Prediction of Solar Radiation using hybrid algorithm of Random Forest and Particle Swarm Optimization

Sunayana Gupta, Anudeep Reddy, Yashi Baldaniya, and Rajesh Kumar

***Abstract*—Due to increased pollution, greenhouse effect and global warming resulting from power production using fossil fuels, there is increased penetration of renewable energy sources into the power production system. Over the last few years, solar radiation has become a significant means of power production using solar panels and the concept of microgrids has made solar power an indispensable source of power in the distribution system. The power production using solar energy is highly variable and weather dependent which creates a power imbalance into the system when it is penetrated without forecasting. Therefore, solar power prediction plays a critical role in the proper usage of solar energy while keeping the system stable. For automating the power system the forecast needs to be very accurate and thus, it is needed to improve the existing forecasting techniques. In this study, we have proposed a solar radiation scheme based on various meteorological factors, including temperature, humidity, wind speed, and others and used this data for building a machine learning model for prediction. We introduced a hybrid model for prediction which optimizes the parameters of Random Forest using Particle Swarm Optimization technique. The results show empirically that the hybrid RF-PSO model significantly improves the prediction accuracy and reduces the MAE error.**

***Index Terms*—Radiation, Solar prediction, Microgrids, Distribution Automation, Decision Tree, Random Forest, Particle Swarm Optimization**

I. Introduction

The continuous depletion of natural resources and increasing pollution due to power production using conventional methods has increased the emphasis of solar power over the years. Out of the total 350 GW installed capacity of the Indian power system, 63.7% is done by using fossil fuels and 21.2% is done using renewable energy resources [1]. Generation of power using these fossil fuels, coal, natural gas, and oil, releases greenhouse gases like carbon monoxide, carbon dioxide, methane, sulfur dioxide, etc which are very harmful and toxic for the environment and result in global warming. Therefore, it is needed to increase renewable penetration to tackle these problems. The Indian Government has targeted to increase solar penetration to 175 GW by 2022 [2], [3]. The new policies of the govt. the intent in developing a smart grid with integrated micro grids throughout the nation. All these microgrids have an installed solar power generation system that makes the two-way current flow possible, making the system more reliable and increasing its efficiency.

The electric power system is highly complex as it is stable only when demand matches supply. With solar power generation, the problem is that it is highly erratic and weather dependent. So for the proper functioning of the power system accurate generation prediction is needed. Solar power generation is directly related to the solar radiation, hence our paper here focuses on radiation prediction.

The researchers have explored three statistical ways of predicting the solar radiation, which are namely, time series modeling, cloud motion vector based prediction and NWP-based radiation forecast [4]. Initially the focus was on time series, an adaptive linear time-series model was suggested to predict the hourly values of solar power for horizons of up to 36 h using 15- min observations of solar power from photovoltaic (PV) systems and the global radiation forecast information provided every 3 h by the Danish Meteorological Institute as input data [5]. The other time series model was made by Kraas et al. using model predictive control [6]. It was later found out that time series models were not robust enough for predicting solar power. Thus, to incorporate more variables on which radiation depends a more robust technique was required and hence the forthcoming models were all machine learning based.

Machine learning interprets the pattern between the predicting variables and the response variable using a large data set and an algorithm to forecast the future values of the response variable. It is broadly classified in supervised, unsupervised learning methods and reinforcement learning. In supervised machine learning the algorithm learns under the guidance of labeled data and features or input variables are mapped onto an output. If the output is continuous then it is called as regression and if it's categorical then it is called classification. On the other hand, in unsupervised learning, there is no labeled data and the algorithm needs to figure out the hidden patterns in order to make predictions about the output. Clustering is a well-known example of unsupervised learning methods. In reinforcement learning, there is an agent that interacts with its environment by producing actions and discovers errors or rewards. These rewards are the basis on which the model improves.

After that cloud vector-based models gained popularity, Han Seung Jang, used satellite images and Support Vector Machine(SVM) technique and analyzed 4 years of historical data and predicted solar radiation [7]. This method was unable to account for temperature, pressure, humidity, wind direction, etc and just predicts using the number of clouds present. The Numerical Weather Prediction(NWP) models were made using different machine learning techniques like SVM [8], [9], ensemble methods [10], Artificial Neural Network(ANNs) [11], wavelet neural network, Recurrent Neural

Sunayana Gupta (e-mail: guptasunayana3@gmail.com)
Anudeep Reddy (e-mail: anudeepkatta9@gmail.com)
Yashi Baldaniya (e-mail: yashibaldaniya98@gmail.com)
Rajesh Kumar (e-mail: rkumar.ee@gmail.com)

Network(RNNs) [12], Radial Basis Function(RBF) neural network. The machine learning models were found to be more robust and accurate as compared to the time series models as they automatically detect a pattern in the data. The complex interrelationship between different variables can be understood properly using machine learning models. Also, the prediction using models based on NWP data was more precise for long hours as NWP data models consider all the variables that can have a correlation with solar radiation. Therefore, forecast for more than several hours was done using NWP data.

The prediction of solar radiation is a complex multivariate problem where different weather variables show more correlation with the radiation at different times. This created a need for a model that can take different data samples and measure the importance of the particular attribute for prediction and forecast accordingly. The decision Tree model exactly does this by selecting the appropriate feature with the help of the Gini Index, but it had an associated problem of inaccuracy and instability. Further, this was improved by growing an ensemble of these trees and taking an average of the predictions. This makes the model more stable and efficient and is termed as Random Forest [13]. Decision Trees had inherent problems of large changes in the structure of the model when small changes in the data were made and were subject to overfitting. This was dealt away with using Random Forest as a large number of trees are constructed using random sample space of data with a random selection of some of the features and the output is further averaged out. This whole process is termed as Bagging which is the abbreviation of Bootstrap Averaging.

In [14] an ensemble tree model, Random Forests (RF), is compared with SVR and K-Nearest Neighbors (KNN) in terms of 1 h ahead global solar radiation in Spain. In their hybrid method, Gala et al. [15] used gradient boosted regression and RF ensemble models for solar radiation forecasting in seven regions in Spain. They stated in that ensemble systems are quite promising and relevant models for solar radiation prediction while compared with traditional prediction models. Another model making use of RF ensemble in solar radiation estimation is conducted by Ibrahim and Khatib for hourly global solar radiation estimation in Malaysia [16] . Another RF ensemble-based study is conducted by Huang et al. [17] for global solar radiation estimation in Australia. The authors concluded that the RF ensemble model is most accurate among other competing algorithms.

But still the existing computational intelligence-based models have not yet been able to satisfy researchers in forecast precision, and the generalization capability of these machine learning models needs further improving. Hence, to overcome this problem, in this study, we have focused on improving a Random Forest model by Particle Swarm Optimization (PSO) technique and thus creating a hybrid model.

II. Methodology Used

A. Decision Trees

Decision Tree is a supervised learning algorithm that is used for regression as well as classification. Here, it is possible to see the logic behind the structure of the tree while it is being constructed. It consists of nodes, branches, and leaves. The nodes specify the feature in consideration, branches signify the decision-making and nodes give the results in terms of error. To construct the tree features are to be taken one by one according to their importance. This is done by top to down greedy search for the feature which gives the least impurity when it is used for prediction. This is done by either using the Gini Index or entropy function which are the measures of impurity.

Entropy is the amount of information is needed to accurately describe the sample. So if the sample is homogeneous, which means all the elements are similar than Entropy is 0, else if the sample is equally divided than entropy is maximum 1.

Mathematically it is written as,

$$Entropy = -\sum_{i=1}^{n} p_i * log(p_i) \tag{1}$$

Where p_i is the proportion of the samples belonging to the class i among the total samples in that split.

Gini index is a measure of inequality in the sample. It has a value between 0 and 1. Gini index of value 0 means the sample is perfectly homogeneous and all elements are similar, whereas, Gini index of value 1 means maximal inequality among elements. It is the sum of the squares of the probabilities of each class. It is illustrated as,

$$GiniIndex = 1 - \sum_{i=1}^{n} p_i^2 \tag{2}$$

Where, i is the number of classes.

B. Random Forests

The supervised algorithm of random forests works on the general principle of bootstrap aggregating or bagging on the DT algorithm. Considering a training set $X = x_1, x_2, ..., x_n$ with output $Y = y_1, y_2, ..., y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, ...B$:

1) Sample, with replacement, n training examples from X, Y; call theseX_b, Y_b.
2) Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = 1/B \sum_{b=1}^{B} f_b(x') \tag{3}$$

The fact that the bootstrapping procedure decreases the variance of the model, without increasing the bias, increases the overall performance efficiency of the model. This means that while the predictions of a single tree are highly sensitive to noise in its training set, which is the problem in DTs, the average of many trees is not, as long as the trees are not correlated. This certainly increases the computation cost, but makes the model very robust. And this correlation is ruled out as the samples as well as the features of all the trees are taken randomly.

One more advantage of the RF algorithm is that all those data points which are not used up in the training while bootstrapping are used for cross-validating the model, thus for

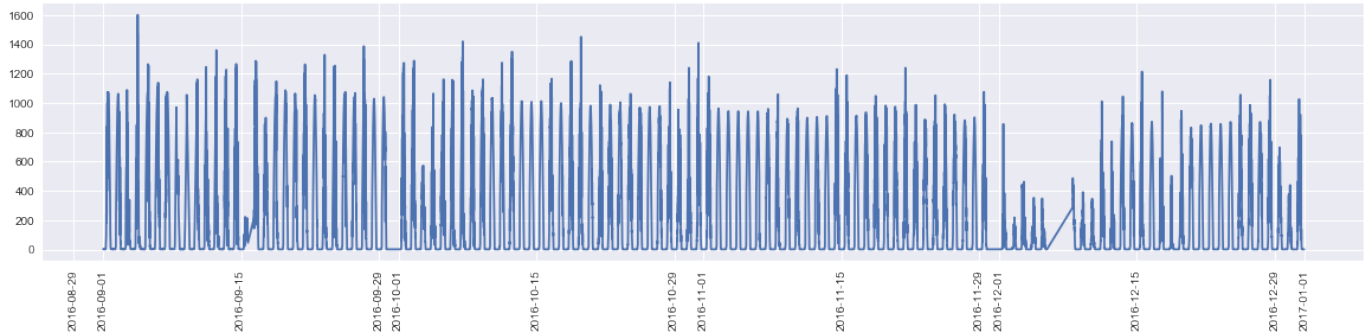


Fig. 1. Figure showing entire Solar Radiation Data

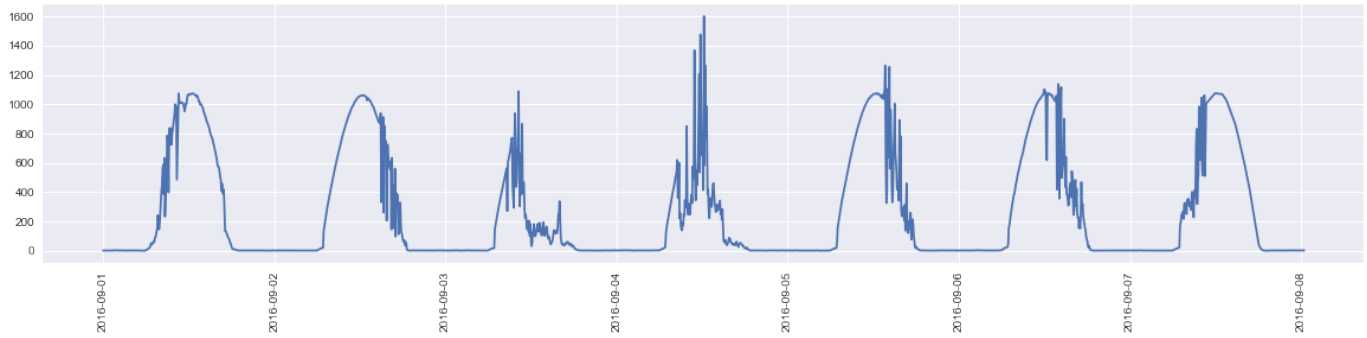


Fig. 2. First week's Solar Radiation Data

cross-validation, more data is not needed, these data points are called out-of-bag points. It is empirically proven that this does not decrease the overall performance of the model [13].

C. Particle Swarm Optimization

PSO is an evolutionary optimization computation technique developed by Kennedy and Eberhart in 1995 [18], [19]. Initially, the particles, probable solutions, are randomly "flown" into the sample space and are given random velocities. The fitness is evaluated and the particles are given velocity in the direction of the best fitness particle along with the random velocity. The acceleration given is also weighed randomly and with time-optimized solution is obtained. The velocity and the position of the particles change according to the equation:

$$v_i(t+1) = w * v_i(t) + c_1 * rand() * (p_i(t) - x_i(t)) + c_2 * rand() * (p_g(t) - x_i(t)) \quad (4)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (5)$$

Where c_1 and c_2 are positive constants, $rand()$ are uniformly distributed random numbers in [0, 1] and w is the inertia weight. The term v_i is limited to the range V_{max}. If the velocity violates this limit, then the velocity of that dimension is set at its maximum limit. Therefore, V_{max} is an important parameter, set by the user, as it decides the fineness with which the regions between the present position and the target positions are searched.

III. Dataset and Data Preprocessing

The dataset was taken from the NASA Hackathon Space App challenge which was held in 2017 [20]. These are meteorological data from the HI-SEAS weather station from four months (September through December 2016) between Mission IV and Mission V, a habitat on an isolated Mars-like site on the Mauna Loa side of the Big Island of Hawaii, relies on photovoltaic panels and solar batteries to power its long-duration Mars analog simulation studies. A total of 32687 data points of 5-minute data were there.

For each data set, the fields are:

A row number (1-n) useful in sorting this export's results, the UNIX time_t date (seconds since Jan 1, 1970). Useful in sorting this export's results with other export's results, the date in yyyy-mm-dd format, the local time of day in hh:mm: ss 24-hour format.

The units of each dataset are:

- Solar radiation: watts per meter^2
- Temperature: degree Fahrenheit
- Humidity: percent
- Barometric pressure: Hg
- Wind direction: degrees
- Wind speed: miles per hour
- Sunrise/sunset: Hawaii time

The whole data is first visualized to find the occurrence of the missing, corrupt or unusual trend in the data. At the time of visualization, it was evident that a week's data was corrupt which is shown inFigure 1.

Further, analysis of the first week's data to understand the trend of solar radiation better which is shown inFigure 2. Thus, it is clear from the week's analysis that radiation is zero before sunrise and after sunset, it is obvious as there is no sunlight during these hours. Also, there is one peak value of radiation

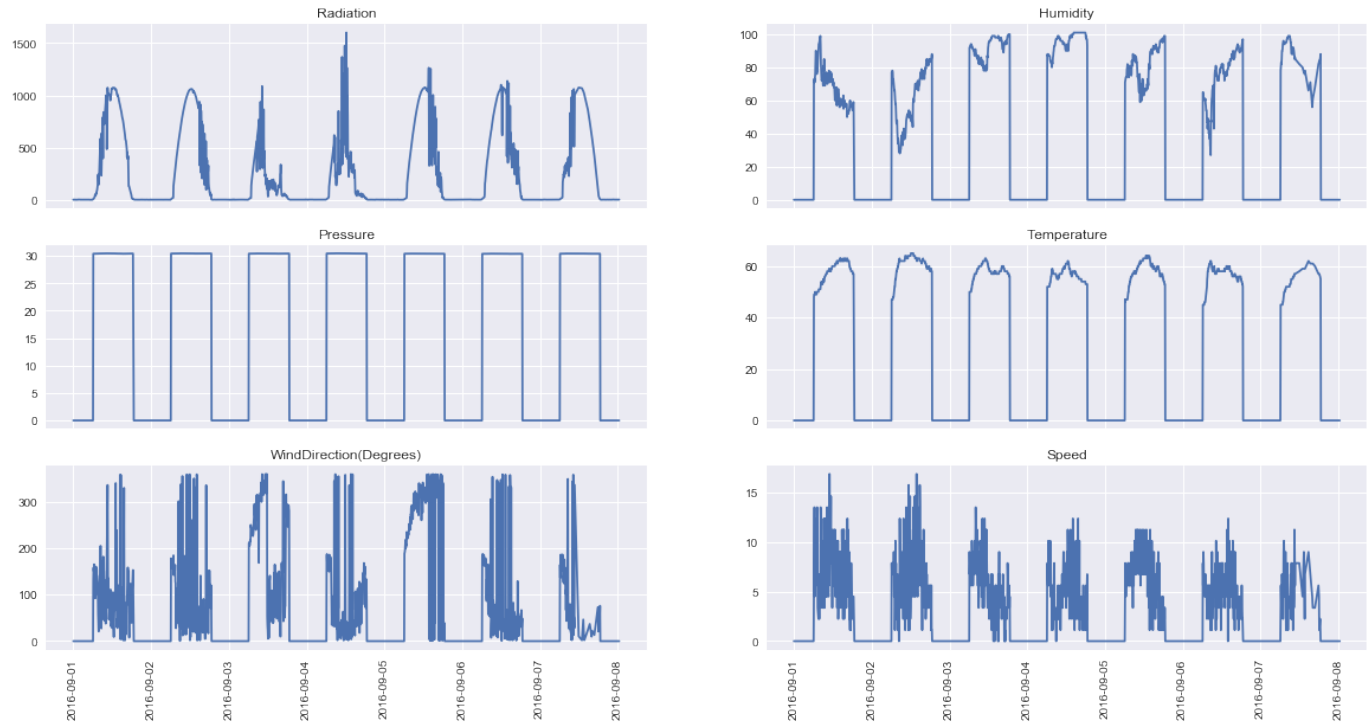


Fig. 3. Data after preprocessing

in a day due to maximum sunlight during peak hours (noon time). From the inference, we have concluded that there is no effect of any features on radiation when it is zero, so we can make the values of features also to zero for better training of the data. The data after this processing is shown in the Figure 3.

A heat map was then generated to evaluate the relative importance of different features, shown in Figure 4. Though both, Wind Direction, and Speed, are characteristics of local weather, they do not make sense as predictors of radiation because they are not much related to radiation as can be seen in the heat map. Wind direction and wind direction has a moderate correlation with temperature, but through engineering judgment, we understand that this is only correlation and not causation.

There are three timescales to consider in this data set:

- Monthly
- Daily
- Hourly

Radiation is expected to vary with the data due to seasonal weather changes. The data set only contains data from autumn and winter, so the model developed from this data might be less capable of predicting radiation during the summer.

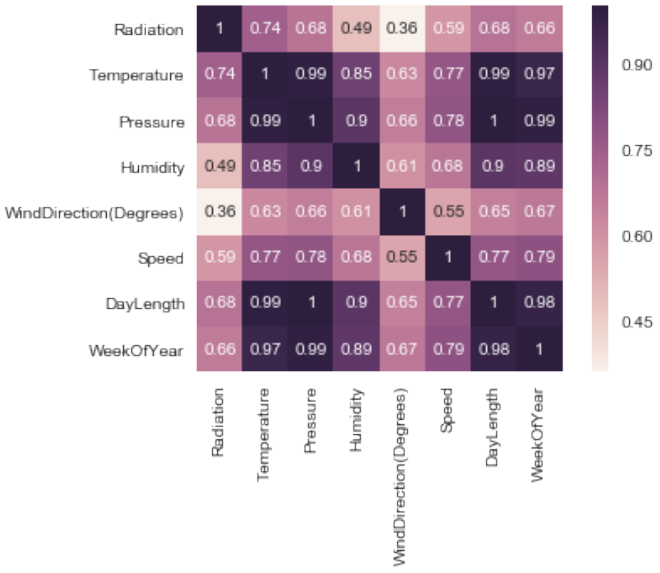


Fig. 4. Heatmap of all the features used for Solar Radiation Prediction

IV. Proposed Solar Radiation Prediction Scheme

Firstly, RF is used to predict solar radiation. For the optimal performance of random forest algorithm, it is necessary to determine the optimal parameter values.

These parameters include-

- Leaf Size
- Number of trees to be grown
- Depth of each tree
- Minimum sample split
- Feature selection

RF does not need cross-validation to estimate the parameters because it has a built-in estimation of accuracy. The optimal parameter search is done by analyzing the model performance by varying each parameter individually. This decreases the error considerably, but is certainly not optimal or satisfactory. Therefore, to improve the algorithm the parameter selection is done by using PSO technique. The algorithm followed is

shown in Figure 5. PSO optimizes all the parameters in one go by increasing the particle velocity towards the local minimum MAE and finally producing optimized results.

40 particles were kept onto the parameter space with limited bounds. The number of iterations were taken to be 50. The optimal number of trees to be grown was found to be 12, maximum depth as 8, features to be taken for prediction were 6, optimal leaf size was 1 and minimum sample split was 2. By keeping these parameters in the model, we achieved minimum error.

A twenty-four hours solar radiation forecasting model was made by Adel Mellit using Multilayer Perceptron Neural Network in Italy [21]. The used MLP consists of one input layer, one or more hidden layer and one output layer. The input layer accepts as parameters the mean daily solar radiation, the mean daily air temperature and the day of the month (i.e. at the time t), while the output layer gives as parameters the 24 h of solar radiation at the next day (i.e. at the time t + 1). This model was implemented using the NASA data that we used in this paper and the results were then compared with that of the hybrid RF-PSO model developed in this study.

V. Performance Evaluation

The performance of the proposed hybrid RF-PSO model is evaluated in this section. We compared it with Decision Tree, Random Forest and Multi Layer Perceptron Neural Network.

A. Performance Evaluation Metrics

In order to evaluate the accuracy of each prediction scheme, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Score are used in this paper.

$$RMSE = \sqrt{\frac{\sum_{i=M+1}^{N}(y_i - f(x_i))^2}{N - M}} \tag{6}$$

$$MAE = \frac{1}{N - M}\sum_{i=M+1}^{N}|(y_i - f(x_i))| \tag{7}$$

Where y_i is the measured solar radiation at each time point and $f(x_i)$ is the prediction result at each time point. For the case in this paper; i is the time sequence; N is the total number of meteorological data samples. $N - M$ is the total number of meteorological data samples for the testing stage.

The RMSE is a frequently used metric to measure the differences between the predicted values and the actual measured values through a model or an estimator. In other words, it represents the quality of the fit between the actual data and the predicted model. The RMSE is a good performance metric for finding the goodness of fit of generalized regression models. The MAE is also a performance evaluation metric used to estimate how close predictions are to the actual values in time series analysis. And, R^2 Score function is used to get the accuracy.

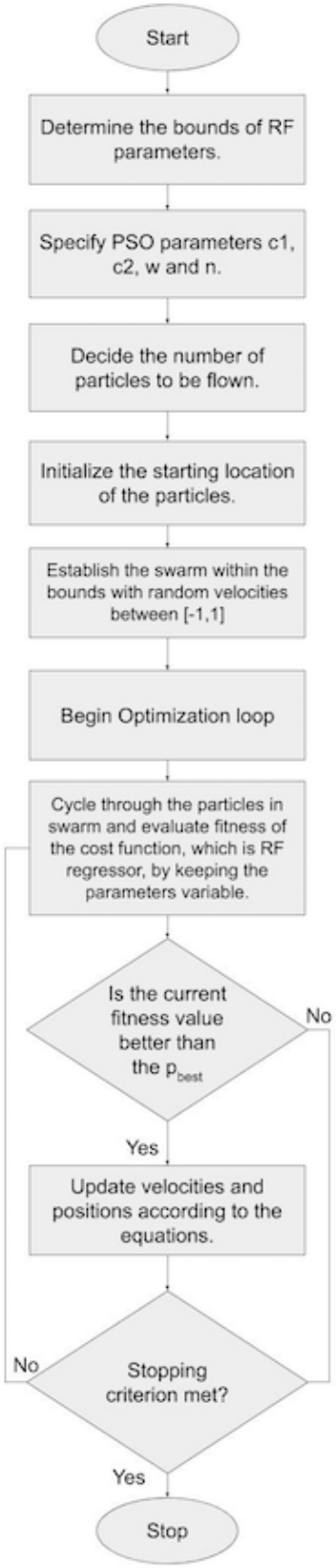


Fig. 5. The flowchart of RF-PSO algorithm.

B. Prediction Results for Solar Radiation

In the testing stage, the results of different prediction schemes (DT, RF, RF-PSO and MLP Neural Network) were compared by using the same data. The solar radiation predictions are in $W-m^{-2}$ and the duration of prediction was taken from 9th December 2016 to 1st January 2017. Figures 6, 7, 8 and 9 shows the prediction results of the DT, RF, RF-PSO, MLP schemes, respectively. It can be clearly observed from the figures that RF-PSO prediction model has been most accurate among others. Table I summarizes the numerical results of the average MRE, average RMSE, and Score of each prediction scheme for testing data. The minimum MAE and RMSE are found to be of RF-PSO scheme and maximum R^2 score of 0.82 was also obtained from RF-PSO model. Based on these results, the model with the best accuracy for solar radiation prediction was RF-PSO hybrid model.

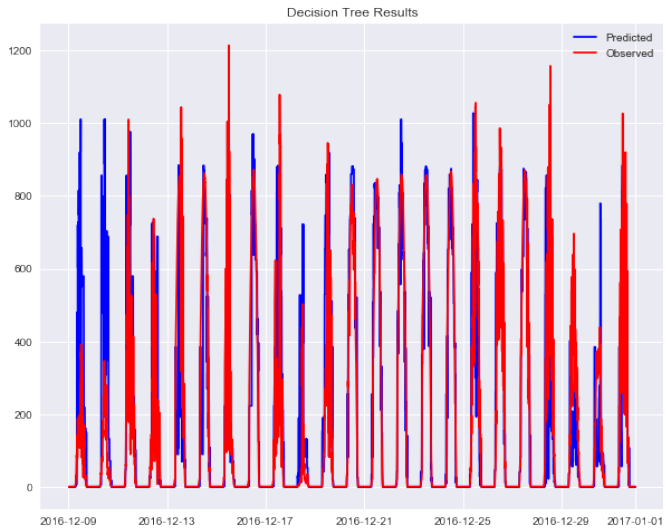


Fig. 6. DT prediction of hourly 3- week-ahead Solar Radiation

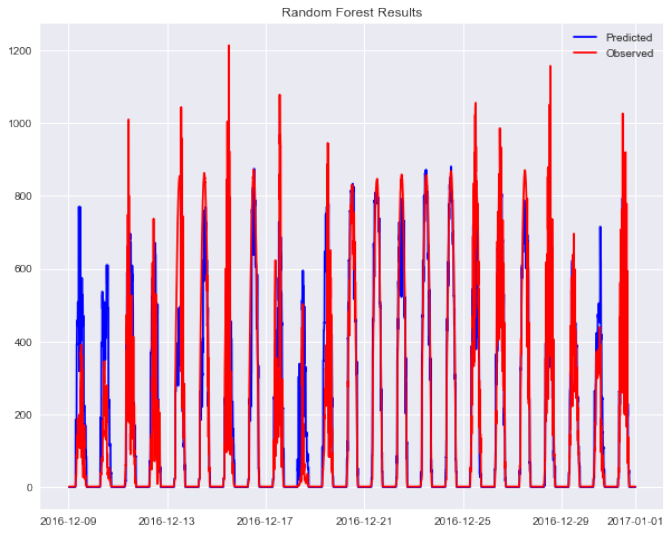


Fig. 7. RF prediction of hourly 3- week-ahead Solar Radiation

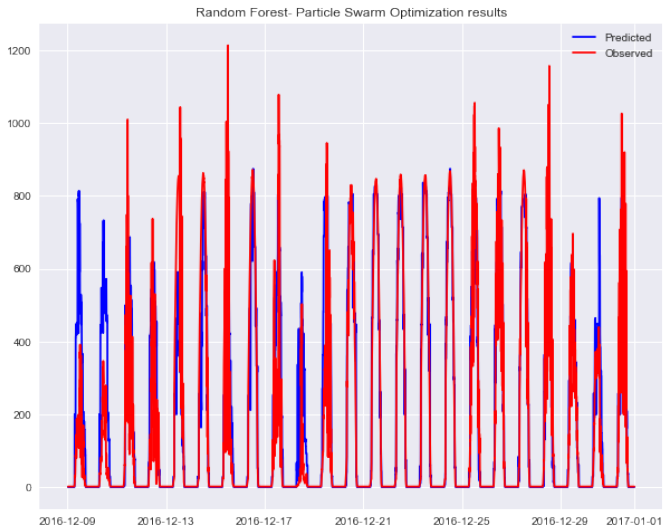


Fig. 8. RF-PSO prediction of hourly 3- week-ahead Solar Radiation

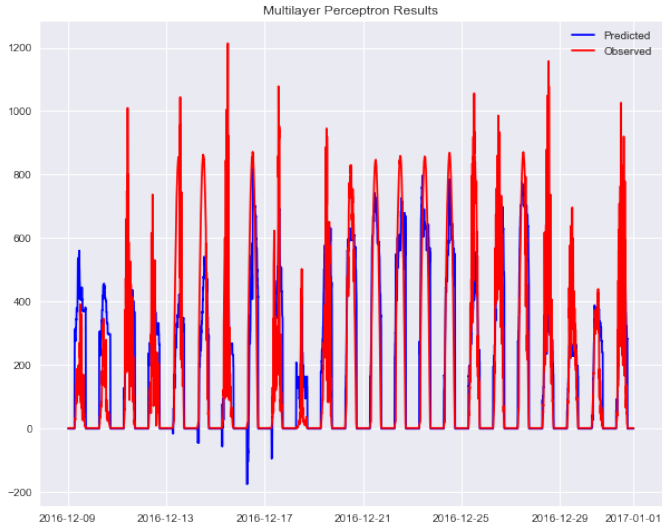


Fig. 9. MLP prediction of hourly 3- week-ahead Solar Radiation

TABLE I
PREDICTIVE ERRORS OF DIFFERENT MACHINE LEARNING MODELS

MODEL	MAE	R^2 Score	RMSE
RF-PSO	43.57	0.91	100.86
RF	49.22	0.84	112.35
DT	65.24	0.76	130.93
MLP	78.12	0.74	144.09

VI. CONCLUSION

We proposed a predictive scheme for the prediction of solar radiation based on the RF-PSO hybrid model. This method explored an optimization technique along with the robust tree based model for solar radiation.

In this paper, meteorological data of the HI-SEAS weather station on the Mauna Loa side of the Big Island of Hawaii is studied. The hybrid algorithm incorporates the advantages of both the algorithms used in it, which are fast computation and faster convergence to optimal results. PSO keeps valuable

information about local and global particles and processes accordingly in finding optimal parameters of RF. This gives the advantage to RF as it does not get stuck into sub-optimal results, but generates global optimal results.

The results of the proposed model are demonstrated and an accurate prediction of solar radiation is obtained. This provides an excellent prediction method for solar radiation monitoring in the modern day world of hybrid power systems.

REFERENCES

[1] [Online]. Available: https://powermin.nic.in/en/content/power-sector-glance-all-india

[2] [Online]. Available: https://niti.gov.in/writereaddata/files/writereaddata/files/document_publication/report-175-GW-RE.pdf

[3] [Online]. Available: https://mnre.gov.in/sites/default/files/uploads/March2019.pdf

[4] [Online]. Available: https://mnre.gov.in/sites/default/files/uploads/draft-report-fscb-remcs.pdf

[5] M. P., N. H., and H.A., "Online short-term solar power forecasting," *Sol. Energy*, pp. 1772–1783, 2009.

[6] M. S.-H. Kraas and R. Madlener, "Economic merits of a state-of-the-art concentrating solar power forecasting system for participation in the Spanish electricity market,," *Solar Energy*, vol. 93, pp. 244–255, 2013.

[7] H. S., K. Y. Bae, H. S. Park, and D. K. Sung, "Solar Power Prediction Based on Satellite Images and Support Vector Machine," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1255–1263, 2016.

[8] K. Y., H. S. Jang, and D. K. Sung, "Hourly Solar Irradiance Prediction Based on Support Vector Machine and Its Error Analysis," *IEEE Transactions on Power Systems*, pp. 1–1, 2016.

[9] D. Sung, "Solar Power Prediction Based on Satellite Images and Support Vector Machine," *Han Seung & Bae, Kuk Yeol & Park, Hongshik*, vol. 7, no. 1255-1263, pp. 10–1109, 2016.

[10] X. and T. Zhang, *Solar Power Prediction in Smart Grid Based on NWP Data and an Improved Boosting Method*, 2017.

[11] K. J.-G., Y. D.-H., L. W.-S., J.-Y., and Y. B. Kim, "Daily prediction of solar power generation based on weather forecast information in Korea," *IET Renewable Power Generation*, vol. 11, no. 10, pp. 1268–1273, 2017.

[12] B. Zhang, P. K., and C. Williams, *Solar radiation prediction based on particle swarm optimization and evolutionary algorithm using recurrent neural networks*, 2013.

[13] *L., Random forests, Mach Learn*, vol. 2001.

[14] R. Urraca, J. Antonanzas, M. A. Martinez, F. J. M. de Pison, and Torres, "F.A.: Smart baseline models for solar irradiation forecasting," *Energy Convers. Manag*, vol. 108, pp. 539–548, 2016.

[15] Y. Gala, A. Fernández, J. Díaz, and Dorronsoro, "J.R.: Hybrid machine learning forecasting of solar radiation values," *Neurocomputing*, vol. 176, pp. 48–59, 2016.

[16] I. A. Ibrahim and Khatib, "T.: A novel hybrid model for hourly global solar radiation prediction using random forest technique and firefly algorithm," *Energy Convers. Manag*, vol. 138, pp. 413–425, 2017.

[17] J. Huang, A. Troccoli, and Coppin, "P.: An analytical comparison of four approaches to modeling the daily variability of solar irradiance using meteorological records," *Renew. Energy*, vol. 72, pp. 195–202, 2014.

[18] J. Kennedy, R. Eberhart, and Y. Shi, *Swarm Intelligence*. San: Morgan Kaufmann, 2001.

[19] R. E. Kennedy, *Particle Swarm Optimization,*, 1995, vol. 4.

[20] [Online]. Available: https://2017.spaceappschallenge.org/challenges/earth-and-us/you-are-my-sunshine/details

[21] A. and A. M. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at," *Solar Energy*, vol. 84, no. 5, pp. 807–821, 2010.