (8 PC's)	3600	720	118.31	132.53	65.75	79.56
ARIMA	3600	720	-	188.32	-	98.54
ARIMA+LS-SVR	3600	720	33.39	199.46	18.90	130.65
SVR (Time Series)	3552	720	4.29	239.29	0.23	181.76
LS-SVR (Time Series)	3552	720	96.92	99.54	53.45	53.26
Hybrid LS-SVR	3552	720	37.18	40.16	20.98	22.34

References

[1] I. PVPS, "Pvps report–a snapshot of global pv–1992-2012," *Report IEA-PVPS T1-22*, vol. 2013, 2013.

[2] S. E. I. Association *et al.*, "Us solar market insight, 2014 year in review executive summary," *Solar Energy Industries Association Retrieved from http://www.seia.org/research-resources/us-solar-market-insight*, 2014.

[3] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, pp. 528–533, IEEE, 2011.

[4] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65–76, 2013.

[5] R. Perez, K. Moore, S. Wilcox, D. Renné, and A. Zelenka, "Forecasting solar radiation–preliminary evaluation of an approach based upon the national forecast database," *Solar Energy*, vol. 81, no. 6, pp. 809–812, 2007.

[6] P. Lauret, E. Fock, R. N. Randrianarivony, and J.-F. Manicom-Ramsamy, "Bayesian neural network approach to short time load forecasting," *Energy conversion and management*, vol. 49, no. 5, pp. 1156–1166, 2008.

[7] Y. Kemmoku, S. Orita, S. Nakagawa, and T. Sakakibara, "Daily insolation forecasting using a multi-stage neural network," *Solar Energy*, vol. 66, no. 3, pp. 193–199, 1999.

[8] G. Mihalakakou, H. A. Flocas, M. Santamouris, and C. G. Helmis, "Application of neural networks to the simulation of the heat island over athens, greece, using synoptic types as a predictor," *Journal of Applied Meteorology*, vol. 41, no. 5, pp. 519–527, 2002.

[9] A. Sfetsos and A. Coonick, "Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques," *Solar Energy*, vol. 68, no. 2, pp. 169–178, 2000.

[10] F. O. Hocaoğlu, Ö. N. Gerek, and M. Kurban, "A novel 2-d model approach for the prediction of hourly solar radiation," in *Computational and Ambient Intelligence*, pp. 749–756, Springer, 2007.

[11] 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.

[12] S. Cao and J. Cao, "Forecast of solar irradiance using recurrent neural networks combined with wavelet analysis," *Applied Thermal Engineering*, vol. 25, no. 2, pp. 161–172, 2005.

[13] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.

[14] G. E. Box and G. M. Jenkins, *Time series analysis: forecasting and control, revised ed*. Holden-Day, 1976.

[15] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, 2009.

[16] J. A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, J. Suykens, and T. Van Gestel, *Least squares support vector machines*, vol. 4. World Scientific, 2002.

[17] Y. Guo, X. Li, G. Bai, and J. Ma, "Time series prediction method based on ls-svr with modified gaussian rbf," in *Neural Information Processing* (T. Huang, Z. Zeng, C. Li, and C. Leung, eds.), vol. 7664 of *Lecture Notes in Computer Science*, pp. 9–17, Springer Berlin Heidelberg, 2012.