Solar Power Probabilistic Forecasting by Using Multiple Linear Regression Analysis

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

Badrul Chowdhury
Deparmtent of Electrical and Computer Engineering
University of North Carolina at Charlotte
Charlotte, USA
Email: b.chowdhury@uncc.edu

***Abstract*—The decreasing capital cost and the advancements of technology led to growing up the deployment of renewable energy resources. The variable renewable generations such as wind and solar energy brought some more challenges to the electrical systems. The variable generation forecasting can be used to mitigate some of these challenges. Wind power forecasting has already embraced advanced research efforts but solar power forecasting is witnessing an ongoing research. This paper proposes a multiple linear regression analysis model build in SAS Enterprise Guide to generate the probabilistic forecasts of the solar power and implemented in global energy forecasting competition 2014.**

I. Introduction

In recent years, the rapid boost of variable generations of wind and solar power into the electrical generation systems causes to present the generating power besides the load to be the main reasons behind the variability that impacting the entire electrical power balance in the grid. The generation and demand balance is required in the economic scheduling of the generating units and the electricity markets trades. When the penetration level of the variable generations is high, the intermittency of these resources becomes important to the electric grid. Thus, wherever the variable generations resources are used, the operating reserve and efficient storage systems are inevitably needed to keep the power balance of the system. The operating reserve which is using the fossil fuel generation units should be kept as low as possible to get the highest benefit from the deployment of the variable generations [1]. Therefore, the forecasting of these renewable resources becomes a vital tool in the operation of the power systems and electricity markets.

A. Variable Generation Forecasting

The practical implementation of variable generations forecasts started in late 1990s, when rapidly increasing levels of wind power penetration took place into some national electric grids in Europe, such as in Denmark and Germany [2]. Currently U.S. is witnessing continuous increase of renewable portfolios in some states, such as California and Texas. So that the renewable energy forecasting is also be critically needed, since these accurate prediction tools would lead to scheduling and dispatching the conventional power plants more efficiently and improving the performance of the entire electrical system [3]. The updating in the policy, such as FERC Order 764 [4], which requires ISOs/RTOs to offer intra-hourly scheduling, helps to growing up the penetration level of the renewable resources plus to introduce more flexibility for the implementation of the variable generations forecasting into power systems operations as well as with the trading of electricity marketing [5]. The timeline and operating mechanism of the variable generation forecasting are different between ISOs and electric utilities. Forecasting timeframe changes according to the scheduling and market operating intervals. For instance, NYISO redispatches the entire system every five minutes, which reduces the variability of the wind resources from one operating interval to the next. While the day-ahead (DA) forecasting is used for reliability assessment and allows NYISO of making day-ahead unit commitment decisions. The real-time forecasts are used in NYISO's real-time security-constrained economic dispatch [6].

B. Variable Generation Forecasting Models

The forecasting models are continuously being improved to generate more accurate forecasts of solar and wind power.

1) The physical approach model: The physical approach describes the physical relationships between weather conditions, topography, solar irradiation, and the solar power outputs of the plant. The input data of it from different sources, outputs of numerical weather prediction (NWP) models, local meteorological measurements as the sky imagers for tracking the clouds movement, and SCADA (the user) data for the observed output power, and additional information about the characteristics the nearby terrain and topography of the site. Furthermore, the specifications of the solar panels type and inverters. The satellite systems and sky imagers are used for tracking the clouds and forecast the solar irradiance up to 3 hours, further than that NWP is usually used to project the irradiance [7].

2) The statistical approach model: Describes the connection between predicted solar irradiance from NWP and solar power production directly by statistical analysis of time series from data in the past without considering the physics of the system. This connection can be used for forecast in the future plant outcomes.

3) The learning model: It uses artificial intelligence (AI) methods to learn the relation between predicted weather conditions and power output generated as time series of the past. Unlike statistical approach, AI methods instead of an explicit statistical analysis, they use algorithms that are able to implicitly describe nonlinear and highly complex relations between these input data (NWP predictions and output power). For both the statistical and AI approach, long and high quality time series of weather predictions and power outputs from the past are of essential importance.

4) The combined approach: Modern practical renewable power forecasting models are usually a combination of physical and statistical models. Since physical approach need the statistics to adjust for more accurate forecasts, while the statistical approach for better forecasts need the physical relations of output power production. The optimal performance of the combined models is achieved by optimal shifting of weights between physical approach forecasts and the statistical forecasts [8], [9].

II. GLOBAL ENERGY FORECASTING COMPETETION 2014 (GEFCOM2014)

The Global Energy Forecasting Competition is the second of its time, since the first was in 2012. It was for load and wind point forecasting. While the current competition GEFCom2014 has four tracks: Probabilistic Electric Load Forecasting, Probabilistic Electricity Price Forecasting, Probabilistic Wind Power Forecasting and Probabilistic Solar Power Forecasting to provide state-of-the-art techniques of energy forecasting [10].

A. Probabilistc Solar Power Forecasting Objective

In this track of competition, the conants are required to submit the probabilistic forecasts of the solar power in hourly steps through a month of forecasts horizon. It's the track of the competition where the proposed model has been implemented and the data come from.

B. Data Descreption

Each submitted task of probabilistic distribution (in quantiles) of the solar power generation is for 3 solar systems (zones).

1) Zones: The three solar systems are adjacent. These farms are also called zones, they have a relatively low output power capacities. Their installation parameters are:

- Zone1: Altitude (595m) Panel type(Solarfun SF160-24-1M195) Panel number (8) Nominal power (1560W) Panel Orientation (38°Clockwise from North) Panel Tilt (36°).
- Zone2: Altitude(602m) Panel type(Suntech STP190S-24/Ad+) Panel number (26) Nominal power (4940W) Panel Orientation (327°Clockwise from North) Panel Tilt (35°).
- Zone3: Altitude(951m) Panel type(Suntech STP200-18/ud) Panel number (20) Nominal power (4000W) Panel Orientation (31°Clockwise from North) Panel Tilt (21°)

2) Weather variables: The target variable is the solar power. There are 12 NWP independent variables are from European Centre for Medium-Range Weather Forecasts (ECMWF). The ECMWF-NWP output to be used as below:

- 078.128 Total column liquid water of cloud (tclw) - (kg/m^2).
- 079.128 Total column ice water of cloud (tciw) - (kg/m^2)
- 134.128 surface pressure (SP) - (Pa).
- 157.128 Relative humidity at 1000 mbar (r) - (%).
- 164.128 total cloud cover (TCC) - (0-1)
- 165.128 10 metre U wind component (10u) - (m/s).
- 166.128 10 metre V wind component (10V) - (m/s).
- 167.128 2 metre temperature (2T°) - (K)
- 169.128 surface solar rad down (SSRD) - (J/m^2) - Accumulated field.
- 175.128 surface thermal rad down (STRD) - (J/m^2)- Accumulated field.
- 178.128 top net solar rad (TSR) - (J/m^2) -Accumulated field.
- 228.128 total precipitation (TP) - (m) - Accumulated field.

C. Evaluation Creteria

Pin ball loss function is used to evaluate the accuracy of probabilistic forecasts. It's a piecewise linear function is often used to evaluate the accuracy of quintile forecasts [11].

$$L_q(\hat{Y}, Y) = \begin{cases} q(\hat{Y} - Y), & \text{if } Y \leq \hat{Y} \\ (1-q)(Y - \hat{Y}), & \text{if } Y > \hat{Y}. \end{cases} \quad (1)$$

Where $L_q(\hat{Y}, Y)$ is the loss function to the probabilistic forecasts for each hour. $\hat{Y}$ is the forecasted value at the certain q quantile of the probabilistic solar power forecasts and Y is the observed value of the solar power. The quantile q is also called the percentile and it has discrete values [0.01 - 0.99]. $\hat{Y}$ and Y are normalized values of the nominal power capacity of each zone. The average of loss function $L_q(\hat{Y}, Y)$ for all forecasting hours should be minimized to yield more accurate forecasts. Therefore, the average loss function value is adopted as a score to evaluate the model performance.

III. SOLAR FORECASTING MODELING

The basic skeleton of the methodology and theoretical background of building the proposed solar power forecasting model in this paper is adopted from reference [12], where the multiple linear regression (MLR) analysis is deployed for short-term load forecasting. SAS Enterprise Guide® is used as the environment where the forecasting model is built in. SAS is a statistical analysis software is suitable for the proposed statistical forecasting model. The flowchart of the solar forecasting model building steps is shown in Fig. 1.

A. Data Preparation

It 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

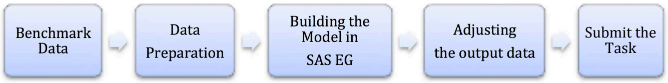


Fig. 1. Flowchart diagram of the solar forecasting modeling

variables for each zone. The data preparation is an important step for treating the data to be ready for the analysis and modeling.

The various steps of the data preparation are shown in Fig. 2. Fig. 3 shows the box plot of the distribution of the historical data (i.e.2012) of solar power for zone1. The other two zones are not different a lot since they are adjacent solar systems. Note, the order of months as it is provided from the competition benchmark data and appears in the box plot does not necessary mean the same order of actual year months.

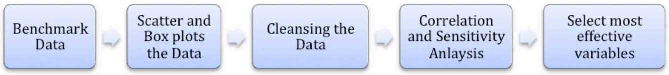

Benchmark Data
Scatter and Box plots the Data
Cleansing the Data
Correlation and Sensitivity Anlaysis
Select most effective variables

Fig. 2. Flowchart diagram of data preparation

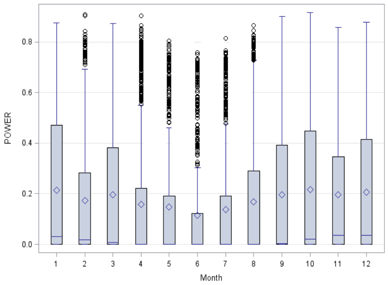
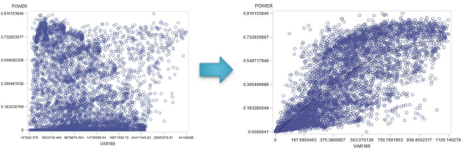


Fig. 3. Box plot of the distribution of observed solar power

The scatter plot is also useful to get 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 since the scatter plots for the observed power with respect to the solar irradiance as it is also called surface solar radiation down (SSRD) for zone1. The scatter plot on the left in Fig. 4 is for the solar irradiance SSRD169.128 as it is provided from the NWP which is in accumulated values (J/m^2), while the plot on the right for the average values of solar irradiance SSRD (W/m^2). The relationships between the variables on the right plot is more obvious and it can tell it is a positive relationship with relatively high positive correlation coefficient. The last four given weather variables (i.e. solar and thermal radiations besides the precipitation) are in accumulated field values not average values. They are increasing in every hour until the end of the day and then start again in accumulation [13]. For getting the average values of these data by applying the formula in equation 2.



169.128 (SSRD) : J/m^2 **Accumulated field (for all day hours)**

169.128 (SSRD) w/m^2 **Average Value (for each hour)**

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

$$Avg(t) = \frac{Acc(t+1) - Acc(t)}{3600} \qquad (2)$$

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

For comparison of the three zones, from Fig 5, zone3 has a strongest positive linear relationship between the solar irradiance and power. The left scatter plot is for zone1.

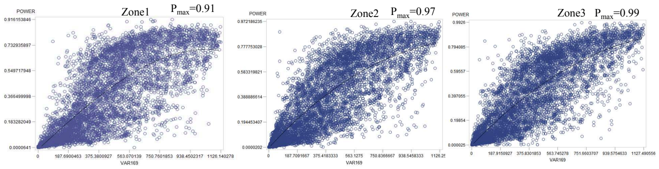


Fig. 5. Scatter plot of the solar irradiance and power for all zones

From the scatter plots, the outliers don't change the general data trends. The majority of the extreme points in the observed data occur at the sun rise and set periods which are the unstable duration of solar panel. By experiment the data cleansing is conducted and led to a tiny improvement in forecasts. This doesn't mean to underestimate the data cleansing stage in data preparation before the modeling since sometimes the outliers could be generated from data entry issues.

B. The Model Building

The main steps of building the forecasting model is shown in Fig 6. ANOVA is the analysis of variance and it's available as a statistical tool in SAS where the multiple linear regression analysis is running.

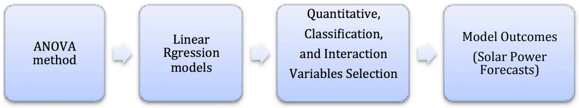

ANOVA method
Linear Rgression models
Quantitative, Classification, and Interaction Variables Selection
Model Outcomes (Solar Power Forecasts)

Fig. 6. Flowchart diagram of the model building

The predictor variables selection for forecasting the solar power as independent or response variable by plotting is cumbersome since there is twelve predictor variables to choose form. So the correlation and sensitivity analysis is carried out for the historical data in SAS to investigate the most effective variables. Fig 7 is the outcomes of the correlation analysis. It's obvious the solar irradiance, in addition to the time (hours), the surface irradiance (VAR169) and net top solar irradiance

(VAR178) and their second order polynomial or quadratic terms have the highest correlation with solar power. The relative humidity (VAR157) has some impact on solar power since the water vapor particles by chance work as small mirrors and increasing the reflection level of the solar irradiance and hence the generated solar power. The temperature at 2m variable (VAR167) is also has an noticeable impact compared to other less effective variables. It is reasonable since the solar panel characteristics are affected by the ambient temperature. In fact, the temperature plus to the solar irradiance are chosen in physical-approach models to forecast the power from PV systems [14].

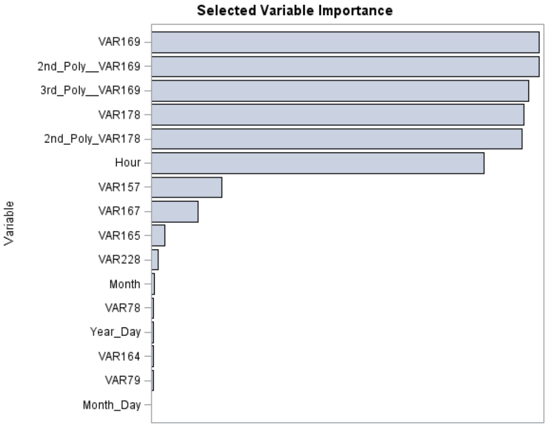


Fig. 7. The order of most effective variable for solar power forecast

The multiple linear regression MLR model can be represented as in equation 3.

$$Y = \beta_o + \beta_1 X_1 + ... + \beta_k X_k + \epsilon \tag{3}$$

Where Y is the response variable and X_k is the k predictor variable. βs are the parameters and ϵ is the error or the variations in Y. The response variable is the solar power, and the predictor variables are selected from the most effective weather variables [12].

In multiple linear regression analysis model, the ability of using the interactions between the variables adding more power to the model to produce more accurate forecasts. In addition, MLR analysis model includes the probability of the forecasts as a prediction interval. The prediction interval includes the variance of the error as well as the variance of the parameter estimates.

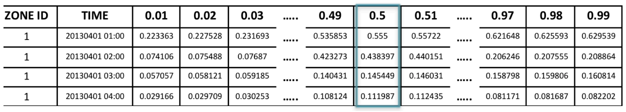
Since the sun azimuth angle changes periodically through the day hours and the sun elevation changes along the seasons [15]. Therefore, the hours and month days and months are considered as classification variables that the chosen predictor variables interact with.

It's worth to mention, that the variables selection is done based on each separated variable alone, but when the variables interact with each other in the MLR analysis model they could loose some correlation power. Therefore, in the total mix of selected variables, the best candidate model is also needed. After the variables selection and their interactions are investigated, the selection of the candidate model of the most efficient performance is found manually by carrying out the three main steps of building the model, training, validation, and testing.

IV. Results

The sample probabilistic forecasts of the solar power for each hour and for all three zones or solar systems are as in Table I. The forecasts in median column ($q = 0.50$) is the point forecast which should be as close as possible to the actual observed power.

TABLE I
Sample of probabilistic forecasts for some hours of zone1



ZONE ID	TIME	0.01	0.02	0.03	.....	0.49	0.5	0.51	.....	0.97	0.98	0.99
1	20130401 01:00	0.223363	0.227528	0.231693	.....	0.535853	0.555	0.55722	.....	0.621648	0.625593	0.629539
1	20130401 02:00	0.074106	0.075488	0.07687	.....	0.423273	0.438397	0.440151	.....	0.206246	0.207555	0.208864
1	20130401 03:00	0.057057	0.058121	0.059185	.....	0.140431	0.145449	0.146031	.....	0.158798	0.159806	0.160814
1	20130401 04:00	0.029166	0.029709	0.030253	.....	0.108124	0.111987	0.112435	.....	0.081171	0.081687	0.082202

Since after submission the forecasts for a task, then in the following task the actual data for these forecasts is provided, the plots can be used as a validation of the model performance. As it is shown in Fig 8 the actual and the point (median) forecasts are plotted with residual plot. The residuals plot has positive and negative values these mean not over-forecasting nor under-forecasts impacts, so the model is not biased. Since the forecasts is probabilistic, the box plots for the distribution of the actual and forecasted power is also shown for each day hour.

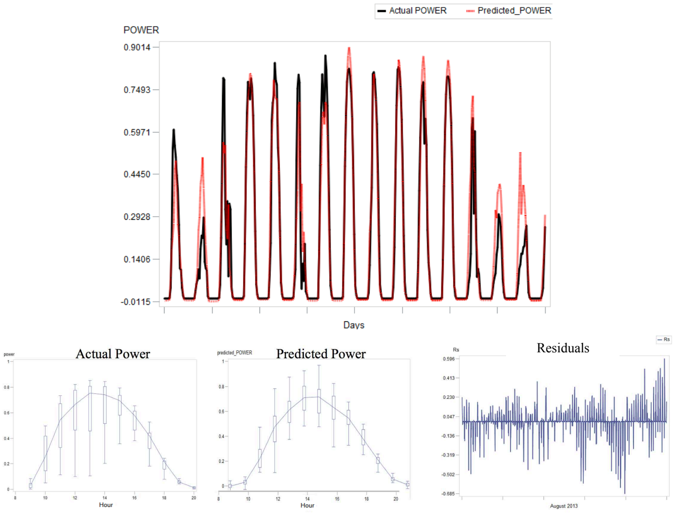


Fig. 8. The results plots for solar power forecasts of the model

The pinball is also built and used for evaluate the mode as well as the root mean square error (RMSE) between the point power forecasts and actual solar power. Table II presents the pinball loss function value and RMSE for each zone and for the entire aggregated system of the three zones.

TABLE II
SEPTEMBER 2013 FORECASTS EVALUATION

Zone	Pinball	RMSE
1	0.0130	0.0725
2	0.0140	0.0741
3	0.0141	0.0742
Aggregated	0.0137	0.0736

The line plot is shown in Fig. 9 for the actual solar power and its corresponding forecast for the median $q = 0.5$ and both the lower and upper limits of prediction interval if the the prediction interval is set up to 95% this makes the prediction interval contains 95% of the observed solar power for each hour. The other quantiles forecasts ($q = 0.02$ to 0.49 and $q = 0.51$ to 0.98) for each hour are distributed linearly between the upper and lower limits ($q = 0.01$ and $q = 0.99$) of the prediction interval.

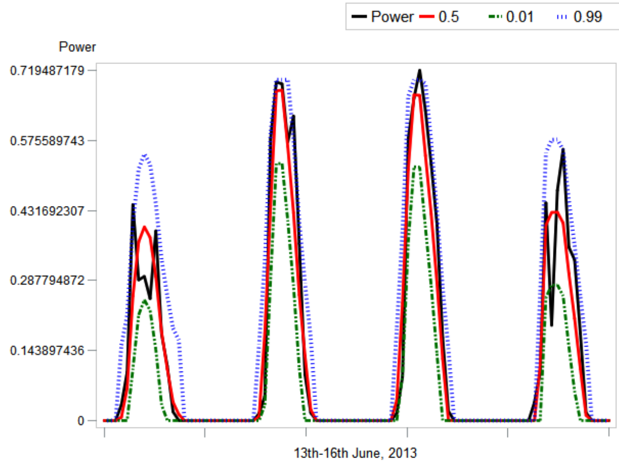


Fig. 9. The line plots for actual solar power and the forecasts where $q = 0.01, 0.5, 0.99$

Meanwhile the scatter plots of the same quantiles for zone1 is shown in Fig. 10. From line and scatter plots, it's obvious that the probabilistic forecasts produced by the model covers the actual solar power.

When the point forecasts ($q = 0.50$) are drawn with respect to the observed power, Pearson correlation coefficient is found 0.9414, this is represented by scatter plot in Fig. 11.

The candidate model is as following: CLASS = Month "Month Day"n Hour; MODEL POWER = VAR167 VAR169 VAR175 VAR228 "2VAR169"n Month Hour VAR79 VAR157 "3VAR169"n VAR169*Hour VAR169*Month "Month Day"n*VAR169 "2VAR169"n*Month "2VAR169"n*Hour VAR167*VAR169*Hour VAR169*VAR175*Hour VAR169*VAR157*Hour VAR169*VAR228*Hour VAR169*VAR79*Hour "Month Day"n "2VAR169"n*"Month Day"n Trend "4VAR169"n "5VAR169"n "Year Day"n "4VAR169"n*Month "4VAR169"n*Hour "4VAR169"n*"Month Day"n VAR78 VAR134

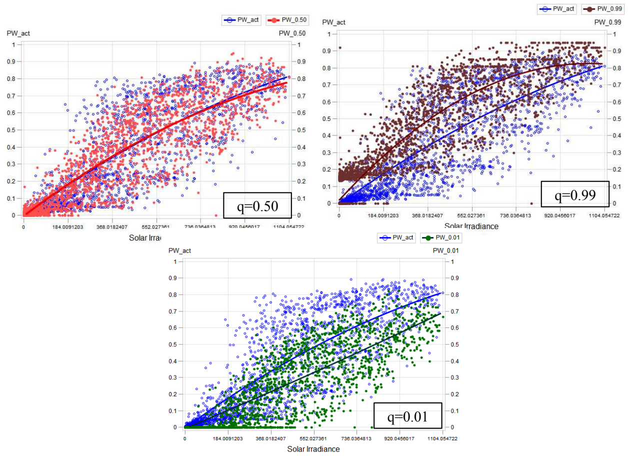


Fig. 10. The scatter plots for actual solar power and the forecasts where $q = 0.01, 0.5, 0.99$

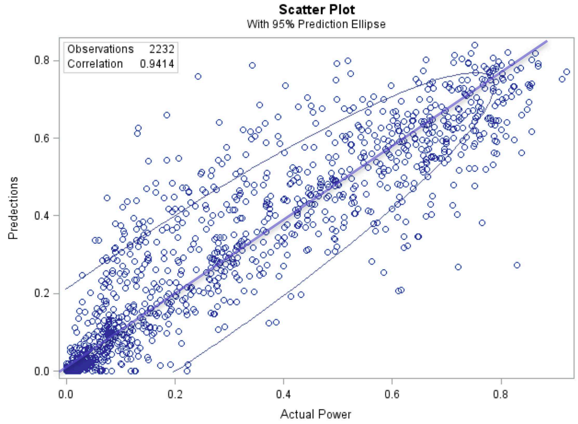


Fig. 11. Scatter plot of the point forecasts and the actual solar power

VAR164 VAR165 VAR166 VAR178 VAR178*Hour VAR178*Month "Month Day"n*VAR178 VAR79*Hour VAR228*Hour VAR157*Hour VAR175*Hour VAR167*Hour VAR134*Hour VAR78*Hour VAR164*Hour VAR165*Hour VAR166*Hour "3VAR169"n*Hour "3VAR169"n*Month "3VAR169"n*"Month Day"n "5VAR169"n*Hour "5VAR169"n*Month

V. CONCLUSION

The adopted multiple linear regression analysis model performed well for solar power forecasts. If the forecasting hours were with clear sky, the model's performance would be better for near forecasting horizon than farther horizon but this is affected by cloudy hours and so the entire performance of the model. Plotting the data, correlation and sensitivity analysis between the variables, as well as data cleansing of outliers if there any are essential data preparation steps before building the forecasting model. For additional historical data the model performance will be improved. By experiment, found the quadratic term of the solar irradiance is doing better job than the cubic term and the natural logarithm term. Since the hours selected as a classification variables, deleting the night hours from data reduces the model's performance.

ACKNOWLEDGMENT

The authors would like to thank Dr. Tao Hong at University of North Carolina at Charlotte and Chair of Global Energy Forecasting Competition 2014 for his cooperation and valuable feedbacks.

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] L. Landberg, A. Joensen, G. Giebel, S. Watson, H. Madsen, T. Nielsen, L. Laursen, J. Jorgensen, D. Lalas, and M. Trombou, "Implementation of short-term prediction," in *EWEC-CONFERENCE-*, pp. 57–62, 1999.

[3] R. Widiss and K. Porter, "Review of variable generation forecasting in the west: July 2013-march 2014," tech. rep., National Renewable Energy Laboratory (NREL), Golden, CO., 2014.

[4] K. Takeaways, "Ferc order 764 and the integration of renewable generation," *Policy*, 2014.

[5] M. MAkhyoun, "Predicting solar power production," 2014.

[6] W. Grant, D. Edelson, J. Dumas, J. Zack, M. Ahlstrom, J. Kehler, P. Storck, J. Lerner, K. Parks, and C. Finley, "Change in the air," *Power and Energy Magazine, IEEE*, vol. 7, no. 6, pp. 47–58, 2009.

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

[8] U. Focken and M. Lange, *Physical approach to short-term wind power prediction*. Springer, 2006.

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

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

[11] J. M. Morales, A. J. Conejo, H. Madsen, P. Pinson, and M. Zugno, *Integrating Renewables in Electricity Markets*. Springer, 2014.

[12] T. Hong, *Short term electric load forecasting*. North Carolina State University, 2010.

[13] "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.

[14] E. Lorenz, T. Scheidsteger, J. Hurka, D. Heinemann, and C. Kurz, "Regional pv power prediction for improved grid integration," *Progress in Photovoltaics: Research and Applications*, vol. 19, no. 7, pp. 757–771, 2011.

[15] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772–1783, 2009.