Probabilistic Forecasting of Solar Power: An Ensemble Learning Approach

Azhar Ahmed Mohammed, Waheeb Yaqub and Zeyar Aung[⋆]

Institute Center for Smart and Sustainable Systems (iSmart),
Department of Electrical Engineering and Computer Science,
Masdar Institute of Science and Technology, Abu Dhabi, UAE.
{amohammed,wyaqub,zaung}@masdar.ac.ae

Abstract. Probabilistic forecasts account for the uncertainty in the prediction helping the decision makers take optimal decisions. With the emergence of renewable technologies and the uncertainties involved with the power generated through them, probabilistic forecasts can come to the rescue. Wind power is a mature technology and is in place for decades now, various probabilistic forecasting techniques are used here. On the other hand solar power is an emerging technology and as the technology matures there will be a need for forecasting the power generated days ahead. In this study, we utilize some of the probabilistic forecasting techniques in the field of solar power forecasting. An ensemble approach is used with different machine learning algorithms and different initial settings assuming normal distribution for the forecasts. It is observed that having multiple models with different initial settings gives exceedingly better results when compared to individual models. Getting accurate forecasts will be of great help where the large scale solar farms are integrated into the power grid.

Keywords: Solar Power, Probabilistic Forecasting, Pinball Loss Function, Ensemble Learning.

1 Introduction

Renewable energy sources have gained popularity in the past decade because of the increasing global warming. The environmental impact of the average barrel of oil is much more today than it was in 1950. The production of fossil fuels has some environmental impact as it releases large amounts of carbon dioxide into the atmosphere. The impact is only going to increase with the depletion of resources as we try to access less accessible resources which results in more costs on transport etc. [5]. According to the U.S. Energy Information Administration, the energy consumption across the world is going to increase by 56 percent between 2010 and 2040. Renewable energy along with nuclear power is the fastest growing energy resource, the growth rate is estimated at 2.5 percent every year [14].

[⋆] Corresponding author.

As the world looks for more environmental friendly energy resources, solar energy is viewed as an important clean energy source. The amount of solar energy striking the earth's surface every hour is more than sufficient to meet the energy needs of the entire human population for one year [9]. Solar energy is one of the most abundant resources available and it can help reduce the ever increasing dependency on energy imports. It can also protect us from the price fluctuations of the fossil fuels and help stabilize the electricity generation costs in the long run. Solar photovoltaic (PV) power plants do not emit any green house gases and they use little water, with the rapid increase in air pollution this benefit of solar PV becomes even more important. The technology roadmap for solar PV envisions a 4600 GW of PV capacity by 2050 this will result in the reduction of 4 gigatonnes (Gt) of carbon dioxide per year [15].

When large solar farm(s) are integrated into the power grid, solar power forecasting becomes essential for grid operators who make decisions about the power grid operations and as well as for electric market operators [20, 23]. The output of solar farms varies with every season, month, day and even hour. This can be challenging for grid operators and they have to turn on and off the power plants to balance the grid. Forecasting the power generated ahead of time can help avoid these problems. It also helps in the optimal use of resources [18].With large solar farms coming into the forefront, short-term ramp events will have a higher impact. Ramp events occurs when the power generation suddenly stops because of a cloud cover and also when the cloud cover lifts and the power generation ramps up again [21].

No forecast is perfect. Every forecast carries along with it a certain amount of inaccuracy and this can be attributed to the errors in real-time measurements and model uncertainty. "Probabilistic forecasting" [8] is the forecast that takes probability distributions over future events. Probabilistic forecasts are preferred to point (a.k.a. single-valued) forecasts as they take into account the uncertainties in the predicted values which helps in taking effective decisions. Probability forecasts are being used for quite some in the case of predicting binary events, events like "what is the probability that it rains today?" and other similar events. But the focus is shifting towards applying them in more general events. Some examples include flood risk assessment, weather prediction, financial risk management among others [8].

Traditionally, in the domain of solar forecasting, point forecasting methods [2, 13, 19, 12] has widely been used. Nonetheless, recently, probabilistic forecasting was used for forecasting of solar irradiance by employing stochastic differential equations [16]. In the electricity market high asymmetric costs are associated with the need to continuously balance the grid to avoid grid failure. These costs arise due to the intrinsic uncertainty associated with the emerging renewable power sources such as wind and solar. Hence a thorough understanding of the uncertainty associated with the prediction is necessary to effectively manage the grid [16]. Thus, we believe that probabilistic forecasting can become a power trend in solar power forecasting.

In this paper, as our **research contributions**, we have:

– Proposed a probabilistic forecasting method for solar power forecasting using an ensemble of different machine learning models, and
– Demonstrated that the use of ensemble learning approach offers significantly better results when compared to individual models.

2 Related Work

Bacher et al. [2] introduce a new two stage model for online short-term solar power forecasting. In the first stage, clear sky model is used to normalize the solar power and in the second stage linear time series models are used to forecast. The clear sky model gives the solar power estimation at any given time. The normalized values are obtained by taking the ratio of solar power in clear sky and the observed solar power. This is done to ensure the resulting values are more stationary, the values must be stationary to apply the statistical smoothing algorithms which assume stationarity. It takes the observed power as input and uses statistical smoothing techniques and quantile regression to give the estimated power values. The linear time series models use adaptive recursive least squares (RLS) method. The coefficients of the model are estimated by minimizing the weighted residual sum of errors, the new values are given a higher weight. This is necessary as the models need to adapt to the changing conditions.

Marquez et al. [19] use Artificial Neural Network (ANN) model to forecast the global and direct solar irradiance with the help of National Weather Service's (NWS) database. Eleven input variables are used in total, nine from the NWS database i.e. the meteorological data and two additional variables solar zenith angle and normalized hour angle are also added. To extract the relevant features, Gamma Test (GT) is used and to efficiently search the feature space, Genetic Algorithm (GA) is used. After extracting the relevant features, ANN model is used to generate the forecasts. The most important features are found to be the solar geometry variables, probability of precipitation, sky cover, minimum and maximum temperature. The models are accurate during summer rather than winter as there are more days with clear skies in summer.

Hossain et al. [12] propose a hybrid intelligent predictor for 6 hour ahead solar power prediction. The system uses an ensemble method with 10 widely-used regression models namely Linear Regression (LR), Radial Basis Function (RBF), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Pace Regression (PR), Simple Linear Regression (SLR), Least Median Square (LMS), Additive Regression (AR), Locally Weighted Learning (LWL), and IBk (an implementation of k-th Nearest Neighbor, KNN, Algorithm). Their results shows that, with respect to mean absolute error (MAE) and mean absolute percentage error (MAPE), the top most accurately performing regression models are LMS, MLP, and SVM.

E. B. Iversen et al. [16] propose a framework for calculating the probabilistic forecasts of solar irradiance using stochastic differential equations (SDE). They construct a process which is limited to a bounded state space and it gives zero probability to all the events outside this state space. To start with, a simple SDE model is used which tracks the solar irradiance from the numerical weather prediction model. This basic model is further improved by normalizing the predicted values using the maximum solar irradiance, this helps in capturing seasonality and trend. The proposed model outperformed some of the complex benchmarks.

3 Data Set

The data used in this study is taken from the Global Energy Forecasting Competition (GEFCOM) 2014 [11, 7]. This is the first competition on probabilistic forecasting in the power and energy industry. The competition lasted for 16 weeks which included 15 tasks. Each week the task is to forecast the power values for the next period. For the first task, hourly data is provided for each of the 3 zones from April 1, 2012 until April 1, 2013. For each of the remaining tasks, the data for the next month is provided. By the end of the last task, the data contained 56,953 records ranging from April 1, 2012 until May 31, 2014.

Table 1 shows the installation parameters for the three solar farms (one in each zone). Table 2 shows the independent variables (a.k.a features or attributes) that are given as part of the training and testing data sets. The objective is to predict the probabilistic distribution of the solar power generation values. The power values are normalized to range between 0 and 1 as the nominal power value for each of the solar farms is different. The location and time zones of the solar farms are not disclosed.

Table 1. Installation parameters for the solar farms. (Note: the actual names and locations of the zones and solar farms are not disclosed by the GEFCOM organizers.)

Zone	*Type*	*Number*	*Power*	*Orientation*	*Tilt*
1	Solarfun SF160-24-1M195	8	1,560W	38° clockwise from North	36°
2	Suntech STP190S-24/Ad+	26	4,940W	327° clockwise from North	25°
3	Suntech STP200-18/ud	20	4,000W	31° clockwise from North	21°

4 Methodology

This section describes the approach that we use to obtain probabilistic forecasts. The below sections describe how the data is grouped and the models used in the process followed by the methods to generate probabilistic forecasts.

4.1 Grouping the Data

The data provided contains no missing values and hence there is no need to handle missing values. The solar power values in the data are normalized to

Table 2. Description of independent variables.

Sr.	*Description*	*Unit*
1	Total column liquid water	$kg/\ m^2$
2	Total column ice water	$kg/\ m^2$
3	Surface pressure	Pa
4	Relative humidity at 1000 $mbar$	%
5	Total cloud cover	$0-1$
6	10 metre U wind component	m/s
7	10 metre V wind component	m/s
8	2 metre temperature	K
9	Surface solar rad down	J/m^2
10	Surface thermal rad down	J/m^2
11	Top net solar rad	J/m^2
12	Total precipitation	m

scale appropriately across all the three solar farms. No other data preprocessing techniques are used. All the variables provided have an impact on the solar power generated. However the data in each zone is grouped based on hour. Hence for each zone 24 different models are used.

Figure 1 shows the solar power values for the month of April 2012. A clear trend can be seen with the values going to zero during the time and slowly peaking in the afternoon, there are more fluctuations in the data and it is widely spread. Whereas Figure 2 shows the hourly data for a particular time (1 am) observed for the year 2013. This shows that the values are not as dispersed as in the case of daily values. This will help avoid the problem of outliers while fitting the models.

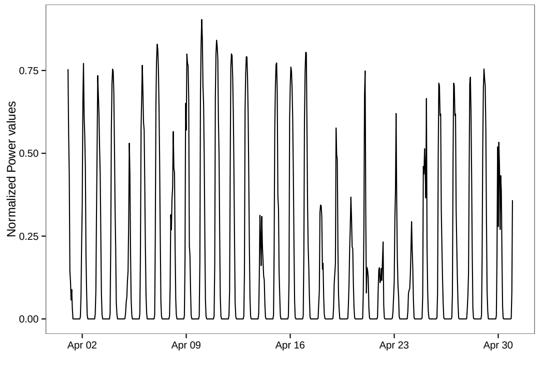

Fig. 1. Observed solar power values for the month of April, 2012 in Zone 1.

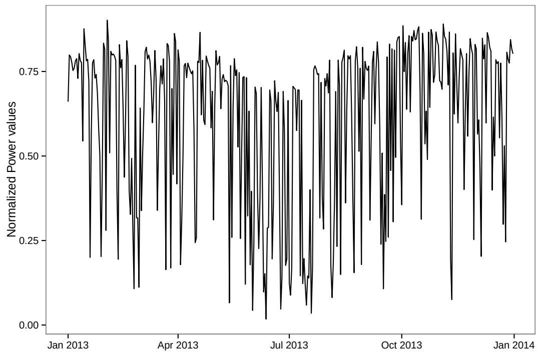

Fig. 2. Solar power values recorded at 1 am across the year 2013 in Zone 1.

4.2 Base Models Used

The following seven individual machine learning methods are used as base models for ensemble learning in this study.

- **Decision Tree Regressor:** A model is fit using each of the independent variables. For each of the individual variables, mean squared error is used to determine the best split. Maximum number of features to be considered at each split is set to the total number of features [4].
- **Random Forest Regressor:** An ensemble approach that works on he principle that a group of weak learners when combined would give a strong learner. The weak learners used in random forest are decision trees. Breiman's bagger is used in which at each split all the variables are taken into consideration [3].
- **KNN Regressor (Uniform):** The output is predicted using the values from the k-nearest neighbors (KNNs) [1]. In the uniform model, all the neighbors are given an equal weight. Five nearest neighbors are used in this models i.e the 'k' value is set to five. Distance metric "Minkowski" is used in finding the neighbors.
- **KNN Regressor (Distance):** In this variant of KNN, the neighbors closer to the target are given higher weights. The choice of k and distance metric are same as above.
- **Ridge Regression:** Penalizes the use of large number of dimensions in the dataset using linear least squares to minimize the error [10].
- **Lasso Regression:** A variation of linear regression that uses shrinkage and selection method. Sum of squares error is minimized but with a constraint on the absolute value of the coefficients [22].
- **Gradient Boosting Regressor:** An ensemble model usually using the decision trees as weak learners, it builds the model in stage-wise manner by optimizing the loss function [6].

4.3 Generating Probabilistic Forecasts using Ensemble Learning

Three ensemble learning approaches are used to generate the probability forecasts using the values generated from the models mentioned above.

- **Naive model:** A cumulative probability distribution where the first quantile is the lowest among the values and 99th quantile was the highest.
- **Normal distribution:** The mean and standard deviation of the point forecasts from above seven individual models are used to generate 99 quantiles assuming normal distribution.
- **Normal distribution with different initial settings:** This method is similar to the above normal distribution method but the models are run with two different initial settings, including the month as a variable and taking only the values for the 30 most recent days as inputs.

5 Evaluation Metric

Pinball loss function [17] is used as an evaluation metric as we are dealing with probabilistic forecasts and not point forecasts. Let the 99 quantiles generated 0.01, 0.02,..., 0.99 be defined as $q_1, q_2 \ldots, q_{99}$ respectively and $q_0 = -\infty$ the natural lower bound and $q_100 = +\infty$ the natural upper bound. Then the score L for q_i is defined as:

$$L(q_i, y) = \begin{cases} (1 - i/100)(q_i - y) & \text{if } y < q_i \\ i/100(y - q_i) & \text{if } y \geq q_i \end{cases}$$

where y is the observed value and $i = 1, 2, \ldots, 99$. To evaluate the overall performance, this score is averaged across all target quantiles. Lower scores indicate better forecasts.

6 Results and Discussions

The results when using only the individual models are shown in Table 3. The average pinball loss in the table refers to average value across 15 months starting from April 2013 for the 3 zones. Random forests and gradient boosting regression gave the best results. A small part of the results for a 24 hour period is shown in Figure 3 for better resolution. The point forecasts are very close to the observed values but they do not take into account the error associated with them and hence we need probabilistic forecasts. There is a significant improvement in the results when the ensemble methods are used as shown in table 4.

Among the three zones, Zone1 had the best results. For example the average error value in Zone1 for Naive model is 0.019673 whereas it is 0.022617 and 0.022531 for Zone2 and 3 respectively. The hourly error values also varied significantly with hours 11 through 18 showing a zero error value since the power generated during that times is zero. The highest error values are observed during the hours zero through four. The monthly error values are shown in figure

Table 3. Individual performance of different models.

Individual Model	*Average Pinball Loss*
Decision Tree	0.0249
Random Forest	0.0194
KNN (Uniform)	0.0223
KNN (Distance)	0.0224
Ridge Regression	0.0206
Lasso Regression	0.0218
Gradient Boosting	0.0193

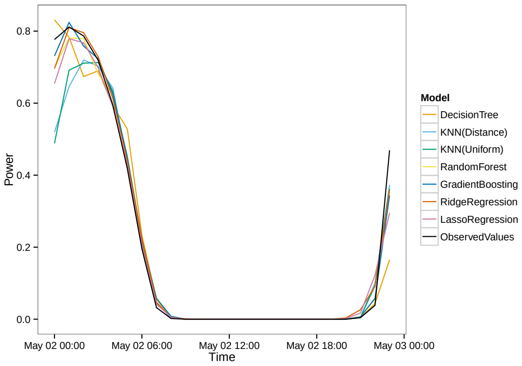

Fig. 3. Point forecasts of different models for 24 hour period on May 2nd, 2013.

Table 4. Ensemble performance of the three methods

Ensemble Model	*Average Pinball Loss*
Naive model	0.0157
Normal Distribution	0.0152
Normal Distribution with different initial settings	0.0148

4. Very low error are observed in the months of May and June whereas August has the highest error rate. These fluctuations in the error rates are caused by the cloud cover. In general better forecasts are achieved in summer because of clear sky and there are high error rates during the winter with more cloud cover during the daytime.

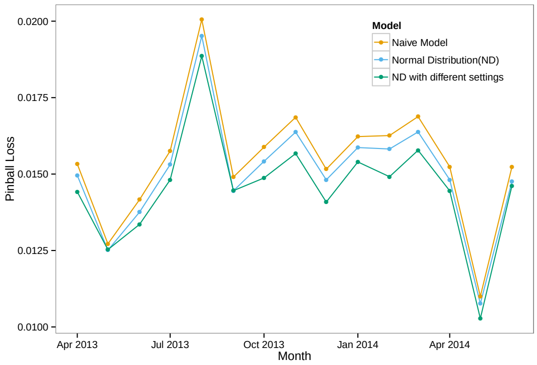

Fig. 4. Pinball loss values for different months.

7 Conclusion

In this study, we generated the probabilistic forecasts using ensemble methods assuming normal distribution for the point forecasts obtained from the individual models. For each hour in each zone, a different model is used to avoid problems with outliers. There is a significant improvement in the performance of ensemble model when compared to individual models. However, these models can be further improved as the assumption that the forecasts follow normal distribution is too restrictive.

References

1. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46(3), 175–185 (1992)
2. Bacher, P., Madsen, H., Nielsen, H.A.: Online short-term solar power forecasting. Solar Energy 83(10), 1772–1783 (2009)
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. Breiman, L., Friedman, J., , Stone, C., Olshen, R.A.: Classification and Regression Trees. Taylor & Francis (1984)
5. Davidson, D.J., Andrews, J.: Not all about consumption. Science 339(6125), 1286–1287 (2013)

6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics pp. 1189–1232 (2001)
7. GEFCOM: Global energy forecasting competition 2014 (2014), http://www.drhongtao.com/gefcom
8. Gneiting, T., Katzfuss, M.: Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151 (2014)
9. Goldemberg, J., Johansson, T.B., Anderson, D.: World Energy Assessment: Overview: 2004 Update. United Nations Development Programme, Bureau for Development Policy (2004)
10. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1), 55–67 (1970)
11. Hong, T.: Energy forecasting: Past, present, and future. Foresight: The International Journal of Applied Forecasting Winter 2014, 43–48 (2014)
12. Hossain, M.R., Oo, A.M.T., Shawkat Ali, A.B.M.: Hybrid prediction method for solar power using different computational intelligence algorithms. Smart Grid and Renewable Energy 4(1), 76–87 (2013)
13. Huang, Y., Lu, J., Liu, C., Xu, X., Wang, W., Zhou, X.: Comparative study of power forecasting methods for PV stations. In: Proceedings of the 2010 IEEE International Conference on Power System Technology (POWERCON). pp. 1–6. IEEE (2010)
14. International Energy Agency: International energy outlook 2013 (2013), http://www.eia.gov/forecasts/archive/ieo13
15. International Energy Agency: Technology roadmap: Solar photovoltaic energy - 2014 edition (2014), http://www.iea.org/publications/freepublications/publication/technology-roadmap-solar-photovoltaic-energy---2014-edition.html
16. Iversen, E.B., Morales, J.M., Møller, J.K., Madsen, H.: Probabilistic forecasts of solar irradiance using stochastic differential equations. Environmetrics 25(3), 152–164 (2014)
17. Koenker, R.: Quantile Regression. Cambridge University Press (2005)
18. Letendre, S.E.: Grab the low-hanging fruit: Use solar forecasting before storage to stabilize the grid (2014), http://www.renewableenergyworld.com/rea/news/article/2014/10/grab-the-low-hanging-fruit-of-grid-integration-with-solar-forecasting
19. Marquez, R., Coimbra, C.F.M.: Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. Solar Energy 85(5), 746–756 (2011)
20. Perera, K.S., Aung, Z., Woon, W.L.: Machine learning techniques for supporting renewable energy generation and integration: A survey. In: Data Analytics for Renewable Energy Integration - Second ECML PKDD Workshop, DARE 2014, Lecture Notes in Computer Science, vol. 8817, pp. 81–96 (2014)
21. Runyon, J.: Transparency and better forecasting tools needed for the solar industry (2015), http://www.renewableenergyworld.com/rea/news/article/2012/12/transparency-and-better-forecasting-tools-needed-for-the-solar-industry
22. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267–288 (1996)
23. Wikipedia: Solar power forecasting (2015), http://en.wikipedia.org/wiki/Solar_power_forecasting