Forecasting Error Modelling for Improving PV Generation Prediction

Happy Aprillia[1] Hong-Tzer Yang[2]
[1]Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Cheng Kung University, Tainan. Email: n28057049@mail.ncku.edu.tw
[2]Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Cheng Kung University, Tainan. Email: htyang@mail.ncku.edu.tw

Abstract
Accurate forecasting of Photovoltaic (PV) generation output is important in operation of high PV-penetrated power systems. In this paper, an adaptive uncertainty modelling method for forecasting error is proposed to improve the prediction accuracy of PV generation. The proposed method models the uncertainty in forecast data using Kernel Density Estimator and guarantee the provision of accurate expected value. Neural Network model is then constructed by the developed uncertainty model to forecast the PV output. The actual confidence level is traced within the day and injected as an input to the Neural Network model by observing the Mean Absolute Prediction Error (MAPE) and Unscaled Mean Bounded Relative Absolute Error (UMBRAE). The proposed method is tested with various significant changes of weather condition and proved to have promising performance on PV generation forecasting. Thus, the developed adaptive uncertainty model can be further used in power system planning that have high-penetration energy sources with stochastic behavior.

Keywords: uncertainty model, PV generation prediction, forecast error modeling

Abstrak
Peramalan yang akurat dari keluaran Photovoltaic (PV) penting dalam pengoperasian sistem tenaga dengan penetrasi PV yang tinggi. Dalam makalah ini, metode pemodelan ketidakpastian adaptif untuk peramalan galat diusulkan untuk meningkatkan akurasi prediksi keluaran PV. Metode yang diusulkan memodelkan ketidakpastian dalam data perkiraan menggunakan Kernel Density Estimator (KDE) dan menjamin penyedian hasil prediksi yang akurat. Model Jaringan Syaraf Tiruan kemudian dibangun dengan model ketidakpastian yang dikembangkan untuk meramalkan keluaran PV. Tingkat kepercayaan aktual dilacak di siang hari dan disuntikkan sebagai masukan pada model Jaringan Syaraf Tiruan dengan mengamati Mean Absolute Prediction Error (MAPE) dan Unscaled Mean Bounded Relative Absolute Error (UMBRAE). Metode yang diusulkan diuji dengan berbagai perubahan kondisi cuaca yang signifikan dan terbukti memiliki kinerja yang menjanjikan pada peramalan PV. Dengan demikian, model ketidakpastian adaptif yang dikembangkan dapat digunakan lebih lanjut dalam perencanaan sistem tenaga yang memiliki sumber energi penetrasi tinggi dengan perilaku stokastik.

Kata Kunci: model ketidakpastian, peramalan, prediksi galat

1. Introduction

As the effect to achieve sustainable development in developing county, penetration of renewable energy source (RES) is increasing significantly up to 20% of country's total energy mixture. Handling this RES in the system means to handle uncertainty in operation. For Power System Operator (PSO), RES injects stochastic power to the system which need some spare of reserve if the actual output of RES falls outside the uncertainty range. The greater the uncertainty leads to a greater spare capacity which is increasing the cost of electricity generation. Thus, providing uncertainty modeling will give PSO an insight to handle stochastic behavior of RES even though it will not necessarily solve uncertainty in the analysis because it still contains forecast error from the actual data.

To provide reliable strategy, PSO needs to foresee the system's condition in next time horizon (hour, day, week, month or even year). However, the forecasting result will be polluted by prediction error. Especially, when prediction only based on short range data. By this, PSO will see this forecast-related strategy as a less favorable option and go with the old conventional strategy. In fact, there are several techniques to have accurately forecast such as Artificial Intelligent (AI), Evolutionary Algorithm (EA) and Statistical Method (SM). Neural Network (NN) is one of well-known AI to forecast load demand

or RES, but it needs long historical data to build a proper model. This leads to an inflexible model that need to change when the input actual data much differs from the trained data. EA such as Particle Swarm Optimization, Genetic Algorithm, Bat Algorithm has also proved to handle forecasting problem but they need an iterative process to find best single objective or multi objectives. In a practice with high renewable energy penetration, EA (Hansen *et al,* 2009) will be time-consuming and provide worse prediction because of input data doesn't remain constant. Meanwhile, the statistical method with regression as the base of error between each data point also has some drawback. The prediction will fail when the system is nonlinear. Besides, SM needs at least 7 days 15-minutes data as training data which also face the similar inflexible issue as NN when the data changes dynamically.

Uncertainty modeling has been applied in many works such as power system reserve planning, chemical solution process, combustion chamber process and medical research by two distinct approaches: simulation technique and statistical analysis. The most reliable simulation technique is done by Monte Carlo Simulation (MCS). MCS will generate some scenarios (several hundred or thousand) from random number following designated probability distribution and find the best and worst scenario to be handled by an operator. (Nikoobakht et al, 2017) presents the stochastic wind power generation with possible scenarios generated by MCS. Wind power uncertainty is assumed as discrete distribution in the form of Weibull distribution and continuous probability distribution functions for security constrained unit commitment problem. (Ding *et al*, 2017) performed Monte Carlo simulation and used confidence level to interpret PV's generation uncertainty. Confidence interval of forecasted value in 10,000 generated scenarios. The result shows that by modeling the uncertainty, less PV generation is curtailed and fewer network losses are achieved in comparison with a deterministic approach.

While most of the methods used to have initial probability distribution and remains constant till the end of simulation, there are several distinct methods emphasized the importance of adaptability. (Huang *et al*, 2012) used analytical method to model uncertainty in basic system load to determine electricity price. This paper emphasized the needs of proper proposal distribution to find correct target distribution for probabilistic value of system load. Even though it is time efficient, this method failed to operate in high probability event due to its algorithm. The complicated ways to have sequential and adaptive learning to determine the importance region will trapped the solution into the incorrect solution when the importance region is slightly shifted. (Aien *et al,* 2014) used the analytical method due to probability distribution of Solar Cell Generator (SCG) and Wind Turbine Generator (WTG). The algorithm is worked by assumption of using beta distribution for SCG and Weibull distribution for WTG. The author emphasized the importance of correlation between uncertain variables in the system. However, this correlation is not well expressed. (Negnevitsky et al, 2015) used normal distribution to describe wind power generation uncertainty even though the wind power generation prediction errors do not fit a normal distribution. Thus, this can affect the accuracy of the result if improper error distribution is used.

It can be concluded that, for modeling uncertainty, both simulation and statistical method need to use proper initial probability distribution function (PDF), meanwhile, the real-time probability distribution might not always the same with common PDF such as normal distribution, Weibull distribution, beta distribution etc. In MCS, the range of uncertainty is controlled by the degree of confidence level which means all observation point will vary to the same uncertainty degree. There is no adaptive ability to sense the actual uncertainty level. In statistical analysis, for some case with importance-relation matrix in between uncertain variables, the method is prone to inaccurate results if the initial importance or quantile is incorrect. Thus, proper modeling of uncertainty can be considered not only from understanding input data characteristics such as standard deviation, proper probability distribution, and quantile (time interval) but also adaptability to sense real time uncertainty and further prepare corrective-ability. Thus, this work will consider the adaptive correction of prediction error that used only day ahead prediction and real-time data for uncertainty modeling. By this, the proportion of uncertainty in the system will be easily investigated for further used in every forecasting method. The rest of this paper is organized as follows. Section 2 will present steps of proposed method. The result and discussion will be described in Section 3. The conclusion will be presented in Section 4.

2. Methods

The non-parametric distribution function will be used to model the uncertainty, overcoming the data shortage (likely 24 hourlies previous day data). By non-parametric method, the analysis can sense the variation of the model if any change happened in the actual metering. Kernel distribution (Bowman and Azzalini, 1997) will adapt the change in the data and generate a PDF by a Kernel Density Estimator (KDE). The smoothness of density curve can be controlled from its bandwidth value h correlate to n sample size and $K(\cdot)$ kernel smoothing function. The KDE can be described as:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right); \quad 0 < x < \infty \qquad (1)$$

where $\hat{f}_h(x)$ is a prediction function of KDE with time series input x. For the proposed application, KDE will be immune to the data loss because of discretization. KDE will provide smooth and continuous probability curve of the sample data. Furthermore, when there are several uncertainties later involve in the system, KDE can have different individual distribution for each component of uncertainties, summing their smooth curves and provide one single continuous probability density function. From (Greco and Pagnotta, 2009), skewness of probability distribution function (λ) relates to standard normal density φ and distribution function Φ that can be described as:

$$g(z;\lambda) = 2\varphi(z)\Phi(\lambda z); \quad \lambda \in IR \qquad (2)$$

Thus, equation (1) will be modified as:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K \frac{g\left(\frac{x - x_i}{h} - \lambda\right)}{G(h^x - \lambda)} \qquad (3)$$

The calculation of the real confidence interval of uncertainty will be done by the following approach.

1. Find the minimal error of actual interval by differentiating the actual and forecast temperature per hour. The minimal error will be placed when the forecast and actual temperature curve is intersected)
   $$P_1 = P_2 \qquad (4)$$

2. The probability of actual condition will follow KDE mentioned in Eq. (1), and joint probability will be calculated as:
   $$p(x) = \frac{P_1}{\sqrt{2\pi}\sigma_1} e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{P_2}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \qquad (5)$$

3. Adopted from the way adaptive threshold in image processing (Gonzales & Woods, 2002), The confidence interval will be as follows:
   $$AT^2 + BT + C = 0 \qquad (6)$$
   $$A = \sigma_1^2 - \sigma_2^2 \qquad (7)$$
   $$B = 2(\mu_1\sigma_2^2 - \mu_2\sigma_1^2) \qquad (8)$$
   $$C = \sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2 + 2\sigma_1^2\sigma_2^2 \ln\left(\frac{\sigma_2 P_2}{\sigma_1 P_2}\right) \qquad (9)$$

This probability distribution function will be used in day ahead forecasting on photovoltaic Alternating Current (AC) power production. Meanwhile the available forecast data usually only consist of 1 point for 1 day and 7 points for next 7 days, Back Propagation Neural Network (NN) is used to determine the prediction accuracy of uncertainty of PV output. There are three inputs in the training process which are forecast temperature, actual site's temperature and past day PV output. Forecast temperature are scraped from Central Weather Bureau of Taiwan (CWB). The CWB data is usually 7 days ahead prediction in a form of prediction range (following normal distribution with a

confidence level). Data of actual site's temperature and PV output are gathered from PV plant in Kaohsiung. This location is selected because it matched with the latitude and longitude of CWB's weather station. The actual temperature data will be 24 point hourly online data and so does the PV output. The difference is in PV output data, historical data of past day is used. These three inputs will be feed to NN and predict the probability distribution of PV output. By this process, the confidence level of next day and 7 days can be accurately predicted. From (Chen *et al,* 2017) mentioned in equation (10) and (11), Mean Absolute Percentage Error (MAPE) and Unscaled Mean Bounded Relative Error (UMBRAE) will be used to calculate prediction error. The proposed strategy is shown in Figure 1.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Actual - Forecast|}{|Actual|} \times 100\% \qquad (10)$$

$$UMBRAE = \frac{MBRAE}{1 - MBRAE} \qquad (11)$$

Where the Mean Bounded Relative Error (MBRAE) is

$$MBRAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Actual - Forecast|}{|Actual - Forecast| + |Actual_t - Actual_{t-1}|}$$

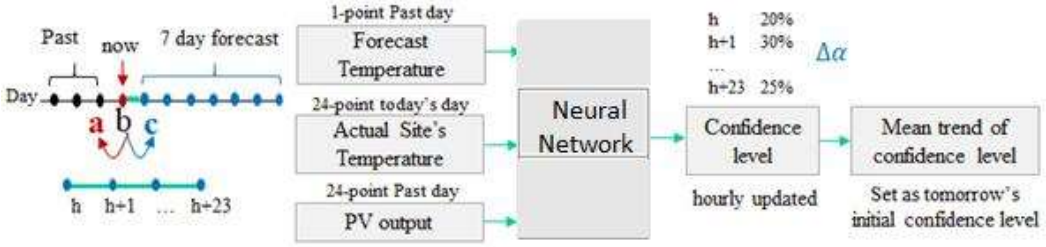

Figure 1. Proposed Strategy for Day Ahead PV Output Forecasting

3. Result and discussion

The proposed method is evaluated for period between October 31st, 2016 to November 2nd, 2016. The input data will be from the historical data of PV plant in Kaohsiung and weather data from Central Weather Bureau of Taiwan at the same period. The 7-day ahead prediction will also use to be compared with the historical data and further used to know the real confidence error. Mismatch between actual and forecast value will be used to find confidence interval.

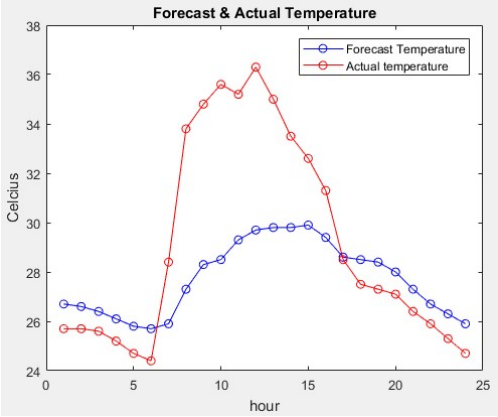

Figure 2. Actual & Forecast Temperature Mismatch on November 1st, 2016

To see the probability distribution within the day, the real value of the mismatch is extracted to Kernel's PDF as shown in Figure 2. From this figure, the forecast temperature is following normal distribution which much differ to actual temperature. To ensure whether the data is following normal distribution or not, the data is plot in the Quantile-Quantile Plot (QQ-plot). The red line shown is the reference of normality. The blue dots following the normality means that the data is closer to normal distribution. In Figure 3, the forecast data is more likely closer to the red line in comparison to actual temperature. To get accurate prediction, the actual probability distribution is used as reference point to calculate real confidence interval.

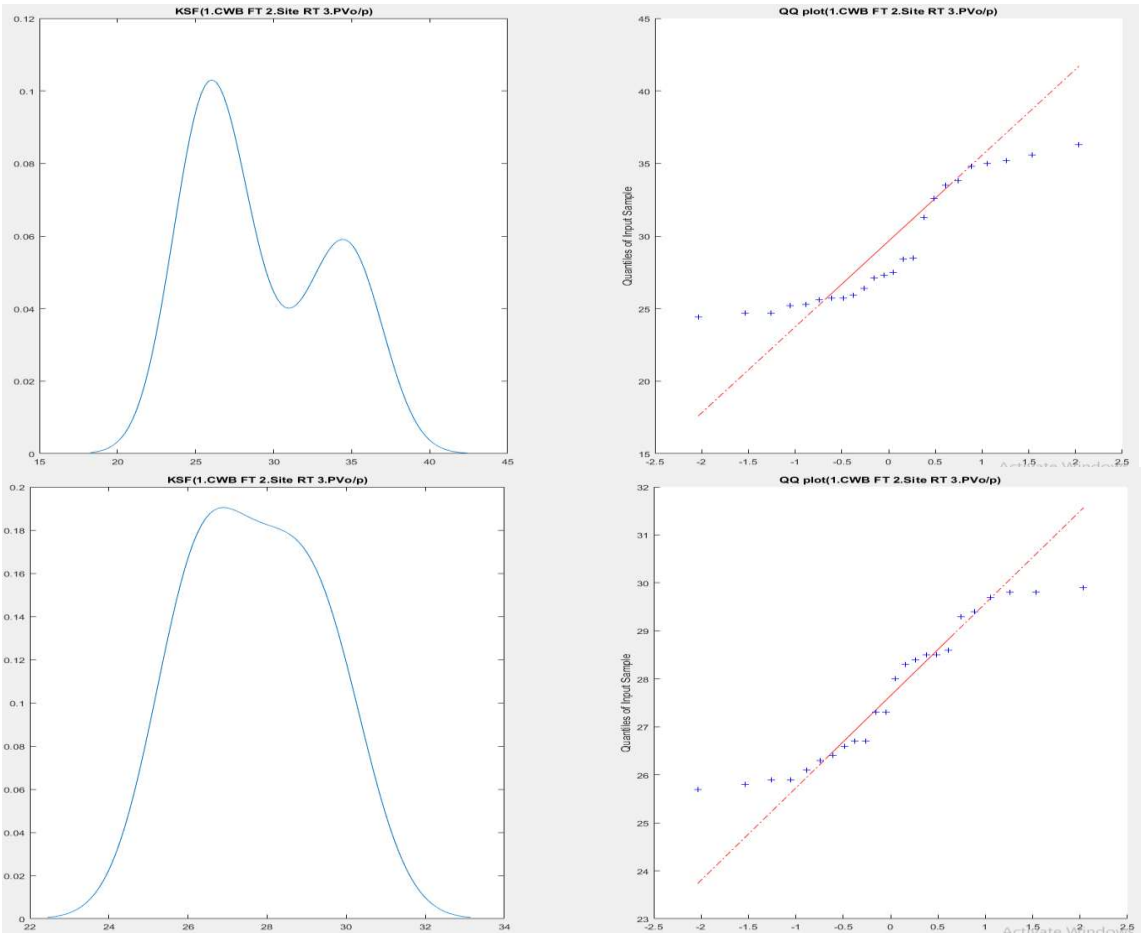

Figure 3. Probability Distribution Function of Actual & Forecast Temperature (Upper and Lower)

To conduct the analysis, the forecast value that available in CWB is usually in the form of points. It is only provided as the lowest and highest prediction of the day following normal distribution. To see how differ the probability of actual and forecast value. these two points are plotted in Figure 4 following the actual probability distribution of previous day. It is shown that there is a mismatch in the temperature's lower and upper bound. Thus, this temperature mismatch will be used to calculate the skewness of the PDF. The PV's output is shown in the Figure 4 (Right).

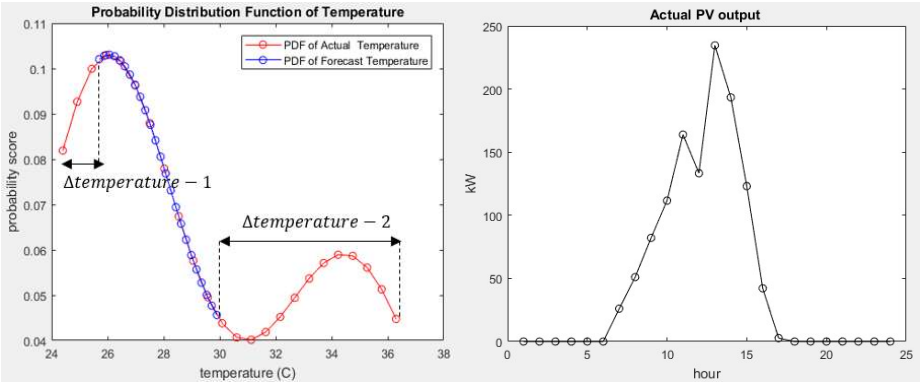

Figure 4. Forecast Temperature in Kernel's Distribution Estimator (Left) and PV's output on November 1st, 2016 (Right)

To forecast PV's output, Back Propagation Neural Network model is used with two inputs which is forecast temperature and Actual PV output respectively. The data of forecast temperature are from CWB's site measurement while the data of actual PV's temperature are from the PV's site. The hidden layer is set as 5 layers and Levenberg-Marquardt optimization is used as a network training function to update the weight and bias values of the model. the performance of NN training is evaluated by Mean Squared Error (MSE) with random data division. After 10 iterations and 6 times validation checks, the results from MAPE prediction are shown in Figure 5. In Figure 5 (right), it can be observed that MAPE of 24-hour forecast is 21.74% with the highest error happened in hour-8 by the prediction's mismatch of 49.23 kW when the real output shows 51 kW. The observation needs to be done carefully because the even though the gap between prediction curve and real output looks wide, if the original output have relatively small value, the error will increase drastically in comparison to high value original output. The simulation in Figure 5 (left) used the historical data to forecast. This simulation will be used as reference that will be compared with the NN's result when probability distribution is used.

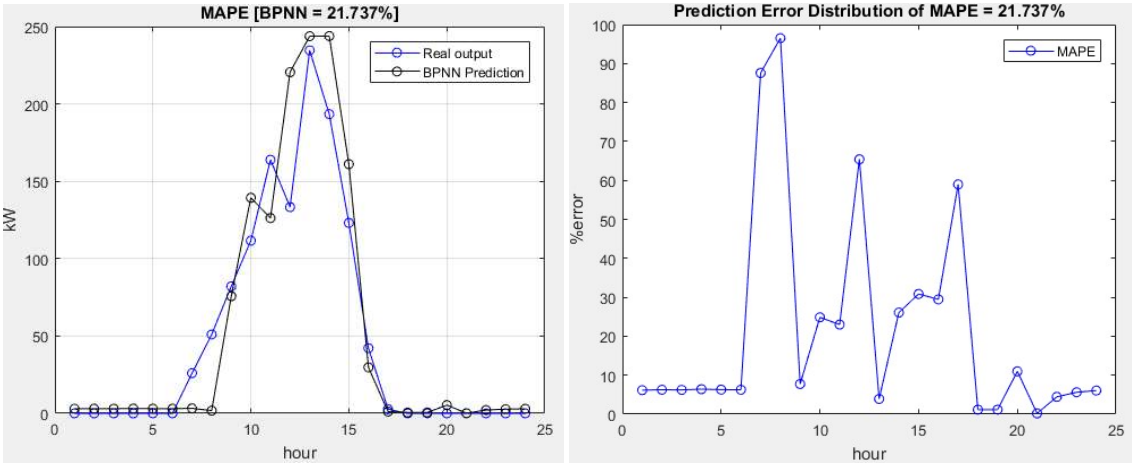

Figure 5. MAPE of PV's Real Output and BPNN's Prediction (left) and MAPE Distribution (Right)

For observing how significance the selection of appropriate PDF to the forecasting output, the NN model is trained and tested with temperature model of two different probability distributions which are normal distribution and Kernel's Density Estimator shown in Figure 6. Both PDF are normalized into [0,1] to have even comparison. Both PDF are extracted from the actual past day historical data which has $24°C$ and $36°C$ as the lowest and highest temperature of the past day. The mean of temperature of normal distribution is $30.25°C$ whose probability score is 1. The higher the probability score means that the temperature have higher possibility to be happened. While in the KDE, there are two

peaks whose highest probability score are **29°C** and **32°C**. The range of this PDF will be modified as the next day temperature prediction.

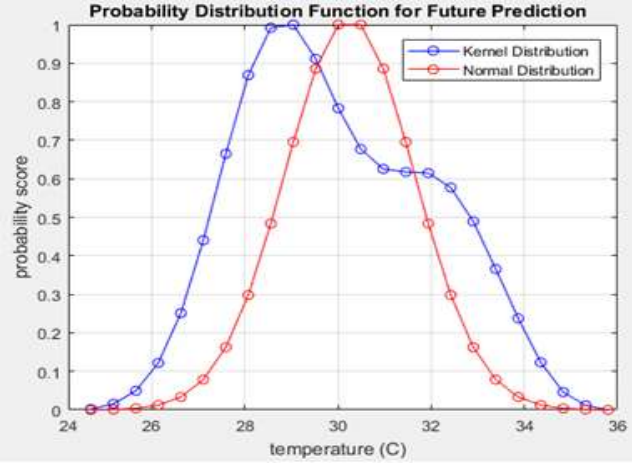

Figure 6. Probability Function for Next Day Prediction

Prediction error is evaluated by MAPE and UMBRAE calculation. In Figure 7, the comparison of forecast result between the usage of KDE and normal distribution as probability distribution is presented. By modelling the temperature with normal distribution, the MAPE reach 46.28% with the highest error is located at hour-8. In contrasts, error of KDE probability distribution at the same point was only 25% which is approximately only 10.63% of MAPE error at hour-8 following normal distribution. By using KDE probability distribution, the MAPE can be suppressed to 21.41% which is 53.74% error reduction. When doing the simulation, authors find that the MAPE error somehow elevated the error calculation when the actual value has very small value near to zero. Thus, UMBRAE will also be used to prove the evaluation.

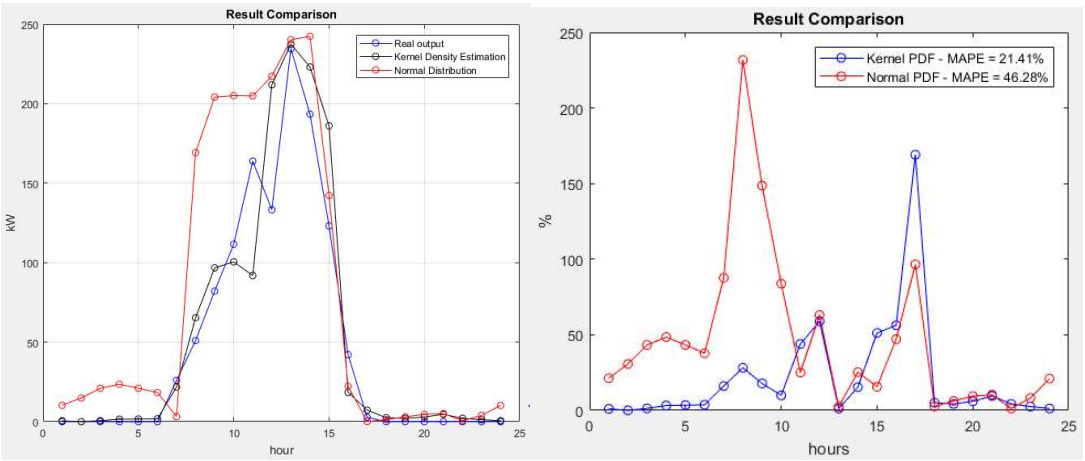

Figure 7. Prediction Comparison between Kernel's PDF and Normal's PDF (left) and Error Comparison using MAPE (right)

In Figure 8, UMBRAE is employed to evaluate the performance of forecast method. The MBRAE of each hour are plotted and later used equation (11) to calculate the unscaled performance of MBRAE. For the following point, errors are bounded into [0,1]. When the MBRAE=1, it means that the actual value is nearly zero while when the MBRAE ≤ 0, the mismatch of the prediction and the actual value is nearly zero which is very accurate. In Figure 8, prediction error that using normal distribution shows 2.64% UMBRAE with the highest error is at point hour-8. Meanwhile, by using KDE, prediction error can be reduced to 2.06% which is 21.97% reduction in UMBRAE Error.

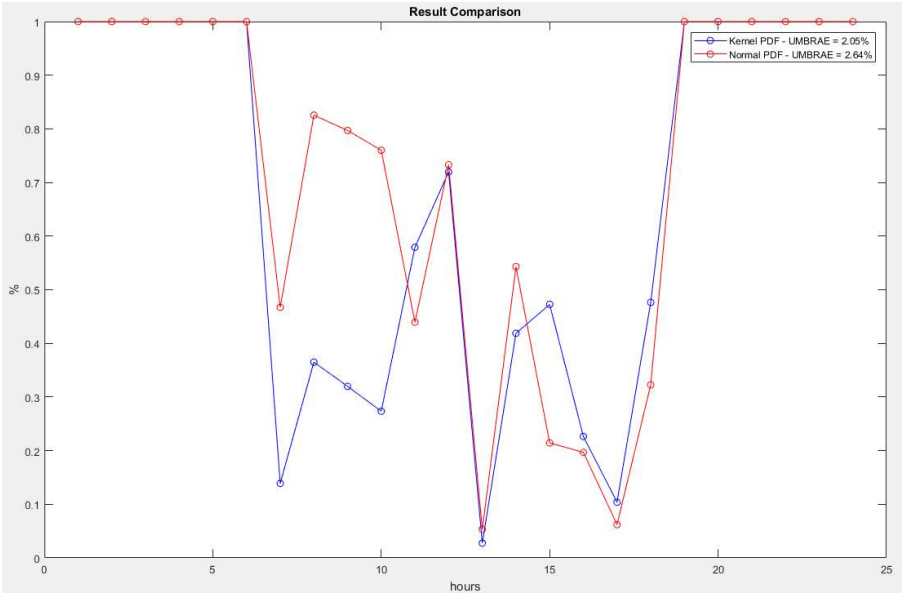

Figure 8. Error Comparison using UMBRAE

4. Conclusion

Several conclusions can be made from the discussion of simulation and result. Firstly, by proposed strategy, the importance of long historical data can be withdrawn by using only previous day to extract its probability distribution function. This will benefit the field's implementation of PV output prediction when the long historical data are not available. Secondly, the accuracy of the forecast is determined by its input. It is necessary to observe not only the pattern of the input data but also its range thus active modification of its probability distribution is needed. The proposed method can adjust the mismatch between the actual measurement and prediction that reduced the prediction error significantly. In the future, proposed strategy will be directly integrated to prediction tool's model structure to further reduce its error prediction.

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan, under Grants MOST 106-3113-E-006-010. Authors would like to thank Indonesia Endowment Fund for Education (LPDP) for funding her study in NCKU.

References

Aien M., Fotuhi-Firuzabad M. and Rashidinejad M. (2014) 'Probabilistic Optimal Power Flow in Correlated Hybrid Wind–Photovoltaic Power Systems," in IEEE Transactions on Smart Grid, vol. 5, no. 1, pp. 130-138.

Bowman A. W. and Azzalini A. (1997). 'Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations', Oxford: Clarendon Press.

Central Weather Bureau (2017), '7-day Forecast', Central Weather Bureau, available from http://www.cwb.gov.tw/V7e/forecast/. Accessed on August 28th, 2017 at 5:32 PM.

Chen C., Twycross J., Garibaldi J.M. (2017). 'A new accuracy measure based on bounded relative error for time series forecasting,' PLoS ONE 12(3): e0174202.

Ding T., Li C., Yang Y., Jiang J., Bie Z. and Blaabjerg F. (2017) 'A Two-Stage Robust Optimization for Centralized-Optimal Dispatch of Photovoltaic Inverters in Active Distribution Networks,' in IEEE Transactions on Sustainable Energy, vol. 8, no. 2, pp. 744-754.

Greco L., Pagnotta S. M. (2009). 'Density estimation by skew-normal kernels', in Proceedings of Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction – S.Co 2009. Milan, Italy: Maggeoli Editore. Vol.7, No. 1: 209-215

Hansen N., Niederberger A. S. P., Guzzella L. and Koumoutsakos P. (2009) 'A Method for Handling Uncertainty in Evolutionary Optimization With an Application to Feedback Control of Combustion,' in IEEE Transactions on Evolutionary Computation, vol. 13, no. 1, pp. 180-197.

Huang J., Xue Y., Dong Z. Y. and Wong K. P. (2012) 'An Efficient Probabilistic Assessment Method for Electricity Market Risk Management,' in IEEE Transactions on Power Systems, vol. 27, no. 3, pp. 1485-1493.

Negnevitsky M., Nguyen D. H. and Piekutowski M. (2015) "Risk Assessment for Power System Operation Planning with High Wind Power Penetration," in IEEE Transactions on Power Systems, vol. 30, no. 3, pp. 1359-1368.

Nikoobakht A., Mardaneh M., Aghaei J., Guerrero-Mestre V. and Contreras J. (2017) 'Flexible power system operation accommodating uncertain wind power generation using transmission topology control: an improved linearised AC SCUC model,' in IET Generation, Transmission & Distribution, vol. 11, no. 1, pp. 142-153, 1 5,