A Machine-Learning Approach for Regional Photovoltaic Power Forecasting

Yuan Li[1,2], Qian Sun[2], Brad Lehman[2], Siyuan Lu[3], Hendrik F. Hamann[3], Joseph Simmons[4,5], Jon Black[6]
1. Sichuan University, Chengdu, Sichuan 610065, China
2. Northeastern University, Boston, MA 02115, USA
3. IBM TJ Watson Research Center, Yorktown Heights, NY 10598, USA
4. University of Arizona, Tucson, AZ 85721, USA
5. Florida Gulf Coast University, Fort Myers, FL 33965, USA
6. ISO New England, Holyoke, MA 01040, USA

***Abstract*— This paper presents a machine-learning approach for regional photovoltaic (PV) power forecasting of up to 2 days ahead with hourly resolution. Physical PV power model is aggregated by geographical clusters and then summed for an entire ISO load zone. Numerical weather prediction (NWP) forecasts provide parameters, such as irradiance, temperature, barometric pressure, and wind speed, which are used as inputs to calculate plane of array (POA) irradiance and PV output power. A machine-learning approach is then developed. Bias correction for calculated power is conducted using linear regression method. During this procedure, categorization in accordance to critical parameters is employed to obtain a fine approximation. With optimized blending coefficients, adaptive mixture of correction results following different NWP methods is introduced to obtain an intelligent and adaptable output power PV forecast. A case study for the period from June 12, 2014 to January 24, 2015 of Southeastern Massachusetts (SEMA) load zone is carried out. Normalized Root Mean Square Error (NRMSE) is 5.28% for day-ahead forecast horizon, which is reduced by 30.6% compared to the baseline that the best individual model is used.**

***Index Terms*—machine-learning; solar power forecasting; photovoltaic; weather prediction**

I. INTRODUCTION

With the rapid growth of integration of solar energy into all levels of electricity supply systems, challenges include not only photovoltaic (PV) technologies, but also planning, organization and management on the system level. Unlike conventional unrenewable energy sources, solar energy production is largely determined by prevailing weather conditions, such as irradiance and temperature, which is variable. This feature makes its integration to the power system essentially different from conventional generation. Fig. 1 shows hourly net load curve of one summer day for peak load analysis presented by independent system operator (ISO) New England to demonstrate new operating conditions brought up by increasing PV penetration to the power system. Apparently, accurate forecast of renewable/PV generation is vital for the utility to reshape the electric demand profiles, and carry out further technologies to manage a green grid.

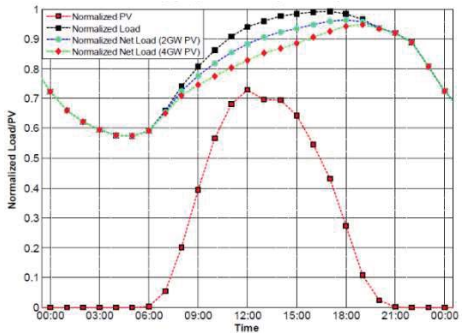


Figure 1. The net load curve for the summer peak load analysis considering different PV penetration level

Vast literature discusses PV power prediction approaches [2-14]. The approaches depend significantly on the forecasting timescale and geographic horizon. Regarding to the forecasting timescale, different approaches take diverse inputs: 1) for very short-term forecasting with temporal resolution from minutes to a few hours, on-site measured irradiance, power data and sky-camera imagery are utilized [2]-[6]; 2) for temporal resolution of 30 min to 6 hour, cloud motion vectors offered by satellite images are introduced into the model [7]; and 3) for forecasting that reaches out 6 hour and further, numerical weather prediction (NWP) model based methods are dominant [8]-[11]. Consequently, regional PV power prediction for 24-48 hours ahead are based on NWP models. The North American Mesoscale Forecast System (NAM), Short-Range Ensemble Forecast (SREF), Global Forecast System (GFS), Rapid Refresh (RAP), and Hi-Resolution Rapid Refresh (HRRR), etc. are well known NWP models that widely used in commercial or public weather services in North America. These NWP models provide irradiance and temperature predictions which can be used as inputs to PV power prediction models. However, accuracy of the various NWP is not well documented. In fact, as be documented in [12], various NWP models excel under different weather situations, for example, it could be that in low irradiance conditions, the HRRR forecast is more accurate than in high irradiance conditions. This motivates this present research to use machine learning adaptive model blending to weight each forecast differently to minimize the error.

Research is partially supported by US Department of Energy Contract #DE-EE0006017

Authorized licensed use limited to: Electricity Generating Authority of Thailand. Downloaded on August 23,2021 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

Specifically for 1–3 day-ahead PV power forecasts with hourly resolution, several PV approaches presented in literature [13]-[16]. Physical modeling derives PV power from predicted irradiance and other NWP parameters [13], [14]. Statistical methodology relies on historic data sets, establishing the relation between output power and predicted irradiance without physical modeling [15]. And a hybrid approach applies physical modeling as the first step, and then correct prediction bias according to historic data sets with a statistical model [16].

This paper presents a machine-learning approach for regional PV power forecasting of up to 2 days ahead. Since the actual weather forecasting is beyond the scope of the paper, results from various NWP models are utilized as a starting point from which all the evaluations and comparisons are made. Our approach to predict the aggregate sum of PV power in a region is to first divide the region into geographical clusters. Then we model the numerous PV installations within each cluster by a "virtual" PV power installation that represents the aggregate PV power for the entire cluster. The tilt angle, orientation, and efficiency are optimized numerically for best prediction results. In actuality, each NWP model has its own "virtual" PV power plant for each cluster, and machine learning is separately performed for multiple weather categories to improve the modeling to determine the adaptive weights of each of the NWP models for the different weather conditions. In the end, the cluster PV forecasts are summed and up-scaled to determine the aggregate PV power for a larger region. A case study of Southeastern Massachusetts (SEMA) load zone with public access to weather and PV site data is carried out. Step by step evaluation validates the effectiveness of this self-learning PV power forecasting approach. Normalized Root Mean Square Error (NRMSE) is 5.28% for day-ahead forecast horizon. Compared to the baseline that the best individual model (NAM in our case) is used, the NRMSE is reduced by 30.6%.

II. MACHINE-LEARNING PROCEDURE

A. Overview

The forecasted region is first divided into geographic clusters. Primarily residential and commercial PV sites are grouped within the cluster according to their distance to the centroid of the cluster. Fig. 2 shows the main procedure of the proposed PV power forecasting approach. Generally, the following steps are taken to obtain a final PV power forecasting result:

①. For the centroid of each cluster, NWP irradiance data of corresponding weather site are entered into a software program. POA irradiance is calculated with respect to cluster parameters (day of the year, location, module's position, etc.).

②. Together with other parameters (forecasted temperature, wind speed, barometric pressure, etc.), POA irradiance is used as the input of physical modeling of a "virtual" PV installation out from each cluster. The total calculated PV power of the whole region, P_{cal}, is then calculated as the summation of all the clusters' output power.

③. Data cleaning is carried out in order to eliminate bad data from the measured output power, P_{mea}, per entire zone. P_{mea} is onsite measurement, which is aggregated measured power per site via real-time PV data service.

④. Regional forecasted POA irradiance, I_{POA_f} (mean value of all the clusters' forecasted POA irradiance), P_{cal}, and P_{mea} are utilized to correct bias of the physical model via linear regression. P_{cal}, and P_{mea} are categorized into various bins, based on which linear coefficients of the bin are computed, and subsequently applied to bias correction. P_{cor} is the result of this step.

⑤. Each P_{cor} from each NWP model are then blended for optimized weighting coefficients. Forecasted errors are calculated and compared with iterating coefficients. The weight coefficients with the minimum errors are returned as the optimized coefficients after self-learning.

⑥. Step ① and ② are repeated with NWP parameters for coming 24 ~ 48 hours. Linear coefficients and weighting coefficients that trained in step ④ and ⑤ are then applied to get the final power forecasting result.

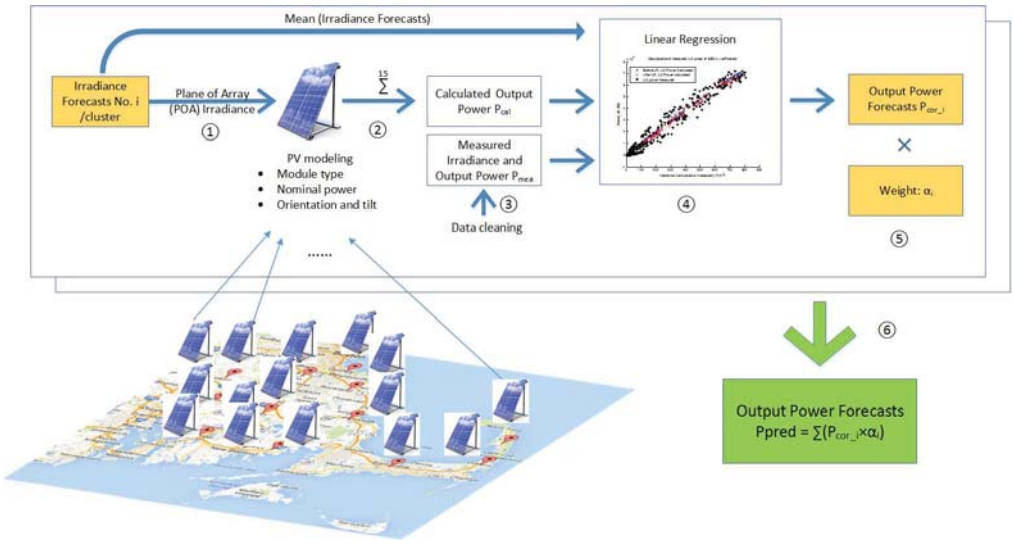


Figure 2. Overview of PV power forecasting procedure

Authorized licensed use limited to: Electricity Generating Authority of Thailand. Downloaded on August 23,2021 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

B. Physical PV Modelling

1) POA calculation

POA irradiance I_{POA_f} is first calculated/forecasted for a virtual PV panel, as shown in (1).

$$I_{POA_f} = I_{rr_b} + I_{rr_d} + I_{rr_r} \quad (1)$$

Hay, Davies, Klucher, Reindl (HDKR) model [17] is used for POA irradiance calculation. Symbols are defined as I_{rr_b} – beam irradiance; I_{rr_d} – diffuse irradiance; and I_{rr_r} – reflected irradiance of a tilted PV panel; where

$$I_{rr_b} = I_b \cdot R_b ; \quad (2)$$

$$I_{rr_d} = I_d \cdot A_i \cdot R_b + I_d \cdot (1 - A_i) \cdot (1 + \cos\beta)/2 \cdot [1 + f \cdot \sin(\beta/2)^3] ; \quad (3)$$

$$\text{and } I_{rr_r} = I \cdot \rho_g \cdot (1 - \cos\beta)/2 . \quad (4)$$

In (2) ~ (4), I, I_b and I_d stands for hourly Global Horizontal Irradiance (GHI); beam radiation that equals to Direct Normal Irradiance (DNI) $\cdot \cos(\theta_Z)$; and Diffuse Horizontal Irradiance (DHI), respectively. Other symbols are defined as follows: R_b – ratio of beam radiation on a tilted plane that equals to $\cos(\theta) / \cos(\theta_Z)$; A_i – anisotropy index that equals to I_b / I_o; I_O – hourly extraterrestrial irradiance on horizontal plane; θ – angle of incidence on the surface; θ_Z – zenith angle of the sun; β – tilted angle of the PV panel; ρ_g – ground reflectance. It should be noticed that for a "virtual" PV installation, parameters such as θ_Z, β, etc. may be quite different from any practical value. For instance, typical β is often around latitude or roof tilt of ~40 deg. in real installations. However, to minimize PV power prediction error it is possible that the virtual PV plant is anywhere from 5~50 degrees, depending on the weather situation after machine-learning approach.

2) PV power calculation

King's model [18] is applied to calculate the forecasted DC power and the AC power referring to inverter's derating efficiency curves. The DC power of one PV model is computed as:

$$P_{DC_\text{model}} = I_{POA_f} \cdot \tau_{PV} \cdot S \cdot \eta_C ; \quad (5)$$

$$\text{and } \eta_C = \eta_{ref}[1 - \beta_{ref}(T_c - T_{ref}) + \gamma_{ref} \lg I_{POA_f}] ; \quad (6)$$

where τ_{PV} is the transmittance of PV module; S is the area of solar cells on one panel; η_C is calculated cell efficiency; η_{ref} is the electrical efficiency at T_{ref} and irradiance of 1000 W/m^2; β_{ref} is the temperature coefficient of the material; T_c is the cell temperature; T_{ref} is the reference temperature that equals to 25 °C; and γ_{ref} is the solar radiation coefficient of the material. DC power of the model is then up-scaled by capacity of the cluster and transduced to AC power as

$$P_{DC_\text{cluster}} = D \cdot N \cdot P_{DC_\text{model}} ; \quad (7)$$

$$\text{and } P_{AC_\text{cluster}} = \eta_{inv} \cdot P_{DC_\text{cluster}} ; \quad (8)$$

where D is the derating factor for losses; N is the number of PV modules in one cluster; η_{inv} is the conversion efficiency of the inverter. The calculation of regional PV power P_{cal} can be obtained by

$$P_{cal} = \sum_{i=1}^{m} P_{DC_cluster,i} , \quad (9)$$

where m is the number of clusters in the studied region.

C. Linear Regression for Bias Correction

Bias correction is another important machine-learning procedure, which has significant effect on training the physical PV model to approximate measured value. In the training data sets, estimated error is obtained by $P'_{cal} - P'_{mea}$, where P'_{cal}, P'_{mea} are the linear approximation of calculated and measured output power versus forecasted irradiance, respectively:

$$E_{rr_est} = P'_{cal} - P'_{mea} ; \quad (10)$$

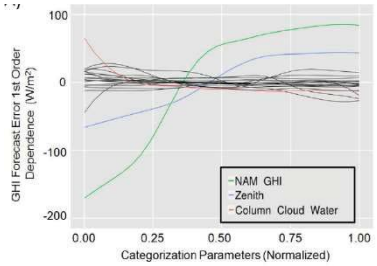
$$P'_{cal} = A_{cal} \cdot I_{rr_POA} + B_{cal} ; \quad (11)$$

$$\text{and } P'_{mea} = A_{mea} \cdot I_{rr_POA} + B_{mea} . \quad (12)$$

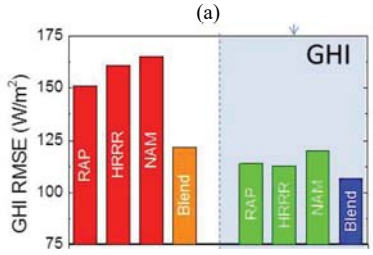
Corrected power is obtained by

$$P_{cor} = P_{cal} - E_{rr_est}. \quad (13)$$

In order to approximate measured PV power values better, multiple "bins" are utilized to get the first-order analytic expression of output power. "Bins" refer to multiple categorizations in accordance to parameters that have significant influence on PV output. This research project has discovered that forecast accuracy can depend strongly on "weather situation" category. For example, Fig. 3 (a) shows the 1st order dependence on several parameters, and concludes that GHI from NAM, and solar zenith angle are the top two critical parameters that have the most influence on forecast error. And Fig. 3 (b) shows that the error is significantly reduced once GHI categorization is used for machine-learning.



(a)



(b)

Figure. 3 (a) Forecast error of GHI from NAM model: 1st order dependence on input parameters derived using functional analysis of variance. (b) Error before and after using GHI for situation categorization.

This very important result motivate us to use multi-dimensional "bins" in accordance to model parameters, e.g. forecasted irradiance, tilted angle, clear sky index, etc. to improve forecast accuracy. For each "bin", linear coefficients A

Authorized licensed use limited to: Electricity Generating Authority of Thailand. Downloaded on August 23,2021 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

and B are obtained via linear regression. And P_{cor} after bias correction is prepared for final process.

D. Adaptive Blending

To minimize the forecast error, machine-learning approach is carried on by adaptive blending of forecasting results from different NWP models. Since different NWP models provide varying accuracies depending on weather situation, adaptive blending creates a way to let the prediction from the "right" category (with less error under certain condition) weight more in the result, but at the mean time keeps diversity of the model. So on one hand side the forecasting approach minimizes error within the training data sets, and on the other hand side it is also adaptable and scalable.

Define blending coefficient vector as

$$\boldsymbol{\alpha} = [\alpha_1 \quad \ldots \quad \alpha_n], \tag{14}$$

where n refers to the number of NWP forecasting models. Blended forecasting power is calculated as

$$P_{pred} = \boldsymbol{\alpha}[P_{cor,1} \quad \ldots \quad P_{cor,n}]', \tag{15}$$

where $P_{cor,k}$ (k = *1*, ..., n) refers to the corrected PV output power corresponding to the k^{th} NWP forecasting model.

As addressed in [19], statistical metrics (root mean square error(RMSE), normalized RMSE, mean absolute error (MAE) , mean absolute percentage error (MAPE), etc.), uncertainty quantification metrics, ramp characterization and economic metrics are generally used to evaluate a certain forecast. Here we use normalized RMSE as an example for machine-learning approach, which is defined as

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{hours} (P_{mea,i} - P_{pred,i})^2 / hours}}{Capacity}, \tag{16}$$

where *hours* refers to the predicted timescale, *Capacity* refers to the total capacity of the PV sites in evaluation, $P_{mea,i}$ and $P_{pred,i}$ refer to the i^{th} hour power in the data sets, respectively. Nelder-Mead method is used to find the minimum forecast error, which is known as a multidimensional unconstrained nonlinear minimization algorithm [20]. For the objective function

$$NRMSE = f(\boldsymbol{\alpha}), \tag{17}$$

a set of n + 1 test points are arranged as a simplex. Initial blending coefficients $\boldsymbol{\alpha}_0$ are firstly assigned. It then extrapolates the behavior of the objective function measured at each test point, and find a new test point to replace the worst one, and so the technique progresses. Returned $\boldsymbol{\alpha}$ after iteration is used for predicted power calculation.

E. Up-Scaling for Regional PV Power Forecasting

For regional PV power forecasting, the task of downloading measured PV data from all the PV sites is tedious, sometimes impossible. Representative PV sites are sampled, and the measured power of which make up training data sets. To obtain the predicted power of entire region, up-scaling is presented as

$$P_{pred,region} = \frac{P_{pred,sample}}{Capacity_{sample}} \times Capacity_{region}. \tag{18}$$

The sampled PV sites should provide representation of the PV actual production in the region. In our ongoing research for New England load zone, uncertainty of forecast result is used to lay statistical foundation for sampling. As shown in Fig. 4, if the sample size exceeds 5% (42 out of 832 sites), the uncertainty of error (out of 100 times calculation using randomly chosen PV sites) is less than 1%. It indicates that once sample size is adequate, difference among the results is insignificant, therefore up-scaling is reliable.

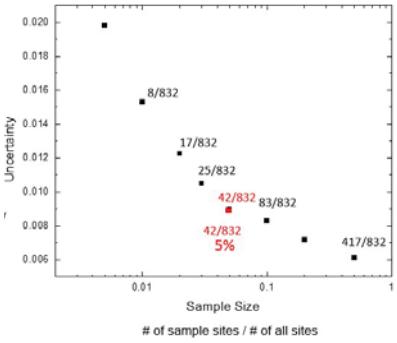


Figure 4. Uncertainty dependence on sample size

III. Case Study and Evaluation

A. Overview of SEMA Region Forecasting

PV power forecasting for Southeastern Massachusetts (SEMA) region is carried out for evaluation for the period June 12, 2014 to January 24, 2015. SEMA load zone is composed of the Cape Cod region as well as southeast Massachusetts. As shown in Fig. 5, there are 15 clusters in SEMA region, consisting of 158 PV sites in total. Location of clusters is determined comprehensively by geographic position and PV capacity. The hourly measured power of entire load zone P_{mea} is obtained from publicly available Solren view real-time PV data service by aggregating the output power of 158 PV sites. Since the training data sets consist of all the 158 sites, up-scaling procedure is not included in this case.

B. Step by Step Evaluation

As an example, the forecasting irradiance based on NAM, SREF and GFS models are currently used for PV power

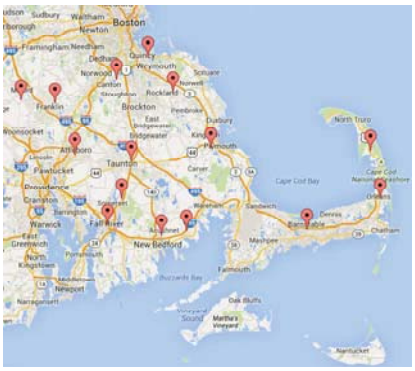


Figure 5. SEMA region and clusters

Authorized licensed use limited to: Electricity Generating Authority of Thailand. Downloaded on August 23,2021 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

forecasting. Fig. 6 shows the machine-learning approach step by step with NAM model: step I) P_{cal} by individual model; step II), P_{cor} by bias correction; and step III) P_{pred} by blending. SREF and GFS models show similar results. In the figure the blue "*" dots indicate forecasting power out of each step, and the red "+" dots indicate measured output power as a reference. It can be seen clearly that blue "*" dots made their way to approximate red "+" step by step. Fig. 7 provides NRMSE of every step for three engaged NWP models. After blending NRMSE was 5.28% for day-ahead forecast horizon, which is reduced by 30.6%, 42.3%, and 44.9% compared to the baseline that the individual model is used, respectively.

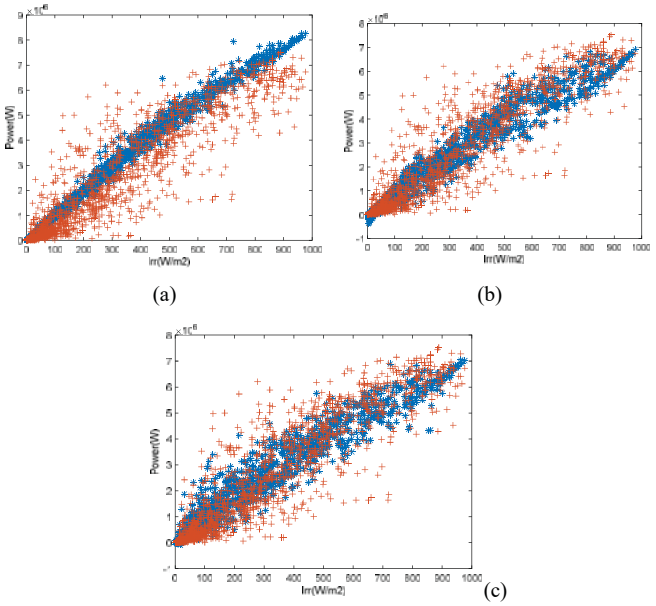


Figure 6. Step by step result (blue "*") of machine learning where P_{mea} (red "+") is used as reference (a) P_{cal} vs. I_{POA_f}; (b) P_{cor} vs. I_{POA_f}; (c) P_{pred} vs. I_{POA_f}

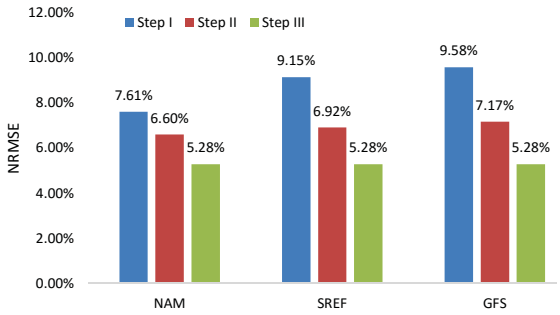


Figure 7. NRMSE of every step for NAM, SREF and GFS models

IV. CONCLUSION

This paper presents a machine-learning approach for regional PV power forecasting, which consists of 1) virtual PV installation with optimized model parameters; 2) multiple categorizations in accordance to model parameters that have significant influence on PV output; and 3) adaptive blending of forecasting results from different NWP models. A case study of PV power forecasting for SEMA load zone is provided to validate the effectiveness of the proposed method. The NRMSE is reduced by 30.6% compared to the best individual model is used.

REFERENCES

[1] Jon Black, ISO New England, Update on Solar PV and Other DG in New England, Holyoke, MA. [Online]. Available: http://www.iso-ne.com/committees/comm_wkgrps/prtcpnts_comm/pac/mtrls/2013/jun-192013/a7_solar_dg_update.pdf.

[2] J. Kleissl, *Solar Energy Forecasting and Resource Assessment*, 1st ed., Waltham, Massachusetts: Academic Press, 2013.

[3] S. Pelland, J. Remund, J. Kleissl, T. Oozeki, K. D. Brabandere, "Photovoltaic and Solar Forecasting: State of the Art," *Technical report*, IEA PVPS T14-01: 2013.

[4] G. Reikard, "Predicting solar radiation at high resolutions: a comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, March 2009.

[5] R. Marquez, C. F. M. Coimbra, "Intra-hour DNI forecasting based on cloud tracking image analysis". *Solar Energy*, vol. 91, pp. 327 – 336, May 2013.

[6] Y. Chu, L. Nonnenmacher, R. H. Inman, Z. Liao, H. T. C. Pedro, C. F. M. Coimbra, "A smart image-based cloud detection system for intra-hour solar irradiance forecasts," J. Atmos. Ocean. Technol. vol.31, pp.1995 – 2007, Sep. 2014.

[7] S. Cros, O. Liandrat, N. Sebastien, and N. Schmutz, "Extracting cloud motion vectors from satellite images for solar power forecasting," *in Proc. of IEEE International Geoscience and Remote Sensing Symposium*, pp. 4123-4126, Quebec City, QC, 2014.

[8] 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, pp. 757-771, Sep. 2010.

[9] C. Chen, S. Duan, T. Cai, B. Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," *Solar Energy*, vol. 85, pp. 2856 – 2870, Nov. 2011.

[10] Y. Zhang, M. Beaudin; R. Taheri, H. Zareipour; and D. Wood, "Day-Ahead Power Output Forecasting for Small-Scale Solar Photovoltaic Electricity Generators," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2253-2262, Sep. 2015.

[11] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker Jr., D. Renne´ and TE. Hoff. "Validation of short and medium term operational solar radiation forecasts in the US," *in Proc. 2009 SES Annual Conf.*, pp. 2161-2172, Buffalo, US.

[12] L. S. Lu, Y. Hwang, I. Khabibrakhmanov, F. J. Marianno, X. Shao, J. Zhang, B.-M. Hodge, H. F. Hamann, "Machine Learning Based Multi-Physical-Model Blending for Enhancing Renewable Energy Forecast – Improvement via Situation Dependent Error Correction." *Proceeding of European Control Conference 2015*,WeB11.3, in press (2015).

[13] E. Lorenz, J. Hurka, G. Karampela, D. Heinemann, HG. Beyer, and M. Schneider. "Qualified forecast of ensemble power production by spatially dispersed grid-connected PV systems," *in Proc. 2008 23rd European Photovoltaic Solar Energy Conf.*, pp. 1-7, Valencia, Spain.

[14] J. Remund, C. Schilter, S. Dierer, S. Stettler, and P. Toggweiler. "Operational forecast of PV production," *in Proc. 2008 23rd European Photovoltaic Solar Energy Conf.*, pp. 1-3, Valencia, Spain.

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

[16] X. Le Pivert, L. Sicot, and J. Merten. "A tool for the 24 hours forecast of photovoltaic production," *in Proc. 2009 24th European Photovoltaic Solar Energy Conf.*, Hamburg, Germany.

[17] J. A. Duffie and W. A. Beckman, *Solar Engineering of Thermal Processes*, 3rd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.

[18] D. L. King, J. A. Kratchvil, and W. E. Boyson, "Field experience with a new performance characterization procedure for photovoltaic arrays," *in Proc. 2nd World Conf. and Exhibition on Photovoltaic Solar Energy Conversion*, pp. 1947-1952, Vienna, Austria.

[19] J. Zhang, B. M hodge, S. Y. Lu, H. F. Hamann, B. Lehman J. Simmons, and etc, "Baseline and target for regional and point solar power forecasts: toward improved solar power forecasting," Solar Energy, in press (2015)

[20] McKinnon, K.I.M, "Convergence of the Nelder–Mead simplex method to a non-stationary point," *SIAM J Optimization*, vol. 9, pp: 148–158, 1999.

Authorized licensed use limited to: Electricity Generating Authority of Thailand. Downloaded on August 23,2021 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.