Solar Power Forecasting Using Artificial Neural Networks

Mohamed Abuella, Student Member, IEEE
Department of Electrical and Computer Engineering
University of North Carolina at Charlotte
Charlotte, USA
Email: mabuella@uncc.edu

Badrul Chowdhury, Senior Member, IEEE
Energy Production & Infrastructure Center
University of North Carolina at Charlotte
Charlotte, USA
Email: b.chowdhury@uncc.edu

***Abstract*—In recent years, the rapid boost of variable energy generations particularly from wind and solar energy resources in the power grid has led to these generations becoming a noteworthy source of uncertainty with load behavior still being the main source of variability. Generation and load balance is required in the economic scheduling of the generating units and in electricity market trades. Energy forecasting can be used to mitigate some of the challenges that arise from the uncertainty in the resource. Solar power forecasting is witnessing a growing attention from the research community. The paper presents an artificial neural network model to produce solar power forecasts. Sensitivity analysis of several input variables for best selection, and comparison of the model performance with multiple linear regression and persistence models are also shown.**

Keywords*—Sensitivity Analysis, artificial neural networks, solar power forecasts.*

I. Introduction

Variable energy generations, particularly from renewable energy resources such as wind and solar energy plants have created operational challenges for the electric power grid because of the uncertainty involved in their output in the short term. When the penetration level of the variable generation is high, the intermittency of these resources may adversely affect the operation of the electric grid. Thus, wherever the variable generation resources are used, it becomes highly desirable to maintain higher than normal operating reserves and efficient energy storage systems to manage the power balance in the system. The operating reserves that use fossil fuel generating units should be kept as low as possible to get the highest benefit from the deployment of the variable generations [1]. Therefore, forecasting these renewable resources takes on a vital role in the operation of power systems and electricity markets.

The rest of the paper is organized as follows: Section II includes a review of statistical forecasting models for variable generations and a brief introduction to artificial neural networks (ANN). Section III describes the data used to build the ANN. Section IV discusses the various solar power forecasting modeling stages. Section V presents the results and evaluation of the models. Section VI provides the conclusions.

II. Statistical Variable Generation Forecasting Models

Forecasting models are continuously being improved to generate more accurate forecasts of solar and wind power. In this section, the statistical models that use both non-learning and learning approaches are described.

A. Statistical Non-Learning Approach Models

These models describe the connection between predicted solar irradiance from numerical weather predictions (NWP) and solar power production directly by statistical analysis of time series from historical data without considering the physics of the system. This connection can be used for forecasts in the future plant outcomes. Plenty of regression models are already implemented as time-series forecasting models, some of which include autoregressive integrated moving averages (ARIMA), and multiple linear regression (MLR) analysis model [2] to name just two types.

B. Statistical Learning Approach Models

Artificial intelligence (AI) methods are used to learn the relationship between predicted weather conditions and the power output generated as historical time series. Unlike statistical approaches, AI methods use algorithms that are able to implicitly describe nonlinear and highly complex relationship between input data (NWP predictions) and output power instead of an explicit statistical analysis. For both the statistical and AI approaches, high quality time series data consisting of weather predictions and power outputs from the

past are very important [3], [4]. One of the most common statistical learning models is the artificial neural network.

The ANN is loosely a simple biological analogy of the brain. They are implemented in widespread applications with different AI approaches such as supervised, unsupervised, and reinforcement learning approaches. In the supervised learning approach, the ANN learns from the data by training them to approximate and estimate the function or the relationship between the input and the output variables.

The major milestones of neural networks dates back to McCulloch and Pitts (1943), Widrow and Hoff (1960), and then in the mid-1980s to Werbos (1974), Parker (1985) and Rumelhart.(1986), who proposed the back-propagation algorithm [5].

With the help of applied mathematics, the backpropagation algorithm helps train the ANN to recognize similar patterns. In the backpropagation concept, information flows in one direction between the neurons (nodes) and the errors *back-propagate* in the opposite direction, changing the strength (weights) of the synapses (links) between the nodes while attempting to minimize the errors by using an appropriate optimization technique such as the gradient descent method. After sufficient training iterations with known input data, the weights between the nodes are adjusted until they give a correct response. Then, the ANN will give the correct response to the (unknown) input data that it has never seen before. The ANN can learn to generalize in this fashion. More sophisticated algorithms are introduced for training ANNs with different optimization methods to improve the performance [6].

ANNs are capable of providing forecasting for the variable generations of wind and solar power when the historical data is available. The ANN is considered as a black box because it is not providing a sufficient qualitative understanding of the relationship between the input and the output variables. A review of ANN-based forecasting models that are implemented to forecast the solar irradiance and energy can be found in [7].

In this paper, ANN model uses the most widely used "vanilla" feed-forward neural networks, sometimes called the single hidden layer network. The ANN model is used as a nonlinear statistical tool to forecast solar power. Its performance is compared with other models.

III. The Data

A. Data Source

The data is derived from the Global Energy Forecasting Competition 2014 (GEFCOM2014) which also including forecasting in the domains of electric load, wind power, solar power, and electricity prices [8].

B. Data Description

The objective is to determine the solar forecasts in hourly steps through a month of forecast horizon.

The target variable is the solar power. There are 12 independent variables available from the European Centre for Medium-Range Weather Forecasts (ECMWF) that are used to produce solar forecasts. These are:

- Total column liquid water of cloud (tclw) - (kg/m^2).
- Total column ice water of cloud (tciw) - (kg/m^2)
- Surface pressure (SP) - (Pa).
- Relative humidity at 1000 mbar (r) - (%).
- Total cloud cover (TCC) - (0-1)
- 10 meter U wind component (U) - (m/s).
- 10 meter V wind component (V) - (m/s).
- 2 meter temperature ($2T^{\circ}$) - (K)
- Surface solar radiation down (SSRD) - (J/m^2)-accumulated field.
- Surface thermal radiation down (STRD) - (J/m^2)-accumulated field.
- Top net solar radiation (TSR) - (J/m^2) -accumulated field
- Total precipitation (TP) - (m) - accumulated field.

The last four weather variables (i.e. solar and thermal radiations besides the precipitation) are given in accumulated field values, and not average values. They are increasing for every hour until the end of the day and then start again in accumulation [9]. The wind variables are given as two components U and V representing wind directional components. The U-component is for east and west directions, while the V-component is for north and south directions. The resultant vector of both wind components is the wind speed vector.

IV. Solar Forecast Modeling

The flowchart of the general solar forecasting modeling steps is shown in Fig. 1.

A. Data Preparation

It is always a good idea to get the analysis of the historical data before setting up the forecasting model. The available historical data contains the solar power and all 12 weather variables.

Fig. 1. Flowchart diagram of the solar forecasting modeling

The data preparation is an important step for treating the data to be ready for the analysis and modeling steps. The various steps of the data preparation are shown in Fig. 2. Fig. 3. (a) shows the list of the available solar power and weather data observations. The shaded months are used to test the model performance. Fig. 3. (b) shows the box plot of the distribution of the observed solar power data for the complete year of 2012. Note, the order of months that appears in the box plot does not necessary means the same order of months in the calendar year.

Fig. 2. Flowchart diagram of data preparation

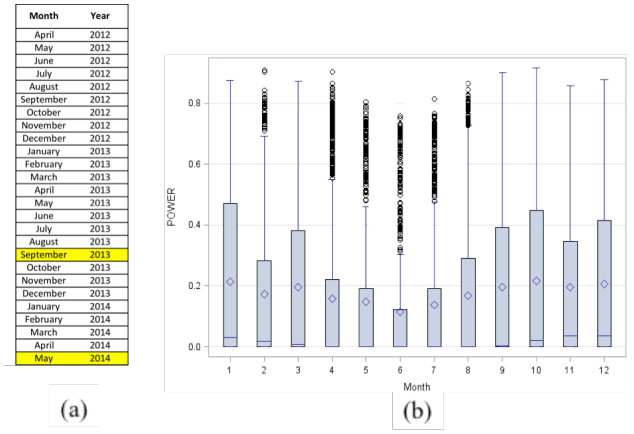

Fig. 3. (a) The entire available data. (b) Box plot of the distribution of observed solar power in the months of 2012.

The scatter plot is also useful to get a sense of the relationships between the predictor variables (the weather variables) and response variable (the solar power). Fig. 4 presents the advantage of plotting the data in scatter plots for the observed power with respect to the solar irradiance, also called surface solar radiation down (SSRD). The scatter plot on the left of Fig. 4 is for the SSRD given in accumulated values (J/m^2), as a solar irradiance of a particular area. The plot on the right is for the average values of the solar irradiance SSRD (W/m^2). The relationship between the variables on the right plot is more obvious, and one can observe the positive relationship with high positive correlation. For getting the average values for the accumulated field data, we apply the formula in Eq. (1).

$$Avg(t) = \frac{Acc(t+1) - Acc(t)}{3600} \qquad (1)$$

Where t is the time in hour steps, Avg and Acc are the average and accumulated values of the data respectively.

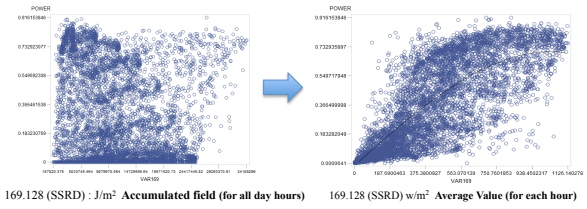

Fig. 4. Scatter plot of the observed solar power vs. Solar Irradiance

From the scatter plots, the outliers don't change the general data trends. The majority of the extreme points in the observed data occurs near sunrise and sunset periods. By experiment, data cleansing is conducted which led to a tiny improvement in forecasts. However, this does not mean one should underestimate the data cleansing stage in data preparation before the modeling stage since sometimes, outliers could be generated from data entry issues.

B. Sensitivity Analysis

Selection of the weather variables as predictors or input variables by plotting is cumbersome since there are twelve predictor variables to choose from. So a correlation and sensitivity analysis is carried out for the historical data to investigate the most effective variables. Table I presents the outcome of the sensitivity analysis. The sensitivity analysis is conducted for each weather variable as an input to the ANN, and by running the ANN a number of times, we get the minimum and the maximum of the root mean square error (RMSE) between the observed and forecasted solar power. The latter is the output variable of the ANN model. The outcome of each run of ANN model is different because for every training run, the initial values of the parameters of the model have different random values.

It was observed that the solar irradiance, in addition to the time of day, the surface irradiance and net top solar irradiance variables and their second order polynomial or quadratic terms have the highest correlation with solar power. The relative humidity and the temperature at 2m also have a noticeable impact on solar power compared to other less impactful variables. In fact, the temperature plus the solar irradiance are chosen in physical-approach models to forecast the power from PV systems [10]. At the top of the input variables rank is the second-degree polynomial of the solar irradiance, which has a correlation with the solar power stronger than the solar irradiance itself. This is because the solar power scatter graph has a parabola trend rather than a straight line.

TABLE I

THE SENSITIVITY ANALYSIS RESULT FOR EACH INPUT VARIABLE OF ANN MODEL

Weather Variables	RMSE	
	Min	Max
2nd Poly. S. Solar Irr.	0.106240	0.107990
Surface Solar Irr.	0.106500	0.108470
3rd Poly. S.Solar Irr.	0.107600	0.112630
Top Solar Irr.	0.110660	0.112250
2nd Poly. Top Solar Irr.	0.111760	0.121760
Hours	0.114930	0.120290
Relative Humidity	0.223370	0.227430
2-m Temperature	0.236330	0.261330
10m- U Wind	0.255670	0.259180
10m- V Wind	0.263570	0.268740
Thermal Irradiance	0.265080	0.265710
Precipitation	0.266860	0.283840
Cloud Cover	0.268020	0.270300
Cloud Water Content	0.269210	0.271260
Cloud Ice Content	0.269240	0.271790
Months	0.270160	0.270760
Month Days	0.270710	0.271520
Surface Pressure	0.270820	0.272463
Year Days	0.271210	0.437640

C. The Model Building

The main steps of building the forecasting model are shown in Fig 5. MATLAB is used for building the ANN model as shown in Fig 6. It is a feed-foreword curve fitting type, which works well when it is not necessary to use the past delayed values of the output as a feedback variable, also several available inputs are applied to extract a better regression. The ANN has the input layer, a hidden layer, and the output layer. The hidden layer has 15 nodes besides the bias node, which is feeding into every node in the hidden and output layers. The bias node is for shifting the activation function left or right, because sometimes the variation in the weights is not enough to minimize the errors and enhance the model performance.

Fig. 5. Flowchart diagram for building the ANN model

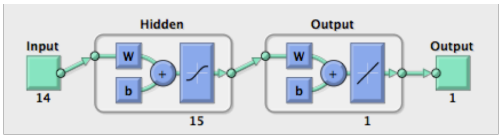

Fig. 6. Block diagram of the ANN topology.

When the predictor variables interact with each other and grouped in the ANN's input layer, they could lose some of their correlation power. Therefore, in the total mix of selected variables, the best candidate model is needed. So every time a new weather variable is added as a new input to the existing list of inputs, the ANN must be run several times to calculate the RMSE until the best group of input variables is found. By carrying out the three main steps of building the model, training, and testing to reduce the dimension of inputs variables, we arrive at the candidate model with most efficient performance. The ANN model with 14 input variables was found to have the least RMSE as shown in the shaded row in Table II. An ANN model with a large number of input variables and nodes could lead to the overfitting issue, which is the situation where the model performs well in the training stage, but produces inaccurate forecasts in the testing stage. For the purpose of solar forecasting, we found the candidate model with 14 input variables to have the least RMSE. For each case listed in Table II, the hidden layer of the ANN had 15 nodes.

TABLE II

CORRELATION ANALYSIS RESULT FOR INPUT VARIABLES OF THE ANN MODEL (DIMENSION REDUCTION OF INPUTS)

Top Grouped Weather Variables	RMSE	
	Min	Max
1	0.1150	0.1170
2	0.0855	0.0876
3	0.0847	0.0856
4	0.0853	0.0862
5	0.0794	0.0809
6	0.0795	0.0837
7	0.0801	0.0819
8	0.0780	0.0799
9	0.0773	0.0796
10	0.0760	0.0818
11	0.0761	0.0784
12	0.0737	0.0804
13	0.0759	0.0949
14	0.0720	0.0794
15	0.0743	0.0771
16	0.0762	0.0853
17	0.0785	0.0998

The training stage includes all the historical data except for the last month before the testing month. The last month of the historical data is used for the validation stage. The validation stage is required to avoid the overfitting issue since the ANN has parameters that are changing their values at the validation stage. Meanwhile, the testing is conducted for two cases, to generate the solar power forecasts by an hourly resolution for September 2013 and May 2014. Keep in mind, the training of each case is carried out separately, May 2014 has more historical data than in September 2013 case. Next, an investigation of the ultimate performance of the model and comparisons with other models is done.

V. Model Results and Evaluation

The following measures are used to evaluate the accuracy of the forecasts and the model performance: plots and graphs, Root Mean Square Error (RMSE), the correlation coefficient (R) between the forecasts and the actual measured solar power, and a comparison with other models. For comparison purposes, the Multiple Linear Regression (MLR) Analysis model [2], and the persistence forecasts model are used. The persistence model as its name implies, is obtained by keeping the actual solar power output at the current hour and using it as a solar power forecast for the next future hour. The RMSE is defined as:

$$RMSE = \sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad (2)$$

Where $\hat{Y}$ is the forecasted value of the solar power forecasts and Y is the observed value of the solar power. $\hat{Y}$ and Y are normalized values of the nominal power capacity of the solar power system. RMSE for all forecasting hours should be minimized to yield more accurate forecasts.

If the training and testing of the model are carried out for just the daylight hours and filtering out the night hours (which have zero solar power generation), the RMSE and the correlation coefficient R should also be determined for these day hours only without including the night hours.

The line plots are shown in Fig. 7 for the actual solar power and its corresponding forecasts from ANN model and compared with MLR and persistence forecasts models. The day-ahead weather variables forecasts are used as input variables for the ANN and they are periodically generated and updated daily to forecast the next days. Therefore, the output of the forecasting model, which is the solar power forecasts, doesn't change much by increasing the horizon time. The zoomed in plot on the right is for a sample day with a lower spike in the solar power generation. The forecasts from the ANN model have tracked the actual power better than the other models.

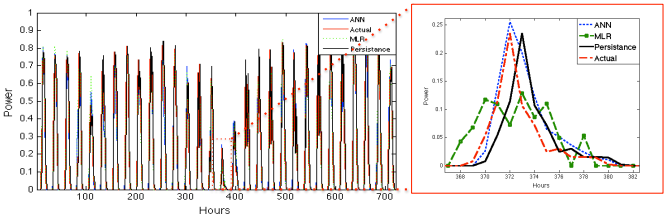

Fig. 7. The line plots for actual solar power and the forecasts from ANN, MLR, and Persistence models

As shown in Fig. 8, the actual and the forecasts are plotted with residuals plot. The residuals plot has both positive and negative values. There appear to be many residuals of the ANN that are lying at or near the zero value as shown on the top right plot which indicates that the generated forecasts are unbiased. The correlation coefficients R between the actual power and the forecasts for all models are also plotted. Table III summarizes the evaluation results of both test cases: September 2013 and May 2014 of the ANN and other model performance. It is obvious that the ANN outperforms other models. In addition, the May 2014 case has accurate forecasts because there are more historical data included in the training and validation stages of the model.

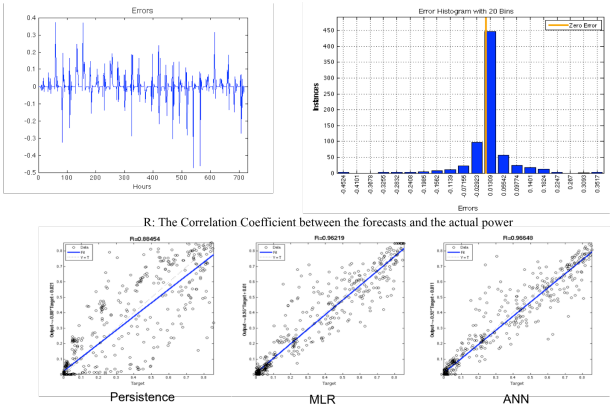

Fig. 8. The residuals plot of ANN model and the correlation coefficient plots for solar power forecasts of ANN, MLR and Persistence models

TABLE III

THE SUMMARY OF FORECASTS FOR BOTH TEST MONTHS

Test Case	September 2013		May 2014	
Model	**RMSE**	**R**	**RMSE**	**R**
ANN	0.0697	0.9665	0.0554	0.97097
MLR	0.0738	0.9622	0.0571	0.96987
Persistent	0.1306	0.8812	0.1125	0.8750

VI. Conclusion

The artificial neural networks model outperforms the multiple linear regression analysis MLR model and the persistence model. The performance of the ANN depends on how well it is trained and on the quality of the data that is used.

The feed-forward ANN with 14 weather variables and with hourly step size for forecasts performed better than the recursive neural networks. The normalized input data doesn't improve the performance, but removing the night hours slightly improves the model performance. Plotting the data, investigating the correlation and sensitivity analysis between the variables, as well as data cleansing of outliers are essential data preparation steps before building the forecasting model. In the clear sky hours, the model produces more accurate forecasts than cloudy hours. The more accurate weather forecasts we use, the more accurate solar power forecasts will be produced. Using the classification variables and the interactions between the variables enhances the performance of the MLR model significantly but this is not the case for the ANN model. With additional historical data, the model performance will improve

References

[1] A. Botterud, J. Wang, V. Miranda, and R. J. Bessa, "Wind power forecasting in US electricity markets," *The Electricity Journal*, vol. 23, no. 3, pp. 71–82, 2010.

[2] M. Abuella and B. Chowdhury, "Solar Power Probabilistic Forecasting by Using Multiple Linear Regression Analysis," in *IEEE Southeastcon Proceedings*, Ft. Lauderdale, FL, 2015.

[3] M. Lange and U. Focken, *Physical Approach to Short-Term Wind Power Prediction*. Springer, 2006.

[4] B. Ernst, B. Oakleaf, M. L. Ahlstrom, M. Lange, C. Moehrlen, B. Lange, U. Focken, and K. Rohrig, "Predicting the wind," *IEEE power energy Mag.*, vol. 5, no. 6, pp. 78–89, 2007.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2 Edition. Springer-Verlag New York, 2009.

[6] J. Kleissl, *Solar Energy Forecasting and Resource Assessment*. Elsevier, 2013.

[7] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," *Prog. Energy Combust. Sci.*, vol. 39, no. 6, pp. 535–576, Dec. 2013.

[8] "Global Energy Forecasting Competition 2014, Probabilistic solar power forecasting." [Online]. Available: http://www.crowdanalytix.com/contests/global-energy-forecasting-competition-2014.

[9] "Many fields have seconds in their units e.g. radiation fields. How can instantaneous values be calculated?" [Online]. Available: http://www.ecmwf.int/en/many-fields-have-seconds-their-units-eg-precipitation-and-radiation-fields-how-can-instantaneous.

[10] E. Lorenz, T. Scheidsteger, J. Hurka, D. Heinemann, and C. Kurz, "Regional PV power prediction for improved grid integration," *Prog. Photovoltaics Res. Appl.*, vol. 19, no. 7, pp. 757–771, 2011.