IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 14, 2020, accepted May 28, 2020, date of publication June 4, 2020, date of current version June 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999903

Similarity-Based Models for Day-Ahead Solar PV Generation Forecasting

HOSSEIN SANGRODY, (Student Member, IEEE), NING ZHOU, (Senior Member, IEEE), AND ZIANG ZHANG, (Member, IEEE)

Department of Electrical and Computer Engineering, State University of New York at Binghamton, Binghamton, NY 13905, USA

Corresponding author: Hossein Sangrody (hsangro1@binghamton.edu)

This work was supported by the New York State Energy Research and Development Authority (NYSERDA) through the High Performing Grid Program, and in part by the NSF under Grant #1845523.

ABSTRACT Accurate forecasting of solar photovoltaic (PV) power for the next day plays an important role in unit commitment, economic dispatch, and storage system management. However, forecasting solar PV power in high temporal resolution such as five-minute resolution is challenging because most of PV forecasting models can only achieve the same temporal resolution as their predictors(i.e., weather variables), whose temporal resolution is usually low (i.e., hourly). To address this challenge, similarity-based forecasting models (SBFMs) are advocated in this paper to forecast PV power in high temporal resolution using low temporal resolution weather variables. To effectively generalize the model for different scenarios of available weather data, three forecasting models (i.e., basic SBFM, categorical SBFM, and hierarchical SBFM) are proposed. As a case study, the PV power generated by the solar panels on the rooftop of a commercial building is forecasted for the next day with a five-minute resolution under three different scenarios of available weather data. The leave-one-out cross-validation analysis reveals that using only two or three weather variables, the proposed SBFMs can achieve higher forecasting accuracy than several benchmark models.

INDEX TERMS Solar PV forecasting, similarity analysis, hierarchical similarity, high temporal resolution solar forecasting, day-ahead forecasting.

I. INTRODUCTION

Variability and uncertainty of unprecedentedly growing solar photovoltaic (PV) generation have incurred serious stability, reliability, and integration costs in power system operation, planning, and market [1]–[4]. The accelerating penetration of solar PV generation in the power grid calls for accurate solar PV power forecasting, which provides authorities an understanding of generation at different forecast horizons. Based on forecast horizons, the PV power forecasting methods can be classified into three categories: short-term, medium-term, and long-term forecasting [5], [6]. The medium-term and long-term forecasting methods have forecast horizons of one week to several years and are mainly applied for scheduling and planning [7]. On the other hand, short-term forecasting methods cover forecast horizons of few seconds to seven days and are mainly applied in real-time operations, such as economic dispatching, optimal reserves, automatic generation control, unit commitment, scheduling, and spot markets. Among the applications of short-term PV power forecasting, the day-ahead PV power forecasting is of great importance in unit commitment, storage system management, transmission scheduling, and day-ahead markets [8], [9]. For these applications, high temporal resolution and accurate solar generation forecasting is desirable but challenging. To address this challenge, similarity-based forecasting models (SBFMs) are advocated in this paper to forecast solar PV generation for the next day with high temporal resolution using low temporal resolution weather data.

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Wu.

A. RELATED WORKS

Forecasting models of PV power can be divided into three main categories: PV performance models, statistical models, and hybrid models [10]. In PV performance models, the irradiance is forecasted, first. Then, the forecasted irradiance is applied to the PV system model to estimate the generated power [11]. For the review on PV performance models, readers are referred to [12]–[14]. Although the main advantage of a PV performance model is its independence from historical data, the forecasting resolution of PV performance models depends on the resolution of a numerical weather

prediction (NWP), mainly irradiance, which is usually low (i.e., hourly) [13], [15]. On the other hand, in statistical models, the PV power generation is directly forecasted using historical PV generation data through statistical and machine learning methods [16]. Inherently, the statistical models rely on large historical data sets from observed PV power generation and weather data to train and validate the forecasting models. Some studies indicate that the statistical models can achieve higher forecasting accuracy than the PV performance models [17]. In the hybrid models, the statistical models and/or PV performance models are combined to improve forecasting accuracy [18].

Many studies have been conducted to forecast PV power generation in the next day using the three aforementioned models. In [19], an online forecasting model is proposed based on a radial basis function network (RBFN) and classification using self-organized map (SOM) to forecast hourly solar power. Similarly, in [20], the RBFN method is applied in hourly solar PV generation forecast of the next day while the support vector machine (SVM) is implemented for the classification. A hybrid model of statistical models is proposed in [21], in which a combination of an artificial neural network (ANN) based method and an autoregressive integrated moving average (ARIMA) method are applied in the day-ahead solar energy forecasting. In [22], a hybrid model, with a combination of the statistical model and a PV performance model, is proposed to forecast a day-ahead solar PV generation with hourly resolution. In [22], the ANN method is applied with hourly historical data of irradiance, power generation, and NWP to forecast hourly PV power generation of the next 24 hours. In [23], a forecasting model is proposed based on an extreme learning machine for day-ahead PV power forecasting with hourly resolution. An ensemble method is proposed in [24] for hourly day-ahead solar PV generation forecasting using several meteorological and astronomical data. The performances of four state-of-the-art forecasting methods (i.e., k-nearest neighbors (KNN), ANN, SVR, and quantile random forecast methods) to forecast solar PV power in the next 24 hours are compared in [25]. However, in these studies, the temporal resolutions of the input variables and consequent output data are one hour. One of the studies that focused on solar PV power forecasting with high temporal resolution for the next day is [26], which proposed a combination of a PV performance model and an ANN-based method for day-ahead solar PV power forecasting with one-minute resolution. However, the proposed forecasting model in [26] requires a large meteorological data set with one-minute resolution, which is often not available in high accuracy.

Accordingly, although previous studies have proposed valuable forecasting models for day-ahead solar PV generation or irradiance forecasting, they require high temporal and spatial resolutions of input data to yield high temporal resolution forecasts. Yet, both the availability and the accuracy of high temporal resolution of weather variables for the next day are the major barriers of applying these forecasting models. In this study, several SBFMs are proposed to address the challenge of forecasting high temporal resolution solar PV power while only low temporal resolution of weather data is available. we proposed forecasting models which extract indexes of days in historical data which have similar patterns to the next day forecasted weather. By extracting indexes of similar days to the next day in terms of weather variables, the forecasted solar PV power is the weighted averaged of power on similar days. The proposed SBFMs are adapted for three application scenarios that have different available weather variables.

B. CONTRIBUTIONS

The purpose of this study is to accurately forecast solar PV power with a five-minute resolution for the next day when forecasted weather data are available hourly. To fulfill this purpose, SBFMs are applied on historical power data with a five-minute resolution and weather data with a 60-minute resolution. In addition, the availability of different weather data provided by local and national weather stations is considered in the SBFM models. Three forecasting models are proposed for three scenarios of different available weather data. Finally, to quantify the forecasting accuracy, the leave-one-out cross-validation method is applied to derive forecast error metrics and to compare the proposed methods with several benchmark models.The contributions of this study are summarized as follows:

1) A basic SBFM (bSBFM) is proposed to increase the accuracy in forecasting the day-ahead PV power with high temporal resolution. This forecasting model proved to be efficient and simple for the cases where only one or some of the commonly available weather variables (i.e., *temperature*, *humidity*, *dew point*, *air pressure*, *wind speed*, and *precipitation rate*) are *available*.
2) Among the commonly available weather variables, *temperature* and *humidity* are identified as the ''best'' weather variables for the bSBFM to forecast solar PV power.
3) When the forecast of *sky cover* is available, a categorical SBFM (cSBFM) is proposed as an upgrade to the bSBFM to classify the historical data according to the *sky cover*, which improves the forecasting accuracy.
4) When the forecast of *irradiance* is available in addition to the commonly available weather variables, a hierarchical SBFM (hSBFM) is proposed to apply the similarity analysis method in multiple steps, which significantly improves the forecasting accuracy.

The rest of the paper is organized as follows. The SBFMs are proposed in Section II. Benchmark models and evaluation methods are described in Sections III. Section IV presents the case study results. The conclusion is drawn in Section V.

II. FORECASTING METHODOLOGIES

As mentioned in the previous section, to accurately forecast PV generation for the next day with 5-minute resolution, most PV forecasting models require that forecasted weather

variables also have 5-minute temporal resolution. However, the forecasted weather variables are normally available with 60-minute temporal resolution. To overcome this limitation, forecasting models based on similarity analysis are proposed in the following subsections.

To apply the similarity analysis method in forecasting models, the KNN method was chosen among other classification methods because the KNN method is simple and does not require the tuning of many parameters to yield optimal results. Thus, from this point forward, the similarity method refers to the KNN method, while similarity analysis refers to the process of applying the similarity method in the forecast models. The KNN method is chosen as the similarity analysis method. The number of neighbors (i.e., k) of the KNN is tuned in the training process. Because k is the only hyper-parameter in this study, the grid search method is applied in the training process to yield the optimal number of k neighbors.

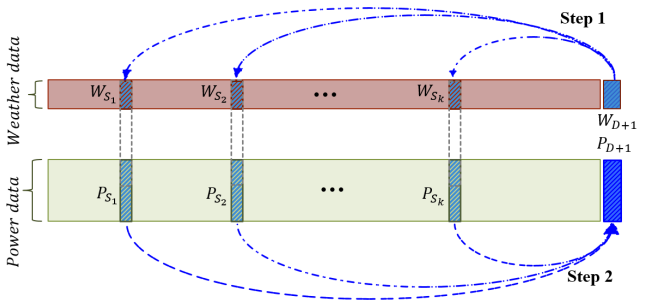


FIGURE 1. Structure of the bSBFM.

A. BASIC SBFM (bSBFM)

In bSBFM, the most commonly available weather variables including *temperature*, *humidity, dew point*, and *wind speed* are applied as predictors to forecast PV power. The bSBFM is illustrated in Fig. 1, in which blue boxes are for the forecasted weather and power data while the red boxes and green boxes represent historical weather and power data, respectively. As shown in Fig. 1, the bSBFM consists of two steps. In the first step, the similarity analysis is performed to identify the indexes of k days in the historical data (with hourly resolution) whose weather patterns are similar to the forecasted weather pattern on day D+1. More specifically, first, vectors $W_i = [V_{i,1}, V_{i,2}, \ldots, V_{i,H}]^T$ and $P_i = [p_{i,1}, p_{i,2}, \ldots, p_{i,M}]^T$ are constructed for day i where $V_{i,h}$ is the recorded weather variables (such as *temperature*, *humidity, dew point*, and *wind speed*) at hour h on day i and $p_{i,m}$ is the recorded PV generation at minute m on day i. Then, the distance between day i and day D+1 (i.e., dis_i) is calculated using (1).

$$\begin{cases} dis_i^2 = (W_i - W_{D+1})^T \times (W_i - W_{D+1}) \\ i \in [\![1, D]\!] \end{cases} \tag{1}$$

The dis_i are sorted in ascending order and the first k days whose distances are the shortest ones are selected as the similar days to day $D+1$. In Fig. 1, the indexes of the similar days are denoted as $S_1, S_2, \ldots, S_k$.

In the second step, the power on the target day is forecasted by the weighted average of the solar PV generation during the identified k similar days. More specifically, the weights are calculated as the inverse of the distances as in (2). The forecasted PV power generation on day $D+1$ (denoted by $\hat{P}_{D+1}$) is calculated using (3), where P_{s_j} represents the PV power generation on day S_j.

$$w_{s_i} = \frac{1}{dis_{s_i}} \tag{2}$$

$$\hat{P}_{D+1} = \frac{\sum_{j=1}^{k} P_{s_j} \times w_{s_j}}{\sum_{j=1}^{k} w_{s_j}} \tag{3}$$

Note that forecasting the power on the target day (i.e. P_{D+1}) using (3) requires knowing the indexes of k similar days, the assigned weight of each similar day (i.e. w_{s_i}), and historical power generation on the similar days (P_{s_i}). Accordingly, in this forecasting model, the temporal resolution of forecasted PV power is independent of the temporal resolution of the weather variables. In other words, although the weather data in (1) have hourly resolution, they are only required to derive the distance between the target day and the similar days. Thus, as illustrated in (3), once the k similar days are identified, the power on the target day (i.e. P_{D+1}) can be forecasted in five-minute resolution using the five-minute resolution historical PV power data (i.e. P_{s_j}) and the constant values of w_{s_j}.

Note that averaging generations of similar days can smooth out the variability caused by noise and improve the overall forecasting accuracy.

The bSBFM is proposed for the commonly available weather variables provided by most weather stations (i.e., numerical weather variables of *temperature*, *dew point*, *humidity*, *wind direction/speed*, *air pressure*, and *precipitation*); however, some weather stations may provide *sky cover* and *irradiance* data. Thus, two additional upgrades on the bSBFM are introduced in the following subsections. Note that to achieve higher forecasting accuracy, it is important to select the relevant weather variables that can better identify the solar PV generation patterns based on similarity analysis and exclude irrelevant or redundant variables. In this paper, the exhaustive search method was used to identify the relevant weather variables because the number of features is limited. Also, note that on the training process, different types of weather variables should be optimally scaled to yield the smallest forecasting errors.

B. CATEGORICAL SBFM (cSBFM)

When the weather variables are categorical or when weather classification on a numerical weather variable is likely to yield higher forecasting accuracy, the classification can be applied to group weather and the corresponding power generation data sets before applying the similarity method. The procedure of the proposed cSBFM is illustrated in Fig. 2 in which blue boxes are for both forecasted weather and power data while red boxes and green boxes represent historical

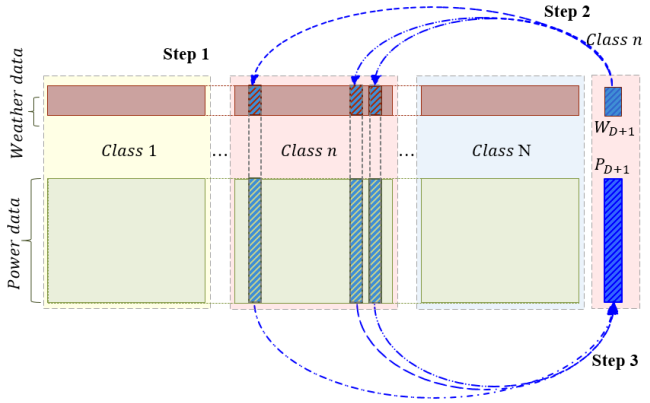


FIGURE 2. Structure of the proposed cSBFM.

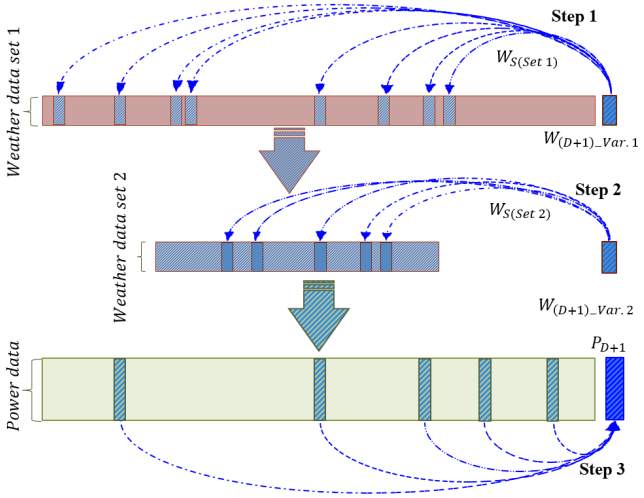


FIGURE 3. Structure of the proposed hSBFM.

weather and power data, respectively. Pink boxes in the background are for days in *Class n*. As illustrated in Fig. 2, the proposed cSBFM consists of the following three steps. In the first step, the class of forecasted weather data (i.e., W_{D+1}) is determined using a classification or clustering method. In Fig. 2, it is assumed that there are N classes and the class of the forecasted weather data is *Class n*. In the second step, the similarity analysis is performed only on the historical data at the same class as day $D+1$ (i.e., *Class n*) to identify the indexes of k similar days that are the nearest to the forecasted weather variables on day $D+1$. More specifically, the distances (i.e., dis_i) of all the days that have the same class as the target day (for this case i.e. *Class n*) are calculated using (4). Then, dis_i are sorted in ascending order. Next, the indexes of the first k days that have the shortest distance are identified as the similar days.

$$\begin{cases} dis_i^2 = (W_i - W_{D+1})^T \times (W_i - W_{D+1}) \\ i \in \{x : class(W_x) = class(W_{D+1}) = Class\, n;\ \ 1 \le x \le D\} \end{cases} \tag{4}$$

In the third step, using (3), the forecasted PV power on day $D+1$ (i.e., P_{D+1}) is estimated by averaging the power generation corresponding to the identified similar days in the historical data.

C. HIERARCHICAL SBFM (hSBFM)

Some weather service providers may supply *irradiance* data in addition to the commonly available weather variables (i.e., *temperature*, *humidity*, *dew point*, and *wind speed*). Under this scenario, a hierarchical forecasting model is proposed based on a step-by-step similarity analysis which effectively distinguishes similarity patterns in power generation and extracts the days similar to the target day. As illustrated in Fig. 3, in this forecasting model, the similarity analysis is conducted hierarchically through the following three steps. In the first step, similarity analysis is performed using weather variable set #1 to identify k_1 candidate days. More specifically, first, the distance of all days in the historical data to the target day is calculated in terms of weather variable set #1, as displayed in (5).

$$\begin{cases} (dis_i^{[V1]})^2 = (W_i^{[V1]} - W_{D+1}^{[V1]})^T \times (W_i^{[V1]} - W_{D+1}^{[V1]})^T \\ i \in [\![1, D]\!] \end{cases} \tag{5}$$

where $W_i^{[V1]}$ is a vector constructed using weather variable set #1 on day i and $W_{D+1}^{[V1]}$ represents weather variable set #1 on target day (i.e., day $D+1$). Following the procedure described in subsection II.A, k_1 days similar to the target day in terms of weather variable set #1 are selected from the historical weather data. As illustrated in Fig. 3, these k_1 selected from all weather data set (i.e., *Weather data set 1*) are *Weather data set 2*. The k_1 similar days, whose indexes are denoted as $S_1^{[V1]}, S_2^{[V1]}, \ldots, S_{k_1}^{[V1]}$, are used as the candidates in the next step.

In the second step, k_2 similar days are selected from the k_1 candidate days by conducting the similarity analysis in terms of weather variable #2 (i.e., $W_{(D+1)\_Var.2}$), as described in (6).

$$\begin{cases} (dis_i^{[V2]})^2 = (W_i^{[V2]} - W_{D+1}^{[V2]})^T \times (W_i^{[V2]} - W_{D+1}^{[V2]}) \\ i \in \{S_1^{[V1]}, S_2^{[V1]}, \ldots, S_{k_1}^{[V1]}\} \end{cases} \tag{6}$$

The results of the second step separate the indexes of the k_2 similar days that have the closest weather patterns (and PV power generation) to the target day (i.e., day $D+1$) in terms of the weather variable variable #2. The indexes are denoted as $S_1^{[V2]}, S_2^{[V2]}, \ldots, S_{k_2}^{[V2]}$. In the third step, the average of PV power generation from the k_2 identified days (illustrated in Fig. 3 as green-blue color boxes in *Power data*) are used to forecast PV power generation on day $D+1$ (i.e., P_{D+1}) using (7). Note that, w_{s_i} in (7) refers to the inverse of the distance between the target and its corresponding neighbor($S_i^{[V2]}$) in the final step of hSBFM (i.e. (8)).

$$\hat{P}_{D+1} = \frac{\sum_{j=1}^{k_2} P_{s_j} \times w_{s_j}}{\sum_{j=1}^{k_2} w_{s_j}} \tag{7}$$

$$w_{s_i} = \frac{1}{dis_i^{[V2]}} \tag{8}$$

Also, note that as it is shown in the numerical results, a forecaster should consider the best weather variable at each step of hierarchical similarity, which reflects weather conditions, seasonality, and geographical locations leading to most similar patterns in historical and target solar PV power generation. In addition, the number of steps and the number of neighbor days to the target day should be optimally selected to improve forecasting accuracy.

III. EVALUATION OF FORECASTING MODELS

A. BENCHMARK FORECASTING MODELS

As one of benchmark models, the SBFM proposed by [27], which was used to forecast solar radiation, is applied to forecast solar PV power. In this SBFM, only historical data of solar PV power is required to forecast the future generation. In this model, at first, the similarity analysis is conducted to identify k days (i.e., day S_1, S_2, …, S_k) in the historical data, whose generation patterns are similar to the current day (denoted as day D) and the k days whose distances are the shortest ones are selected as the similar days to day D.

$$\begin{cases} dis_i^2 = (P_i - P_D)^T \times (P_i - P_D) \\ i \in [\![1, D-1]\!] \end{cases} \tag{9}$$

Then, the solar PV powers that are generated one day after the identified k similar days (i.e., day $S_1 + 1$, $S_2 + 1$, …, S_k+1) are weighted and averaged to forecast the PV power on target day. The forecasted PV power generation on day $D+1$ (i.e., P_{D+1}) is calculated using (2) and (10), where P_{s_j+1} represents the PV power generation on day $S_j + 1$.

$$\hat{P}_{D+1} = \frac{\sum_{j=1}^{k} P_{s_j+1} \times w_{s_j}}{\sum_{j=1}^{k} w_{s_j}} \tag{10}$$

In addition, two other forecasting models (i.e., ANN models and persistence models) are used as the benchmarks to evaluate the forecasting accuracy and computation complexity of the proposed forecasting models. In a recent review paper on solar power forecasting, the ANN models are identified as the most commonly used forecasting model in forecasting solar PV power [10]. Simplicity in applying an ANN model as a black-box forecasting model and its efficiency in modeling the complex nonlinear relationship between target variables and predictors are the major reasons for its pervasive applications [28], [29]. Using an ANN model, the solar PV power is forecasted hourly using hourly forecasted weather data. However, to facilitate the comparison of the ANN model with the proposed models in this study, the hourly weather data is interpolated to increase its temporal resolution to five minutes. More specifically, linear interpolation is used to add 11 new estimates between two points of hourly weather data before applying in the training and testing processes of the ANN model. The results of forecasting by the ANN model using interpolated test data are compared with actual solar generation data to yield forecasting metrics.

The other benchmark forecasting model is the persistence model which uses the recent past values of a target variable to forecast its future values. Specifically, for the proposed day-ahead PV power forecasting, the persistence model forecasts the PV power of the forthcoming day (i.e., day $D+1$) to be the same as the PV power during the same time of the previous day (i.e., day D).

B. CROSS-VALIDATION OF FORECASTING MODELS

In this study, the metrics used to quantify forecast error for the day-ahead forecast include mean absolute error (MAE), mean relative error (MRE), and normalized root mean square error (nRMSE), which are defined as follows [30]–[32].

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{p}_i - p_i| \tag{11}$$

$$MRE = \frac{1}{N} \frac{\sum_{i=1}^{N} |\hat{p}_i - p_i|}{p_o} \times 100\% \tag{12}$$

$$nRMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - p_i)^2}}{p_o} \times 100\% \tag{13}$$

Here, symbols $\hat{p}_i$ and p_i are the forecasted and measured solar PV power respectively at observation i and N is the number of observations. Symbol p_o is the installation capacity of solar PV generation. Since the performance of a forecasting model is assessed by applying it on the test data set, it is a matter of importance how test data set is selected. There are three main cross-validation methods: holdout, k-fold, and leave-one-out, which adopt different ways of dividing the available data into the training data set and the validation data set [33]. The leave-one-out method is a special case of the k-fold method with k set to be the total number of observations. In the leave-one-out method, each day is considered as a test data set once to evaluate the forecasting accuracy. Accordingly, the forecast metrics are calculated for each day, and the final forecast error metrics are the average of the forecast error metrics for all the days. Note that the leave-one-out method is computationally more expensive and has higher variance than the other two cross-validation methods. However, it is less bias in estimating forecasting error. Because this study is focused on forecasting day-ahead solar generation, the training process time is not a major constraint. In this regard, the leave-one-out method is chosen which is statistically more efficient than the other two methods.

IV. NUMERICAL RESULTS

As a case study, the solar PV panels installed on the rooftop of the Engineering and Science building at the State University of New York-Binghamton University (at latitude and longitude of 42.094 and -75.958, respectively) are considered. The power generations with temporal resolution of five minutes were collected for more than two years, from September 2016 to November 2018.

While the PV power data are recorded with five-minute temporal resolution, weather forecasting data are usually

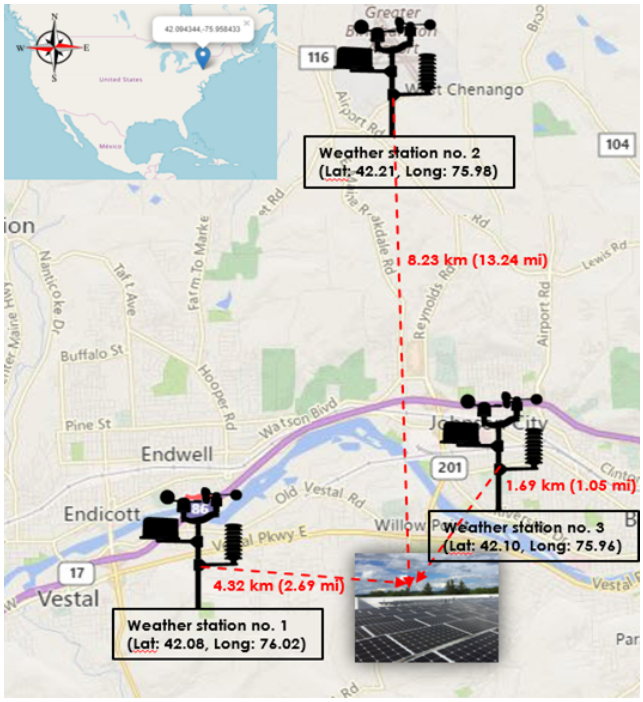


FIGURE 4. Locations of the weather stations and solar PV panels.

available with one-hour resolution. To evaluate the forecasting models in different scenarios, weather data corresponding to the historical PV power data were collected from three different weather stations, which provide different types of weather variables. As illustrated in Fig. 4, the weather stations include two local weather stations (i.e., Weather stations no. 1 and no. 3) and the National Weather Service at Binghamton Regional Airport (i.e., Weather station no. 2). While all three weather stations provide hourly weather variables of *temperature*, *dew point*, *humidity*, *wind direction/speed*, *air pressure*, and *precipitation*, Weather stations no. 2 and no. 3 also have hourly *sky cover* and *solar irradiance*, respectively.

In the training and evaluation of the forecasting models, nighttime data are excluded. In addition, the length of daytime in a target day may not be the same as other days in the historical data set. Accordingly, to ensure that the weather variables during the daytime in a target day are properly compared with the weather variables during daytime of the days in the historical data, only weather data confined to the shortest range of daytime during the entire studying term are applied in similarity analysis. For this case study, there are 630 days from September 2016 to November 2018 (123 days out of 753 days are excluded due to missing data) and there is daylight from 7:32 AM to 4:31 PM for all the days. However, because the temporal resolution of the weather data is hourly, the weather data from 8:00 AM to 4:00 PM are considered in all the forecasting models. In addition, as mentioned previously, to apply weather data variables in the SBFMs, the weather variables should be normalized. In this study, each weather variable is normalized to the Max-Min Feature scaling procedure using (14)

$$W_n = \frac{W_i - W_{min}}{W_{max} - W_{min}} \tag{14}$$

where W_n represents normalized weather variable for W_i.

TABLE 1. Day-ahead forecast of solar PV power generation using weather data from Weather station no. 1.

Forecasting model	Forecasting errors metrics		
	MAE (W)	nRMSE (%)	MRE (%)
Persistence	1309.1	25.5	17.1
ANN	1406.5	27.4	18.7
SBFM [27]	1176.0	20.9	15.3
bSBFM	**826.2**	**15.3**	**10.8**

Among the weather variables provided by Weather station no. 1, *temperature*, *dew point*, *humidity*, and *wind speed* are relevant for solar energy forecasting [34]. Using the weather data at station no. 1, the forecasting results of the day-ahead solar PV power generation with the five-minute resolution is summarized in Table 1. As is evident in Table 1, the SBFM proposed by [27] (i.e., SBFM [27]) has higher forecasting accuracy than the other two benchmark models; thus, this model is more efficient at forecasting solar PV generation in high temporal resolution when the related forecasted weather data with high temporal resolution are unavailable. Furthermore, note that the proposed bSBFM has improved the forecasting accuracy, significantly. Also, note that in the bSBFM, any influential weather variables of *temperature*, *dew point*, *humidity*, and *wind speed*, or the combination of them, can be a candidate predictor set for the model. In other words, the four weather variables of *temperature*, *dew point*, *humidity*, and *wind speed* give 15 (i.e., $2^4 - 1$) combinations of the weather variables as the predictors of the bSBFM. Since the number of choices is small, all 15 choices were tested in the model and their forecast errors were compared. Accordingly, among 15 possible combinations of these predictors, predictor set of [*temperature* , *humidity*] is the ''best'' set in terms of the smallest forecast errors. In addition, because multiple weather variable types are applied in similarity analysis, they are scaled in the training process to avoid bias of a variable over others. In this study, the exhaustive search reveals that the optimal weights of 1 and 0.375 yield the smallest forecasting errors with the predictor set of [*temperature* , *humidity*], respectively.

Weather station no. 2 provides *sky cover* data in addition to the same weather variables provided by Weather station no. 1. In this case, all weather variables are a numerical value except the *sky cover* data in Weather station no. 2, which takes categorical values of vertical visibility (VV), overcast (OVC), broken (BKN), scattered (SCT), few (FEW), and clear (CLR or SKC). In Table 2, the range of okta (i.e., cloud cover amount) for each *skycover* category is summarized.

TABLE 2. Numerical values of categorical *sky cover*.

Sky cover	Rang of okta	Percentage	Class
Clear (CLR or SKC)	0	0	1
Few (FEW)	1/8 - 2/8	25	2
Scattered (SCT)	3/8 - 4/8	50	3
Broken (BKN)	5/8 - 7/8	75	4
Overcast (OVC)	8/8	100	5
Vertical Visibility (VV)	8/8	100	5

In this situation, the cSBFM is applied to the data provided by this weather station along with clustering on *sky cover*.

In the cSBFM, for a target day with a defined *sky cover*, similarity analysis is applied only to the historical weather data that have the same *sky cover* class. Therefore, hourly categorical *sky cover* is quantified into numerical values as displayed in Table 2, and the average *sky cover* on each day results in the corresponding *sky cover* class for that day. Note that because the range of VV and OVC are the same in terms of cloud coverage rate, they are considered in the same class (Class 5). Accordingly, each day in training and test data sets is assigned to one of the five classes described in Table 2 and the forecasting model is trained and validated only on the one corresponding target day's class.

TABLE 3. Day-ahead forecast of solar PV power generation using weather data from Weather station no. 2.

Forecasting model	Forecasting errors metrics		
	MAE (W)	nRMSE (%)	MRE(%)
Persistence	1308.2	25.5	17.1
ANN	1335.8	26.2	17.1
SBFM [27]	1083.6	18.4	13.5
bSBFM	818.2	14.9	10.5
cSBFM	**806.1**	**13.9**	**9.5**

The forecasting results using the data from Weather station no. 2 are summarized in Table 3. Consistent with the forecasting results using the data from weather station no.1, the [*temperature*, *humidity*] predictor set is the "best" predictors among all the most common weather variables (i.e., *temperature*, *dew point*, *humidity*, *wind speed*) and their combination sets. In addition, the forecasting accuracy is improved by including *sky cover*. As displayed in Table 3, the cSBFM which applies the classification based on *sky cover*, and uses predictors of [*temperature*, *humidity*], results in a lower forecast errors than all the other models.

Unlike Weather station no. 2, Weather station no. 3 provides hourly *irradiance* data instead of *sky cover* data. Upon this condition, weather station no. 3 supplies hourly *irradiance* data in addition to the commonly available weather variables of *temperature*, *dew point*, *humidity*, and *wind speed*.

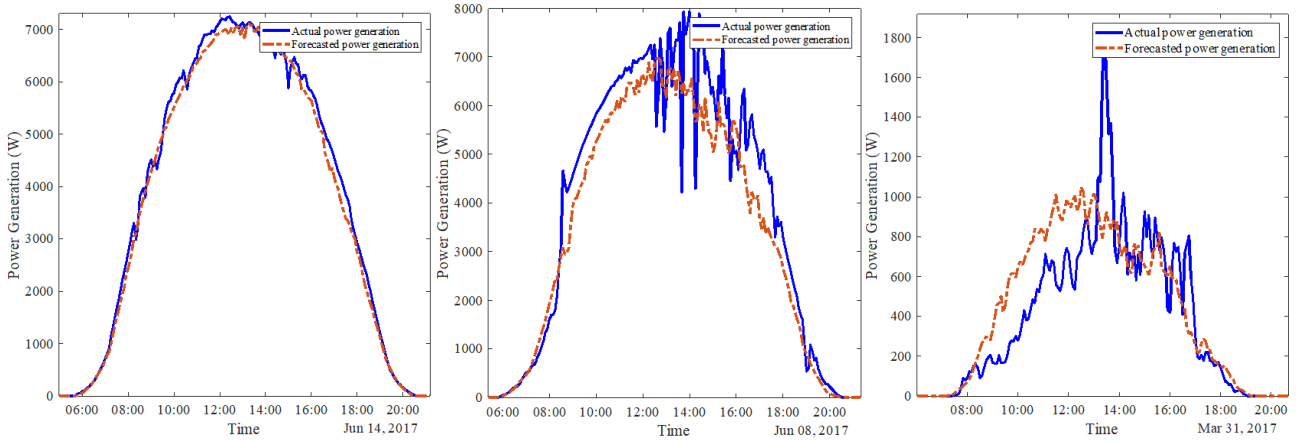
TABLE 4. Day-ahead forecast of solar PV power generation using weather data from Weather station no. 3.

Forecasting model	Forecasting errors metrics		
	MAE (W)	nRMSE (%)	MRE(%)
Persistence	1282.3	25.1	16.8
ANN	1289.4	25.2	16.8
SBFM [27]	1180.5	21.1	15.4
bSBFM	733.9	14.3	9.6
hSBFM	**618.3**	**10.6**	**7.6**

Table 4 shows the day-ahead forecast of the solar PV power generation in five-minute resolution using the benchmark forecasting models (i.e. Persistent, ANN, and SBFM [27]), bSBFM, and hSBFM. Note that in Table 4, for the bSBFM using the data from Weather station no. 3, the predictors of [*temperature*, *humidity, irradiance*] are the "best" predictors, which suggests that the *irradiance* is beneficial to the forecasting accuracy. Also note that, in the hSBFM, the [*temperature* , *humidity*] as the first variable set, then *irradiance* as the second variable set are applied, which results in significant improvement in the forecasting accuracy. In other words, the best forecasting results with the data of Weather station no. 3 are achieved using the hSBFM which first applies the similarity analysis to extract the indexes of some candidate days using [*temperature* , *humidity*], and then applies the final similarity analysis using *irradiance*. Note that other possible combinations of predictors in all the SBFMs, including hierarchical and non-hierarchical models, were assessed, which yielded larger forecast errors. In addition, in the hSBFM, more than two-steps models were also considered, but these have not yielded better forecasting results than the hSBFM with the two steps of [*temperature*, *humidity*], and then *irradiance*.

As mentioned, the input weather data of the bSBFM in Table 4 is the predictor set of [*temperature*, *humidity, irradiance*]; however, the simulation shows that the bSBFM using predictors of [*temperature*, *humidity*] results in MAE, nRMSE, and MRE equal to 810.6, 14.6, and 10.2, respectively. These forecasting error metrics are still smaller than the forecasting error metrics of the bSBFMs using data from Weather station no. 2 (shown in Table 3) and Weather station no. 1 (shown in Table 1). The reason for the forecasting results improvement, with the same forecasting model and input variables, is referred to the improvement in the accuracy of weather data as the Weather station no. 3 is the closest weather station to the solar panels.

Fig. 5 illustrates the performance of the proposed hSBFM in three different weather patterns of sunny, partially_cloudy, and cloudy days. In Fig. 5a, which represents power generation on a sunny day, the solar power output is relatively consistent, with no abrupt change, and the hSBFM model can accurately forecast the PV generation with MAE, nRMSE, and MRE of 159.91 W, 2.89%, and 2.25%, respectively.



(a) Forecasted PV power on a sunny day (b) Forecasted PV power on a partially_cloudy day (c) Forecasted PV power on a cloudy day

FIGURE 5. Forecasted PV power using the hSBFM for different weather patterns.

Fig. 5b illustrates the forecasted PV power on a partially cloudy day. As illustrated, the solar power generation ramps smoothly in the morning and becomes bumpy during the afternoon when the weather pattern dramatically changes. Despite these abrupt changes in the middle of the day, the forecasting model can satisfactorily forecast actual power generation with MAE, nRMSE, and MRE of 578.95 *W*, 10.81%, and 8.29%, respectively. Fig. 5c illustrates the performance of the proposed hSBFM on a cloudy day, during which PV power is low along with dramatic variability. However, it is apparent that the hSBFM can predict the actual solar power generation relatively well with the MAE, nRMSE, and MRE of 739.27 *W*, 14.16%, and 12.46%, respectively.

TABLE 5. Comparison of different SBFMs in terms of required inputs, computation time, and forecast errors.

Forecasting model	Inputs*	Time (*ms*)	MAE (*W*)
SBFM [27]	Pg	2.8	1176.0
bSBFM	Pg, Tm, Hm	3.2	826.2
cSBFM	Pg, Tm, Hm, Sc	3.5	806.1
hSBFM	Pg, Tm, Hm, Ir	4.2	618.3

* Pg: Power generation, Tm: *Temperature,* Hm: *Humidity,* Sc: *Sky cover*, Ir: *Irradiance*

Table 5 compares different SBFMs in terms of required predictors, processing time, and resulted forecast error metrics in MAE. In this table, the results for the SBFM [27] and bSBFM come from the forecasting studies using the data from Weather station no. 1 while the results for the cSBFM and hSBFM come from the forecasting studies using the weather data from Weather stations no. 2 and no. 3, respectively. The study was conducted on a personal computer (PC) with Intel® Core™i7-CPU @3.5 GHz and 16 GB of RAM. In this table, the time refers to the computation time used for forecasting PV generation just for a sample day. It is evident that the SBFM [27] uses the shortest computation time of 1.6 *ms* while the hSBFM, which includes two steps of similarity analysis, uses the longest computation time of 4.2 *ms*. Considering the forecast horizon is the next day, the computation time of all the models is short enough to achieve real-time forecasting. In addition, among these models, the SBFM [27] is the only model which does not require weather data. However, this is also the model with the lowest forecasting accuracy. Moreover, this study reveals that two to three weather variables are sufficient to improve forecasting accuracy.

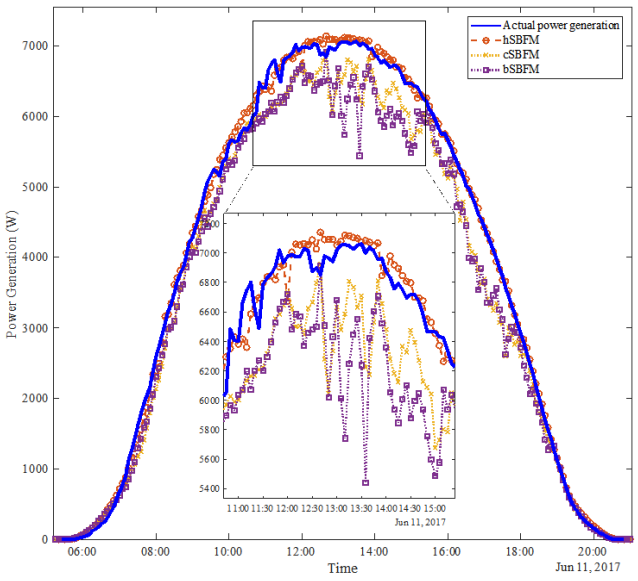


FIGURE 6. Forecasting results of the three SBFMs on a random day.

In Fig. 6, a random day (i.e., a sunny day on June 11, 2017) is selected to illustrate the performance of the bSBFM, cSBFM, and hSBFM. Observe that the hSBFM can forecast the solar PV power generation in a higher accuracy

TABLE 6. Forecast error metrics of the three SBFMs on a random sunny day.

Forecasting model	Forecasting errors metrics		
	MAE (W)	nRMSE (%)	MRE(%)
bSBFM	352.9	6.67	5.24
cSBFM	311.1	5.67	4.4
hSBFM	71.9	1.43	1.02

than the other models. On this sample day, the forecast error metrics for the hSBFM are 71.9 W, 1.43%, and 1.02% for MAE, nRMSE, and MRE, respectively. A comparison of the three models is shown in Table 6. As shown in this table, the bSBFM, which refers to the bSBFM using [*temperature*, *humidity*], has the largest forecast error. While the cESBFM has slightly smaller forecast errors than the bSBFM, the hSBFM has significantly improved the forecasting accuracy.

Note that all the methodologies and forecasting models elaborated in this study can also be applied when the temporal resolution of PV power generation is as low as the temporal resolution of weather data in an hourly resolution. Inherently, with the same forecasting model, the results of the solar PV generation forecasting in hourly resolution are more accurate than forecasting in minute(s) resolution. The reason is sought in the pattern of solar generation in hourly resolution which is more smooth than the pattern of solar generation in 5-minute resolution. As an example, by applying the hSBFM to forecast the hourly PV generation using the previous data sets, the MAE, nRMSE, and MRE of the forecasting results are 451.5, 7.8%, and 5.7%, respectively while the MAE, nRMSE, and MRE for the forecasting with five-minute temporal resolution are 618.3, 10.6%, and 7.6%, respectively.

V. CONCLUSIONS

Accurate forecasting of solar PV generation for the next day in the temporal resolution of minutes is challenging to achieve accurate forecasting results with most forecasting models because normally, the correlation of PV power is not stationary over the next day horizon and the temporal resolution of weather forecasting is low. To overcome this challenge, several SBFMs were proposed in this paper to increase forecasting accuracy. First, the bSBFM was proposed to forecast solar PV power generation during the next day with the temporal resolution of 5 minutes when only one or some of the commonly available weather variables (i.e., *temperature*, *humidity*, *dew point*, and *wind speed*) are available. The study results based on a small-scale solar PV system reveal that the proposed bSBFM is efficient, simple, and more accurate than the benchmark models. In addition, the study results indicate that from the available weather variables, the weather variable data sets of *temperature* and *humidity* yield the most accurate forecasting results using the bSBFM. Then, the study was extended for two other cases, in which one weather station provided categorical values of *sky cover* data and another weather station provided numerical values of *irradiance* data in addition to the commonly available weather data. Leveraging these data, two upgraded forecasting models (i.e., the cSBFM and hSBFM) were proposed. The study results show that the forecasting accuracy of the cSBFM is slightly better than the bSBFM using the *sky cover* data. However, the forecasting accuracy of the hSBFM is significantly better than the bSBFM and cSBFM where *temperature*, *humidity*, and *irradiance* are applied hierarchically in the similarity analysis. In other words, the PV power generation forecasting for the next day with five-minute temporal resolution can yield significantly accurate results if the weather variables of *temperature* and *humidity* are applied at the first level of the proposed hSBFM and *irradiance* data are applied at the second level.

REFERENCES

[1] H. Sadeghian and Z. Wang, "A novel impact-assessment framework for distributed PV installations in low-voltage secondary networks," *Renew. Energy*, vol. 147, pp. 2179–2194, Mar. 2020.

[2] M. Mahoor, A. Majzoobi, and A. Khodaei, "Distribution asset management through coordinated microgrid scheduling," *IET Smart Grid*, vol. 1, no. 4, pp. 159–168, Dec. 2018.

[3] H. Sadeghian and Z. Wang, "Decentralized demand side management with rooftop PV in residential distribution network," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2018, pp. 1–5.

[4] A. Dehghan-Banadaki, T. Taufik, and A. Feliachi, "Big data analytics in a day-ahead electricity price forecasting using TensorFlow in restructured power systems," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2018, pp. 1–5.

[5] U. K. Das, K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. Van Deventer, B. Horan, and A. Stojcevski, "Forecasting of photovoltaic power generation and model optimization: A review," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 912–928, Jan. 2018.

[6] A. Zare-Noghabi, M. Shabanzadeh, and H. Sangrody, "Medium-term load forecasting using support vector regression, feature selection, and symbiotic organism search optimization," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Atlanta, GA, USA, Aug. 2019, pp. 1–5.

[7] H. Sangrody, N. Zhou, and X. Qiao, "Probabilistic models for daily peak loads at distribution feeders," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Chicago, IL, USA, Jul. 2017, pp. 1–5.

[8] C. Pan and J. Tan, "Day-ahead hourly forecasting of solar generation based on cluster analysis and ensemble model," *IEEE Access*, vol. 7, pp. 112921–112930, 2019.

[9] G. Notton, M.-L. Nivet, C. Voyant, C. Paoli, C. Darras, F. Motte, and A. Fouilloy, "Intermittent and stochastic character of renewable energy sources: Consequences, cost of intermittence and benefit of forecasting," *Renew. Sustain. Energy Rev.*, vol. 87, pp. 96–105, May 2018.

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

[11] E. Lorenz, J. Hurka, G. Karampela, D. Heinemann, H. G. Beyer, M. Schneider, "Qualified Forecast of ensemble power production by spatially dispersed grid-connected PV systems," *Measurement*, pp. 1–7, Oct. 2007.

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

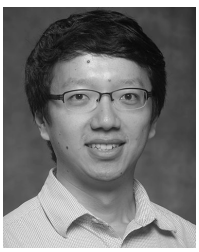
[13] A. Dolara, S. Leva, and G. Manzolini, "Comparison of different physical models for PV power output prediction," *Sol. Energy*, vol. 119, pp. 83–99, Sep. 2015.

[14] E. B. Ssekulima, M. S. El Moursi, A. Al Hinai, and M. B. Anwar, "Wind speed and solar irradiance forecasting techniques for enhanced renewable energy integration with the grid: A review," *IET Renew. Power Gener.*, vol. 10, no. 7, pp. 885–989, Aug. 2016.

[15] R. A. Verzijlbergh, P. W. Heijnen, S. R. de Roode, A. Los, and H. J. J. Jonker, "Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications," *Sol. Energy*, vol. 118, pp. 634–645, Aug. 2015.
[16] H. Sangrody, M. Sarailoo, N. Zhou, N. Tran, M. Motalleb, and E. Foruzan, "Weather forecasting error in solar energy forecasting," *IET Renew. Power Gener.*, vol. 11, no. 10, pp. 1274–1280, Aug. 2017.
[17] G. Graditi, S. Ferlito, and G. Adinolfi, "Comparison of photovoltaic plant power production prediction methods using a large measured dataset," *Renew. Energy*, vol. 90, pp. 513–519, May 2016.
[18] A. G. R. Vaz, B. Elsinga, W. G. J. H. M. van Sark, and M. C. Brito, "An artificial neural network to assess the impact of neighbouring photovoltaic systems in power forecasting in Utrecht, The Netherlands," *Renew. Energy*, vol. 85, pp. 631–641, Jan. 2016.
[19] C. Chen, S. Duan, T. Cai, and B. Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," *Sol. Energy*, vol. 85, no. 11, pp. 2856–2870, Nov. 2011.
[20] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Trans. Ind. Appl.*, vol. 48, no. 3, pp. 1064–1069, May 2012.
[21] P. Ramsami and V. Oree, "A hybrid method for forecasting the energy output of photovoltaic systems," *Energy Convers. Manage.*, vol. 95, pp. 406–413, May 2015.
[22] S. Leva, A. Dolara, F. Grimaccia, M. Mussetta, and E. Ogliari, "Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power," *Math. Comput. Simul.*, vol. 131, pp. 88–100, Jan. 2017.
[23] M. Hossain, S. Mekhilef, M. Danesh, L. Olatomiwa, and S. Shamshirband, "Application of extreme learning machine for short term output power forecasting of three grid-connected PV systems," *J. Cleaner Prod.*, vol. 167, pp. 395–405, Nov. 2017.
[24] X. Zhang, Y. Li, S. Lu, H. F. Hamann, B.-M. Hodge, and B. Lehman, "A solar time based analog ensemble method for regional solar power forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 268–279, Jan. 2019.
[25] L. Gigoni, A. Betti, E. Crisostomi, A. Franco, M. Tucci, F. Bizzarri, and D. Mucci, "Day-ahead hourly forecasting of power generation from photovoltaic plants," *IEEE Trans. Sustain. Energy*, vol. 9, no. 2, pp. 831–842, Apr. 2018.
[26] Y. S. Manjili, R. Vega, and M. M. Jamshidi, "Data-analytic-based adaptive solar energy forecasting framework," *IEEE Syst. J.*, vol. 12, no. 1, pp. 285–296, Mar. 2018.
[27] A. Boilley, C. Thomas, M. Marchand, E. Wey, and P. Blanc, "The solar forecast similarity method: A new method to compute solar radiation forecasts for the next day," *Energy Procedia*, vol. 91, pp. 1018–1023, Jun. 2016.
[28] H. Sangrody, N. Zhou, S. Tutun, B. Khorramdel, M. Motalleb, and M. Sarailoo, "Long term forecasting using machine learning methods," in *Proc. IEEE Power Energy Conf. Illinois (PECI)*, Chicago, IL, USA, Feb. 2018, pp. 1–5.
[29] H. Sangrody and N. Zhou, "An initial study on load forecasting considering economic factors," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Boston, MA, USA, Jul. 2016, pp. 1–5.
[30] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renew. Energy*, vol. 105, pp. 569–582, May 2017.
[31] B. Khorramdel, H. Khorramdel, A. Zare, N. Safari, H. Sangrody, and C. Y. Chung, "A nonparametric probability distribution model for short-term wind power prediction error," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2018, pp. 1–5.
[32] J. Zhang, B.-M. Hodge, A. Florita, S. Lu, H. F. Hamann, and V. Banunarayanan, "Metrics for evaluating the accuracy of solar power forecasting," Nat. Renew. Energy Lab., Golden, CO, USA, Tech. Rep. NREL/CP-5000-60142, 2013.
[33] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, Montreal, QC, Canada, 1995, pp. 1137–1145.
[34] H. Sangrody, M. Sarailoo, N. Zhou, A. Shokrollahi, and E. Foruzan, "On the performance of forecasting models in the presence of input uncertainty," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 2017, pp. 1–6.

HOSSEIN SANGRODY (Student Member, IEEE) is currently pursuing the Ph.D. degree in electrical engineering with the State University of New York (SUNY) at Binghamton. Prior to joining the SUNY at Binghamton, he worked with the Niroo Research Institute. His research interests include predictive modeling, applied machine learning, time series analysis, and integration of renewable energies in power systems.

NING ZHOU (Senior Member, IEEE) received the Ph.D. degree in electrical engineering with a minor in statistics from the University of Wyoming, in 2005. From 2005 to 2013, he worked as a Power System Engineer with the Pacific Northwest National Laboratory. He is currently with the Department of Electrical and Computer Engineering, Binghamton University, as an Associate Professor. His research interests include power system dynamics and statistical signal processing. He is a Senior Member of the IEEE Power and Energy Society (PES).



ZIANG ZHANG (Member, IEEE) received the B.S. degree from the Beijing Institute of Technology, the M.S. degree from Purdue University Northwest, in 2009, and the Ph.D. degree in electrical engineering from North Carolina State University, in 2013. He joined the Department of Electrical and Computer Engineering, Binghamton University—SUNY, as an Assistant Professor, in 2014. Before he joined Binghamton University, he worked at the ABB Corporate Research Center, Raleigh, NC, USA. His current research interests include renewable energy integration, transient stability analysis, battery systems operation, and distributed control algorithms. He is a member of the Smart Grid task Force of the IEEE Industrial Electronics Society and a Task Force Member of the Decision Support Tools for Energy Storage Investment and Operations of the IEEE Power and Energy Society. He is an Associated Editor of IEEE Access.