Short-term Global Horizontal Irradiance Forecasting Based on Sky Imaging and Pattern Recognition

Cong Feng, Mingjian Cui, Meredith Lee, and Jie Zhang
University of Texas at Dallas
Richardson, TX, 75080 USA

Bri-Mathias Hodge
National Renewable Energy Laboratory
Golden, CO 80401, USA

Siyuan Lu, Hendrik F. Hamann
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA

***Abstract*—Accurate short-term forecasting is crucial for solar integration in the power grid. In this paper, a classfication forecasting framework based on pattern recognition is developed for 1-hour-ahead global horizontal irradiance (GHI) forecasting. Three sets of models in the forecasting framework are trained by the data partitioned from the preprocessing analysis. The first two sets of models forecast GHI for the first four daylight hours of each day. Then the GHI values in the remaining hours are forecasted by an optimal machine learning model determined based on a weather pattern classification model in the third model set. The weather pattern is determined by a support vector machine (SVM) classifier. The developed framework is validated by the GHI and sky imaging data from the National Renewable Energy Laboratory (NREL). Results show that the developed short-term forecasting framework outperforms the persistence benchmark by 16% in terms of the normalized mean absolute error and 25% in terms of the normalized root mean square error.**

***Index Terms*—Cassification, solar forecasting, sky imaging, pattern recognition, support vector machine.**

I. Introduction

Solar energy is one of the most promising candidates to tackle the energy crisis, but grid integration is still challenging due to the variability and uncertainty associated with the power output. Thus, accurate solar forecasting is crucial to the economic dispatch and to the reliability of the power grid, as it reduces the uncertainty in power system operations. [1]. Solar generation is mainly affected by the solar irradiance, which in turn is largely driven by the movement of clouds.

Machine learning models have shown better accuracy than physical models [2] for short-term (within 1-hour ahead) GHI forecasting. A number of advanced techniques have been recently used to enhance short-term GHI forecasting, such as total sky images, satellite images, and other numerical weather prediction models. Among these information sources, variables like the lagging data of the forecasting variable [3], cloud indices (CI) [4], and red blue ratio (RBR) features [5] are the most informative inputs to the machine learning models.

The GHI is highly influenced by the weather condition (e.g., sunny, partially cloudy, and cloudy). It is generally challanging to ensure an accurate forecast for different weather types from a single model. Multi-model solar forecasting based on a weather type classfication has shown to be an effective way to solve this challenge. Other information, such as temperature [6], the self-organized map [7] [8], and the weather report [9], have been used as classification criteria. However, classification forecasting is currently mainly used in 1-day-ahead or even longer time horizons. Further, those classification techniques are limited by expensive computation (e.g., weather report) or inaccurate performance (e.g., temperature).

In this paper, a new short-term GHI forecasting framework that utlizes pattern recognition to classify the weather type, is developed to conduct the 1-hour-ahead GHI forecasting.

II. Weather Pattern Recognition

To improve the overall GHI forecasting accuracy over a day, multiple forecasting models can be trained based on the weather type. The weather of a single day is categorized by the average clear sky index (CSI), which is the ratio of actual GHI and clear sky GHI, into three types (sunny: $CSI > 0.75$, cloudy: $CSI < 0.25$, and partially cloudy: $0.25 \leq CSI \leq 0.75$). To choose the most suitable trained model for forecasting future GHI, a simple direct classification can be used based on the calculated CSIs from multiple hours prior to the forecasting data point. This direct classification method determines the weather type of a day by its first i hours CSIs. Figure. 1 illustrates the accuracy of the direct classification with different numbers of hours adopted. Support vector regression is used as the forecasting engine for this illustration. It is shown that the more hours of data used to determine the weather type, the more accurate the weather categorization is. The categorization accuracy highly affects the forecasting performance. From this figure, more than 8 hours of data are needed to achieve a 50% categorization accuracy, which is too many for real-time forecasting, since the first 8 hours of the day cannot be forecasted. Thus, pattern recognition is applied in this paper to identify the weather type of a day by the first few hours' data, which is expected to use fewer hours for the classification. Pattern recognition is a kind of signal identification technique that is popularly applied in different engineering fields. In this paper, a support vector machine (SVM) classifier is adopted as the pattern recognition algorithm.

A. Feature Extraction

To obtain well-performing pattern recognition models, suitable features need to be extracted from different information

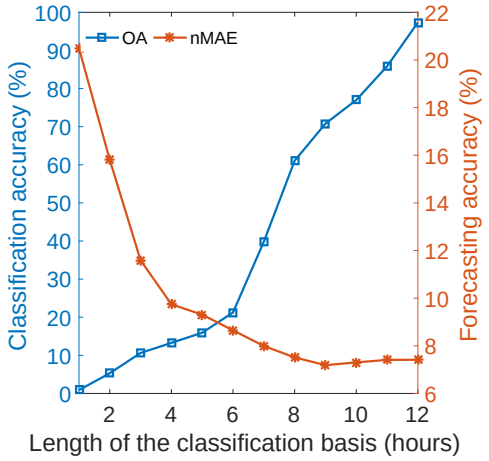


Figure 1: Categorization and forecasting accuracy of the direct classification method. Overall accuracy (OA) and normailized mean absolute error ($nMAE$) are evaluation metrics to measure the categorization and forecasting accuracy, respectively, which are defined in Section IV.

sources and fed into the models. The features selected in this paper are from two categories: i) GHI features: historical GHI (GHI), clear sky GHI (GHI_{clr}), and CSI; and ii) sky imaging features: mean (μ), standard deviation (σ), and Rényi entropy (H) of the normalized sky image pixel RBR ($nRBR$) values. GHI_{clr} is the GHI value under no-cloud condition, which is generated by a clear-sky models. In this paper, the Ineichen and Perez model [10] is selected as the clear-sky model. CSI is the ratio of GHI and GHI_{clr}, which can be used as the criteria to classify the weather types. The other three features are extracted by sky image processing, and are expressed as:

$$RGB = \begin{bmatrix} r_1 & g_1 & b_1 \\ \vdots & \vdots & \vdots \\ r_n & g_n & b_n \end{bmatrix} \tag{1}$$

where r, g, and b represent the red, green, and blue values of one sky image pixel in the RGB color system. n is the number of pixels in each image, which is 1392×1040. Then $nRBR$ of pixels is calculated by:

$$nRBR = \frac{r-b}{r+b} \tag{2}$$

$nRBR$ is the basis to calculate the three sky imaging features μ, σ, and H. H is the Rényi entropy, defined as:

$$H = \frac{1}{1-\alpha} log \left[\sum_{i=1}^{n} (p_i^{\alpha}) \right] \tag{3}$$

when $\alpha = 2$, it is the order of Rényi entropy. p_i^{α} is the frequency for the ith bin (out of 150 evenly spaced bins evenly spaced). These 6 features (i.e., GHI, GHI_{clr}, CSI, μ, σ, and H) compose the feature space serving as the inputs to the pattern recognition model.

B. Pattern Recognition Algorithm

Classification is a type of pattern recognition method. The support vector machine (SVM) classifier is selected for classifying the weather type. To model an SVM classifier, the outputs (weather types) are assumed to take a form of [11]:

$$y_i = \omega_i^T \cdot \kappa(x, x') + \psi \tag{4}$$

where ω_i is an l-dimensional weighted vector. x is the n-dimensional input vector. $n = 6j$ (j is the number of hours chosen as classification basis). ψ is the bias constant. κ is the kernel function that maps the n-dimensional input vector into an l feature space. The radial basis function (RBF) is selected as the kernel function, expressed as:

$$\kappa(x, x') = e^{-\frac{\|x-x'\|}{2\varrho^2}} \tag{5}$$

where ϱ is the kernel parameter. The objective function of the SVM is formulated as:

$$\min \quad \frac{1}{2}\|\omega\|^2 + \mathrm{C}(\sum_{\mathrm{i}=1}^{\mathrm{t}}(\xi_{\mathrm{i}} + \xi_{\mathrm{i}}^*)) \tag{6}$$

subject to:

$$\langle \omega, x_i \rangle + \psi - y_i \leq \epsilon + \xi_i^*, \ \forall i \tag{7a}$$

$$y_i - \langle \omega, x_i \rangle - \psi \leq \epsilon + \xi_i, , \ \forall i \tag{7b}$$

$$\xi_i, \ \xi_i^* \geq 0 \tag{7c}$$

where ξ and ξ^* are the upper and lower ϵ bands of the deviations around the objective function. C is a tradeoff parameter. Once the classfier model is trained, the weather type can be categorized by the inputs vector x with the same features.

III. Forecasting Methodology

The short-term GHI forecasting framework developed in this paper is summarized in Fig. 2. In addition to pattern recognition, the framework contains two other parts: the data preprocessing module and the GHI forecasting module. In the data preprocessing module, a three-step technique is applied to improve the pattern recognition and forecasting performance. The forecasting module is divided into three model sets: Model Set I ($MS–I$), Model Set II ($MS–II$), and Model Set III ($MS–III$).

A. Data Preprocessing Module

A three-step data preprocessing is conducted to enhance the pattern recognition and the GHI forecasting accuracy, including the following steps: i) elimination, ii) normalization, and iii) reconstruction. Data elimination aims to exclude data in the early morning (before 7 : 00 am in this study) and late night (after 7 : 00 pm), since most GHIs are 0. Normalization converts the data to the range between 0 and 1. Data reconstruction aims to group the training data into three model sets as follows:

$$G1 = \left[GHI_1 \ GHI_{14} \ \cdots \ GHI_{1+13(t-1)} \ \right]^T \tag{8}$$

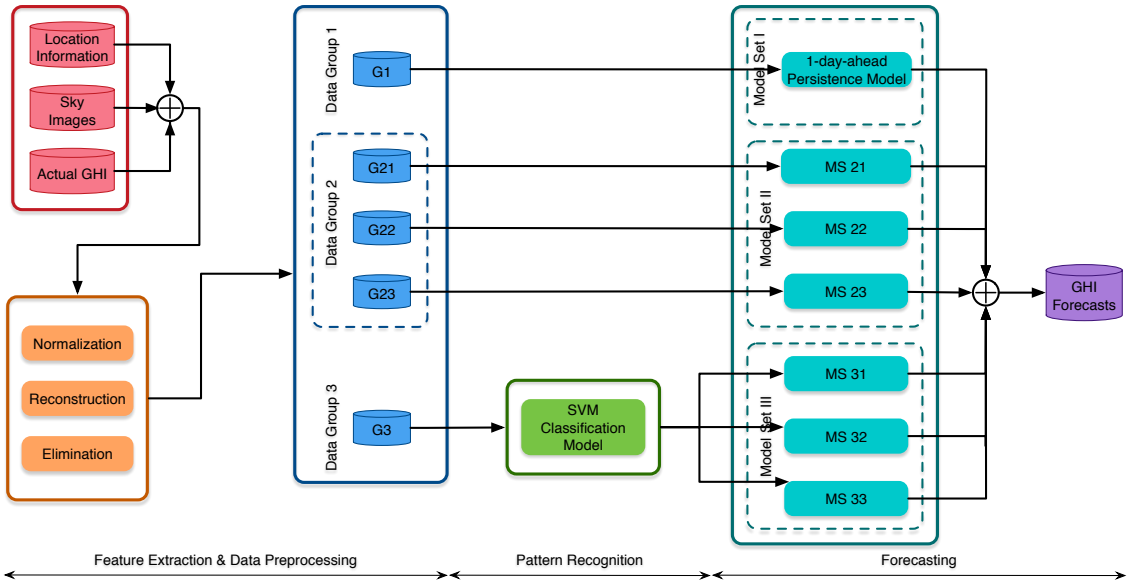


Figure 2: Overall framework of the short-term GHI forecasting based on sky imaging and pattern recognition.

$$G2_i = \begin{bmatrix} X_i & y_i \\ X_{i+13} & y_{i+13} \\ \vdots & \vdots \\ X_{i+13(t-1)} & y_{(i+13(t-1))} \end{bmatrix}, i = 1, 2, 3 \quad (9)$$

$$G3 = \begin{bmatrix} X_1^{'} & Y_1 \\ X_2^{'} & Y_2 \\ \vdots & \vdots \\ X_t^{'} & Y_t \end{bmatrix}, X_i^{'} = \begin{bmatrix} X_{4+9(i-1)} \\ X_{5+9(i-1)} \\ \vdots \\ X_{12+9(i-1)} \end{bmatrix}, Y_i = \begin{bmatrix} y_3 \\ y_4 \\ \vdots \\ y_{12} \end{bmatrix} \quad (10)$$

where t is the number of days in the original training data, $X_i = [GHI_i\ GHI_{clr,i}\ CSI_i\ \mu_i\ \sigma_i\ H_i]$, and $y_i = GHI_{i+1}$. $G2$ has three matrices ($i = 1, 2, 3$), each of which only contains samples at a specific hour (7:00, 8:00, or 9:00 am) and their 1-hour-ahead GHI data. $G3$ includes samples from 10:00 am to 6:00 pm within one day for t days.

B. GHI Forecasting Module

Three sets of forecasting models are developed to compose the forecasting module. MS–I model is a 1-day-ahead persistence model to forecast the first hour's GHIs ($GHI_{7:00\ am}$) every day. In this paper, a persistence of cloudiness model that assumes a constant clear-sky index within the forecasting time horizon is chosen, which is given by:

$$GHI_p(t + \Delta t) = \frac{GHI(t)}{GHI_{clr}(t)} \times GHI_{clr}(t + \Delta t) \quad (11)$$

where $GHI_p(t+\Delta t)$ means the persistent prediction of GHI at time t within the time horizon Δt. GHI_{clr} is the GHI generated by the clear-sky model.

The MS–II and MS–III forecasting models are developed using machine learning algorithms. Both MS–II and MS–III have multiple models. MS–II models (i.e., MS_{21}, MS_{22}, and MS_{23}) are trained by the data at the same historical hour (i.e., $G21$, $G22$, and $G23$) to predict GHIs from 8:00 am to 10:00 am. For example, the machine learning model is trained by all historical 8:00 am GHIs to forecast the 8:00 am GHI. MS–III models (i.e., MS_{31}, MS_{32}, and MS_{33}) are trained by the data of three different weather types, to predict GHIs from 11:00 am to 7:00 pm. The SVM regression is used as the forecasting algorithm to train the MS–II and MS–III models, which conforms to similar principles as described in Section II.

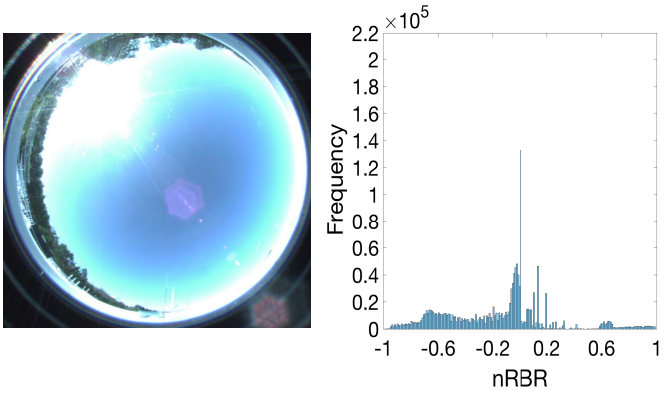
IV. Case Studies

To validate the developed forecasting framework, we used the GHI and sky imaging data (latitude = 39.742° North, longitude = 105.18° West, elevation = 1,828.8 m) released by NREL.

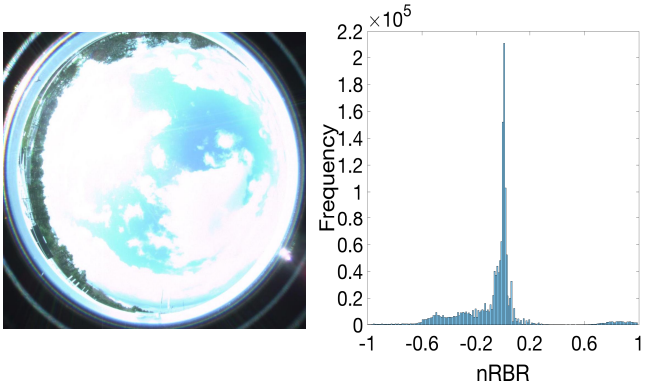
A. Classification Accuracy

The proportion of sunny days, partially cloudy days, and cloudy days in $G3$ are 71%, 23%, and 6%, respectively. The sky images and the corresponding pixel $nRBR$ histograms of three weather types are shown in Fig. 3. It is observed that the sky image of a cloudy day is significantly different from the sky images of the sunny and partially cloudy days. In contrast to Fig. 3b, Fig. 3a has more navy pixels, whose $nRBR$s fall into the range [-1, -0.5]. The other three features are related to the GHIs and GHI_{clr}s. The GHI and GHI_{clr} curves for different weather types are shown in Fig. 4. It may be seen that the GHIs among different weather types vary considerably. Since all six features show a difference among the three weather patterns, all the 6 features are helpful for the pattern recognition.

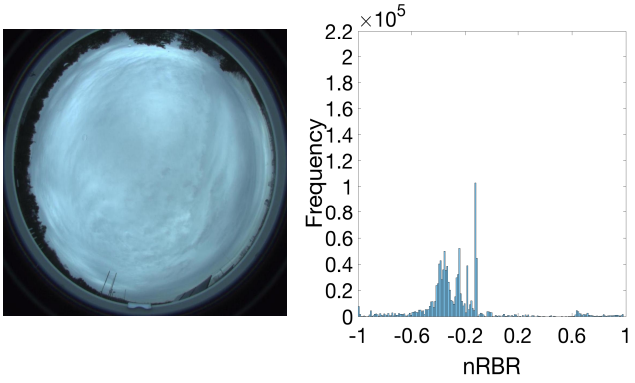
Based on the SVM classification models, the weather pattern of one day is recognized using the first 4 hours' features of the day. The pattern recognition accuracy is evaluated by the



(a) Sunny day



(b) Partially cloudy day



(c) Cloudy day

Figure 3: Sky images and corresponding pixel $nRBR$ histograms of different weather types.

three metrices, which are true positive rate (TP_{rate}), precision (P), and overall accuracy (OA). They are defined as:

$$TP_{rate} = \frac{m_{ii}}{\sum_{j=1}^{n} m_{ij}}, \ i = 1, \cdots, n \tag{12}$$

$$P = \frac{m_{ii}}{\sum_{j=1}^{n} m_{ji}}, \ i = 1, \cdots, n \tag{13}$$

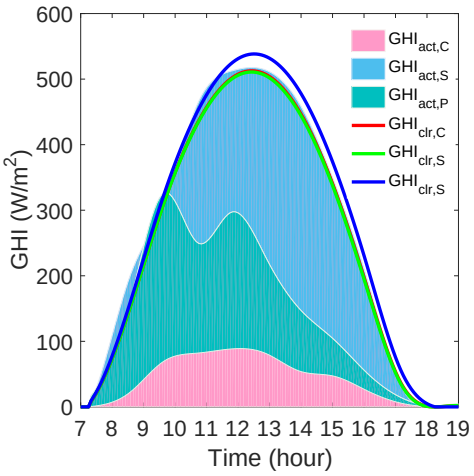


Figure 4: Actual GHIs and clear sky GHIs of different weather types. $GHI_{act,C}$ is the actual GHI in a cloudy day, $GHI_{act,S}$ is the actual GHI in a sunny day, $GHI_{act,P}$ is the actual GHI in a partially cloudy day, $GHI_{clr,C}$ is the clear sky GHI in a cloudy day, $GHI_{clr,S}$ is the clear sky GHI in a sunny day, and $GHI_{clr,P}$ is the clear sky GHI in a partially cloudy day.

$$OA = \frac{\sum_{i=1}^{n} m_{ii}}{\sum_{j=1}^{n} \sum_{i=1}^{n} m_{ij}}, \ i = 1, \cdots, n \tag{14}$$

where n is the number of patterns. m_{ij} represents the objects belonging to the pattern i and being classified to pattern j.

Table I: Pattern recognition results

	Sunny	Partially cloudy	Cloudy
Sunny	87	0	0
Partially cloudy	12	13	0
Cloudy	1	0	0

Note: The actual classifications were determined by the daily average CSIs which are shown in red. The pattern recognition results are shown in green.

Table II: Pattern recognition accuracy

	Sunny	Partially cloudy	Cloudy
TP_{rate} (%)	100	52	0
P (%)	87	100	100
OA (%)	–	88	–

In Tables I and II, the testing data contains 113 days, 87 of which are sunny days. All of the sunny days are correctly recognized. The testing data also contains 25 partially cloudy days, 12 of which are misrecognized as sunny days. And the only cloudy day in the testing data is misrecognized as a sunny day. The reason for mis-recognition of cloudy and partially cloudy days is because the number of these two patterns are too small in the SVM classifier training data to train the SVM classifier. But the OA is 88%, which is a significant improvement compared to the direct classication method (13% with the first 4 hours data).

B. Forecasting Accuracy

Based on the encouraging pattern recognition results, one of the $MS–III$ models is selected for a certain day to forecast 1-hour-ahead GHIs from 11:00 am to 7:00 pm. The first four hours' GHIs are predicted by four other models in the $MS–I$ and $MS–II$. The forecasted GHIs within the day are shown in Fig. 5. It is shown that the $MS–I$ model has relatively large errors comparing to other models in the forecasting module; $MS–II$ models show relatively more accurate forecasting; $MS–III$ models also have encouraging results, except for the late afternoon. The significant forecasting errors may be caused by the seasonal difference between the training and testing data. Compared to the 1-hour-ahead persistence model, the developed framework is more accurate.

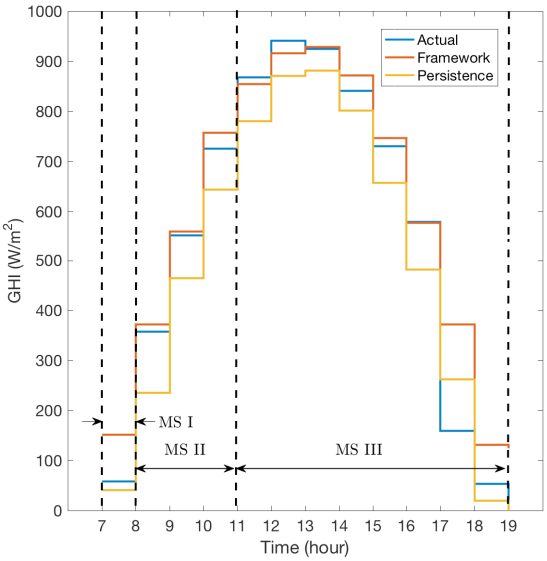


Figure 5: Step plot of forecasted GHIs within one day.

In order to evaluate the overall forecasting accuracy of the developed framework, two error criteria are utilized: the normalized mean absolute error ($nMAE$) and the normalized root mean square error ($nRMSE$), respectively, defined by:

$$nMAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i - y_i}{y_{max}}\right| \tag{15}$$

$$nRMSE = \frac{1}{y_{max}}\sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{n}} \tag{16}$$

where f_i is the forecasted GHI. y_i is the actual GHI. y_{max} is the maximum actual GHI over the year 1,068 W/m^2. n is the sample size 1,469. The evaluation results are listed in Table III. The overall results are calculated based on the time series data compiled from three model sets' results.The overall performance of the proposed framework outperforms the 1-hour-ahead persistence model by 16% in terms of $nMAE$ and 25% in terms of $nRMSE$.

Table III: Forecasting accuracy

	P	MSI	$MSII$	$MSIII$	Overall
$nMAE$ (%)	7.93	2.69	7.40	6.92	**6.70**
$nRMSE$ (%)	12.94	4.14	10.65	9.88	**9.75**

Note: The overall $nMAE$ and $nRMSE$ performance of the developed framework are in boldface. P stands for persistence of cloudiness model.

V. Conclusion

In this paper, a classification forecasting framework in conjunction with pattern recognition was developed for short-term 1-hour-ahead GHI forecasting. The developed framework recognized the weather pattern of each single day using the SVM classifier and selected the most suitable forecasting model. Multi-set of models were applied in the forecasting module to forecast the GHIs from 7:00 am to 7:00 pm. The developed framework has shown a promising pattern recognition performance and performed better than the persistence model for 1-hour-ahead GHI forecasitng. The potential future work is to validate the developed framework with larger data sets and to compare to more benckmark models.

Acknowledgment

This work was supported by the National Renewable Energy Laboratory under Subcontract No. XHQ-6-62546-01 (under the U.S. Department of Energy Prime Contract No. DE-AC36-08GO28308).

References

[1] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," *CSEE Journal of Power and Energy Systems*, vol. 1, no. 4, pp. 38–46, 2015.

[2] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. Martinez-de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016.

[3] R. Azimi, M. Ghayekhloo, and M. Ghofrani, "A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting," *Energy Conversion and Management*, vol. 118, pp. 331–344, 2016.

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

[5] Y. Chu, M. Li, and C. F. Coimbra, "Sun-tracking imaging system for intra-hour dni forecasts," *Renewable Energy*, vol. 96, pp. 792–799, 2016.

[6] M. Ding, L. Wang, and R. Bi, "An ann-based approach for forecasting the power output of photovoltaic system," *Procedia Environmental Sciences*, vol. 11, pp. 1308–1315, 2011.

[7] C. Chen, S. Duan, T. Cai, and B. Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," *Solar Energy*, vol. 85, no. 11, pp. 2856–2870, 2011.

[8] H. Yang, C. Huang, Y. Huang, and Y. Pai, "A weather-based hybrid method for 1-day ahead hourly forecasting of pv power output," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 3, pp. 917–926, 2014.

[9] J. Shi, W. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Transactions on Industry Applications*, vol. 48, no. 3, pp. 1064–1069, 2012.

[10] P. Ineichen and R. Perez, "A new airmass independent formulation for the linke turbidity coefficient," *Solar Energy*, vol. 73, no. 3, pp. 151–157, 2002.

[11] F. Wang, Z. Zhen, Z. Mi, H. Sun, S. Su, and G. Yang, "Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting," *Energy and Buildings*, vol. 86, pp. 427–438, 2015.