IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 21, 2020, accepted September 7, 2020, date of publication September 18, 2020, date of current version October 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024901

Short-Term Photovoltaic Power Forecasting Using an LSTM Neural Network and Synthetic Weather Forecast

MOHAMMAD SAFAYET HOSSAIN, (Graduate Student Member, IEEE), AND HISHAM MAHMOOD, (Member, IEEE)
Department of Electrical and Computer Engineering, Florida Polytechnic University, Lakeland, FL 33805, USA
Corresponding author: Hisham Mahmood (hmahmood@floridapoly.edu)

ABSTRACT In this paper, a forecasting algorithm is proposed to predict photovoltaic (PV) power generation using a long short term memory (LSTM) neural network (NN). A synthetic weather forecast is created for the targeted PV plant location by integrating the statistical knowledge of historical solar irradiance data with the publicly available type of sky forecast of the host city. To achieve this, a *K*-means algorithm is used to classify the historical irradiance data into dynamic type of sky groups that vary from hour to hour in the same season. In other words, the types of sky are defined for each hour uniquely using different levels of irradiance based on the hour of the day and the season. This can mitigate the performance limitations of using fixed type of sky categories by translating them into dynamic and numerical irradiance forecast using historical irradiance data. The proposed synthetic weather forecast is proved to embed the statistical features of the historical weather data, which results in a significant improvement in the forecasting accuracy. The performance of the proposed model is investigated using different intraday horizon lengths in different seasons. It is shown that using the synthetic irradiance forecast can achieve up to 33% improvement in accuracy in comparison to that when an hourly categorical type of sky forecast is used, and up to 44.6% in comparison to that when a daily type of sky forecast is used. This highlights the significance of utilizing the proposed synthetic forecast, and promote a more efficient utilization of the publicly available type of sky forecast to achieve a more reliable PV generation prediction. Moreover, the superiority of the LSTM NN with the proposed features is verified by investigating other machine learning engines, namely the recurrent neural network (RNN), the generalized regression neural network (GRNN) and the extreme learning machine (ELM).

INDEX TERMS PV power forecasting, machine learning, LSTM, neural network, deep learning, synthetic weather forecast.

I. INTRODUCTION

Solar PV generation is one of the most promising renewable energy resources that are expected to mitigate the climate change crisis and improve global energy security. Atmospheric variables, such as solar irradiance, temperature, humidity and cloud properties, directly and indirectly influence PV power generation. These variables make PV generation intermittent and stochastic. Therefore, large-scale PV power penetration in the utility grid requires reliable forecasting models to operate the power grid economically and reliably [1], [2]. The short term PV power forecast, which extends from an hour ahead to 24 hours ahead, is essential for a secured grid operation [1]. On the other hand, long-term forecasting horizons extend from one month to one year, which is used for long term planning [3]. The physical behavior and the time series nature of PV power generation are explored using various types of forecasting models in the state of the art literature. Statistical models use historical data of PV power generation, whereas physical models utilize satellite imagery sources [2]. Artificial intelligence (AI) based models use neural networks, and other machine learning techniques, to capture the stochastic nature of the PV power time series [4]. Recently, these models are combined together and proposed as hybrid models [5]. A prediction performance review of machine learning, mathematical, and

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

hybrid forecasting models is presented in [1]. According to this comprehensive review, the hybrid models show superiority in comparison to models that only use machine learning or mathematical techniques [1]. A hybrid model employing a genetic algorithm (GA) based weight optimization for a support vector machine (SVM) is proposed for a single-step prediction in [6]. A single-step ahead forecasting algorithm combining meta-heuristic optimization and back propagation neural network (BPNN) is proposed in [7]. A probabilistic forecasting approach, which utilizes quantile regression and an ELM, is proposed in [8] for single-step ahead PV power forecasting. The prediction models used in [7] and [8] show high accuracy as the PV generation sampling resolution is only 5 min, which limits their applications to certain real-time grid operations. LSTM NN based models outperform the conventional neural network based models when the sampling resolution is more than 15-min as shown in [9]. The model proposed in [9] employs the attention mechanism and LSTM NN for single-step ahead forecasting of PV generation, while exploring 7.5 min to 1 hour sampling resolutions. The techniques in [6]–[9] rely mainly on the PV generation time series, which can achieve reasonable accuracy in the case of single-step ahead predictions. However, for multi-step ahead predictions, reliance on the generation time series only may result in inadequate performance.

A deep convolutional neural network (CNN) is used in [10] to extract features from the PV power time series. Thereafter, the extracted features and the time series of the weather variables are fed to a support vector regression (SVR) network to forecast PV generation over intraday horizons. However, the direct use of PV power time series gradient is not exploited in this model [11]. An effective algorithm is implemented in [12] using deep NNs and atmospheric data to predict the total generation of the whole next day. However, in recent years, models that predict accumulated generation become less popular in some countries due to the penalties enforced by operators on producers who fail to report accurate generation forecast [5]. Therefore, hourly day ahead and hourly intraday models are adopted to obtain a more accurate prediction, and achieve the intended economic benefit in the energy markets [13]. A hybrid model combining a wavelet transform, a deep neural network (DNN), and an LSTM NN, utilizes temperature data to predict next multiple time steps of PV generation in [14]. The proposed model adopts recursive multi-step ahead forecasting method to predict 12-hour horizons.

In recursive prediction methods, the value of the forecast horizon first step is predicted from direct observations, and then used recursively to predict the rest of the steps in the prediction horizon. On the other hand, the entire prediction horizon is estimated at once, from direct observations, in direct multi-step prediction methods. The direct multi-step ahead forecasting method is more efficient in short-term forecasting, in comparison to the recursive one [15]. A combination of a wavelet transformation and a feed forward neural network (FNN) is used in [16] to forecast day ahead hourly PV generation by utilizing historical weather data. Differential evolution (DE) and particle swarm optimization (PSO) are used to optimize a mathematical forecasting model in [17] to predict a 4-hour horizon PV generation. In general, hybrid models can achieve higher prediction accuracy when incorporating weather forecast data [18]. An LSTM NN based PV power forecasting algorithm is proposed in [19] to predict intraday and 24-hour horizons using a time index as an additional input feature along with the relevant weather variables. A DNN is used in [20] to predict next 24-hour PV generation based on historical weather information and a rolling horizon strategy. The forecasting accuracy of the models proposed in [14], [16], [17], [19], [20] can still be enhanced by considering weather forecast data [18].

Therefore, recent PV power forecasting models use weather forecast data effectively for hourly day ahead predictions. A BPNN based day ahead forecasting model that uses the weather aerosol index as an additional predictor is proposed in [21]. The weather forecast data of a targeted day is used to choose the training samples of the forecasting model. However, the conventional feed forward NN is proven to have its own accuracy limitations with time series prediction, in comparison to the more advanced memory-based neural networks, such as recurrent or LSTM networks. A day ahead hourly prediction algorithm using SVR, self-organizing map (SOM), and learning vector quantization (LVQ) classifiers is proposed in [22]. The 15-hour prediction horizon is divided into five 3-hour segments. On the other hand, a one out of six prediction sub-models is used at a time, depending on the weather conditions. Weather forecast data are utilized with a fuzzy inference machine to decide the sub-model suitable for every 3-hour horizon segment. The algorithm uses PV generation data of previous similar days, while ignoring the same day most recent time series trend.

A combination of a radial basis function neural network (RBFNN) and a fuzzy K-means algorithm is used to develop five models for five types of days [23]. The temperature and precipitation forecast data of a targeted day are used to choose the corresponding model to predict the day ahead hourly PV generation. However, as will be shown in Section II, solar irradiance is more correlated to the type of the day, with respect to PV generation, than temperature and precipitation. The weather classification and SVM are used to develop four PV forecasting models corresponding to four types of days in [24]. The type of the day is selected based on the weather forecast of the day, such as sunny, cloudy, etc. These weather forecasts are mostly available for large geographical area, rather than for a specific PV plant location. This can only give a rough insight into the type of the day which helps in classifying the training data.

Inspired by the previous research efforts, an algorithm is proposed in this paper to leverage the powerful time series processing features of LSTM NNs with a synthesized approximate weather forecast, to predict intraday and day-ahead horizons. The statistical knowledge gained from the historical irradiance data of a PV plant location is integrated with the publicly available type of sky forecast to create a synthetic solar irradiance forecast. A *K*-means algorithm is used to classify the historical irradiance data into dynamic type of sky groups that vary from hour to hour in the same season. In other words, the types of sky are defined for each hour uniquely, using different levels of irradiance, depending on the hour of the day, and the season. The synthetic weather variables, the historical weather and PV generation time series, the time of the day index and the month of the season index are used as input features for the LSTM NN. The performance of the synthetic weather forecast variables as input features is compared to that of the categorical hourly and daily city type of sky forecast. It is shown that using the synthetic irradiance forecast can achieve up to 33% improvement in accuracy in comparison to that when an hourly type of sky forecast is used, and up to 44.6% in comparison to that when a daily type of sky forecast is used. Finally, the proposed algorithm is implemented using other forecasting engines, namely the RNN, the GRNN, and the ELM, to verify the superiority of LSTM NN with the proposed input features.

The rest of the paper is organized as follows. The problem statement and data mining approach are discussed in Section II. The forecasting framework based on LSTM NN is presented in Section III. Section IV presents the performance evaluation metrics. The simulation results are presented and discussed in Section V. Finally, Section VI concludes the paper.

II. PROBLEM STATEMENT AND DATA MINING APPROACH

As mentioned earlier, recent PV power forecasting models take advantage of available weather forecasting utilities. The weather forecast for a city area is available, in hourly and daily resolutions, on public weather websites. The weather variables including temperature, humidity and wind speed are almost homogeneous around the city areas. However, solar irradiance varies from a location to another in the same city due to the cloud effect. Moreover, most weather channels provide hourly and/or daily type of sky categories, rather than the time series of the numerical irradiance forecast. This categorical variable can only give a rough and static insight into the irradiance level since the same sky type can refer to different levels of irradiance depending on the hour of the day, and on the season. Accordingly, the categorical type of sky forecast for a city may not result in an accurate PV generation forecast for a specific plant location. Creating a synthetic numerical irradiance forecast, that can be associated with dynamic types of sky, can be a good starting point towards improving the prediction accuracy. The creation of synthetic weather profiles is considered as a part of the data mining approach of the proposed model. Data mining is very popular in solving time series regression and classification types of problems [25]. In this paper, the data mining approach includes a data-set preparation, a correlation analysis, a statistical analysis, and a synthetic weather forecast preparation. The prepared data set is used with an LSTM NN to predict the PV power generation.

A. APPROACH OVERVIEW

Collecting historical PV generation and weather data, for the considered PV plant, is the first step of the data preparation. The hourly historical weather data, of the Desoto solar farm (25 MW) and the city of Arcadia in Florida, are collected from the national renewable energy laboratory (NREL) website for the period of 2012-2018 [26]. The data-set comprises solar irradiance, temperature, wind speed, precipitable water, pressure and relative humidity. The historical generation and weather data are divided into four seasons, marked as spring (Mar-May), summer (Jun-Aug), autumn (Sep-Nov), and winter (Dec-Feb). A correlation analysis is conducted to choose the suitable predictors for the proposed model. This is created using statistical analysis of the available historical data in this step.

To achieve this, it is proposed to classify the historical irradiance data into dynamic types of sky for every hour of the day in the same season. Therefore, the types of sky are defined for each hour uniquely, using different levels of irradiance, based on the hour of the day and the season. Accordingly, every hour of the day, e.g. 10:00 AM-11:00 AM, has multiple irradiance clusters that are different from the other hours of the day. This is executed using a *K*-means algorithm.

The irradiance clusters are associated with the categorical types of sky provided by weather channels, which correspond to the standard sky condition categories set by the National Oceanic and Atmospheric Association (NOAA) [27]. For each hour of the prediction horizon, the categorical type of sky forecast, for the whole city area, is used to determine the corresponding irradiance cluster. The cluster center points, calculated from the historical data, represent the numerical value of the approximate synthetic forecast for that specific location at that specific hour. The city weather forecast variables including temperature, wind speed and humidity are used directly in the synthetic weather forecast profile. Finally, the historical weather data, the synthetic weather forecast data, the historical PV power time series, and other categorical indices are used as input features for the LSTM NN, as will be discussed in detail in Section III.

B. CORRELATION BETWEEN ATMOSPHERIC VARIABLES AND PV POWER GENERATION

PV power generation is influenced, at different levels, by atmospheric parameters such as solar irradiance (GHI), temperature (T), wind speed (WS), precipitable water (PW), relative humidity (RH) and pressure (P). The Pearson product moment correlation coefficient (PPMCC) method is adopted to calculate the correlation between each of the weather variables and the PV power (P_{PV}) generation [21]. The correlation coefficient (ξ) of two vectors, e.g. x and y, is calculated as

$$\xi = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \overline{y})^2}} \tag{1}$$

TABLE 1. **Correlation between PV power generation and weather variables.**

	GHI	*T*	*WS*	*PW*	*RH*	*P*
P_{PV}	0.9989	0.5434	0.3134	-0.0054	-0.6932	-0.0125

Time series of the PV power and the aforementioned atmospheric variables are considered as vectors, and the correlation coefficients are calculated using Equation (1). As can be seen from Table 1, solar irradiance, temperature, relative humidity and wind speed show a strong correlation with PV power generation, which makes them a good choice for the model input features.

C. STATISTICAL ANALYSIS

The solar irradiance is a stochastic time series signal with characteristics that vary hourly, daily, and seasonally. The NOAA categorizes sky conditions, into five types, as sunny, mostly sunny, partly sunny/partly cloudy, mostly cloudy and cloudy [27]. The aforementioned types of sky are provided by most weather forecast utilities for each hour. The objective of the statistical analysis is to translate these categories into their corresponding solar irradiance clusters using the historical data collected from the geographic location under consideration. The solar irradiance time series of each season, given as $X = [x_1, x_2, \ldots, x_{nm}]$, is rearranged to create an $n \times m$ matrix by grouping and aligning same time step data points as follows:

$$\widehat{X} = \begin{pmatrix} x_1 & x_2 & \ldots & x_m \\ x_{m+1} & x_{m+2} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{(n-1)m+1} & x_{(n-1)m+2} & \ldots & x_{nm} \end{pmatrix} \qquad (2)$$

where, m represents the total time steps in a day, i.e. 24 in this paper, and n represents the number of days in a season. Thereafter, the K-means algorithm is applied to the irradiance data subset of every hour to cluster the observations. The K-means algorithm clusters the data according to the minimum distance of each data point from randomly selected k number of center points [25]. The algorithm assigns each observation to its nearest cluster with a center point C_k which is calculated as follows:

$$C_k = \frac{1}{n}\sum_{j=1}^{n} z_j^k \qquad (3)$$

where, z_j^k is the j^{th} observation of the k^{th} cluster which contains n number of data points. In this paper, the number of clusters is five, corresponding to the NOAA's five classes. The five center points for each hour in a day are determined and kept in record. The data preparation is detailed in Algorithm 1. The K-means center points vary with the hour of the day as shown in Fig. 1. Each hour of a day has different five center points that are associated with the five types of sky. To clearly show the dynamic variation of the clusters for different hours of the day, the probability distribution of the irradiance is estimated for each cluster of each hour of the day [28]. The irradiance probability distributions for the five clusters of hours 8:00 AM, 11:00 AM, and 3:00 PM are shown in Fig. 2. It is worth mentioning that the statistical analysis is performed with keeping the testing data set excluded.

Algorithm 1 Data Preparation and K-Means

```
Input: $X \in \mathbb{R}^{nm}$ contains the irradiance data of a season
Output: $Q \in \mathbb{R}^{m\times k}$ is a $(m \times k)$ matrix with $k$ centers for
    each hour of a day
$m \leftarrow 24$
$k \leftarrow 5$
$\widehat{X} \leftarrow \{Reshape(X)\}'$    ▷ $\widehat{X} \in \mathbb{R}^{n\times m}$ is an $(n\times m)$ matrix
    as shown in Equation (2)
for $i \leftarrow 1, 2, \ldots, m$ do
    $Z \leftarrow \emptyset$    ▷ $Z \in \mathbb{R}^n$
    $Z \leftarrow \tilde{X} \subset \widehat{X}$    ▷ $\tilde{X} \in \mathbb{R}^n$ is the $i^{th}$ column of $\widehat{X}$
    $C \leftarrow Kmeans(Z)$    ▷ $C \in \mathbb{R}^k$
    $\overline{C} \leftarrow Bubblesort(C)$
    $(\tilde{Q} \leftarrow \overline{C}) \subset Q$    ▷ $\tilde{Q} \in \mathbb{R}^k$ is the $i^{th}$ row of $Q$
end for
return $Q$
```

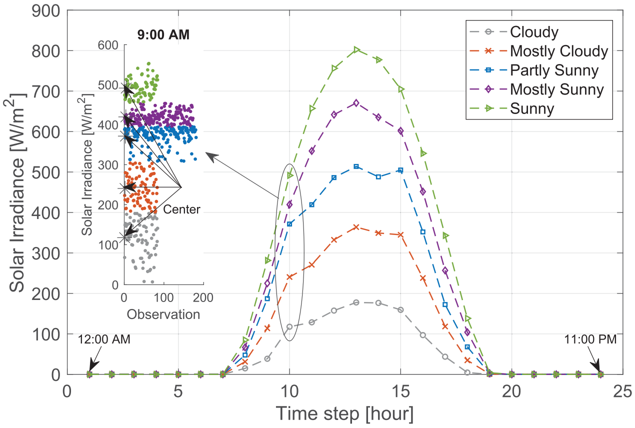

FIGURE 1. ***K*-means centers for the solar irradiance of each hour in a winter season - observed data are for the winters of 2012-2018 at the Desoto solar farm in Florida.**

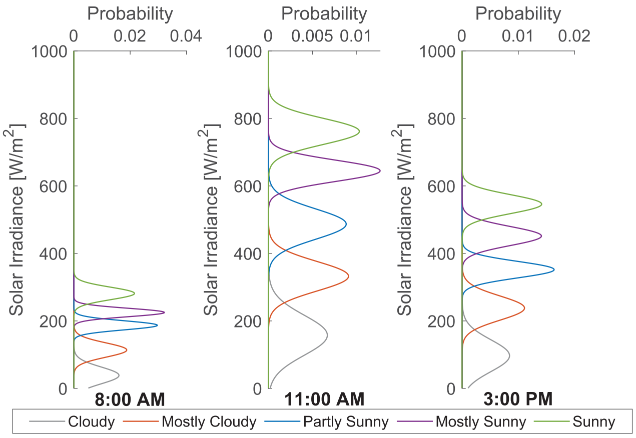

FIGURE 2. **Irradiance probability density function for each of the five clusters at three different hours of a winter day.**

Algorithm 2 Numerical Solar Irradiance Forecasting of an M-Hour Prediction Horizon

```
Input: $Y \in \mathbb{R}^M$ contains categorical solar irradiance forecast of the city (cloudy=1, mostly cloudy=2, partly sunny/partly cloudy=3, mostly sunny=4 and sunny=5), $I \in \mathbb{R}^M$ contains time index of each hour of the prediction horizon, $T \in \mathbb{R}^{24\times5}$ is a $(24 \times 5)$ matrix which contains the cluster center points for each hour of a day
Output: $Q \in \mathbb{R}^M$ contains numerical solar irradiance forecast for the solar farm location
$Q \leftarrow \emptyset$        $\triangleright\ Q \in \mathbb{R}^M$
for $i \leftarrow 1, 2, \ldots, M$ do
    $\bar{\tau} \leftarrow I[i]$
    $C \leftarrow \tilde{T} \subset T$    $\triangleright\ \tilde{T} \in \mathbb{R}^5$ is the $\bar{\tau}^{th}$ row of $T \in \mathbb{R}^{24\times5}$
    for $j \leftarrow 1, 2, \ldots, 5$ do
        if $Y[i] = j$ then
            $Q[i] = C[j]$
        end if
    end for
end for
return $Q$
```

D. SYNTHETIC WEATHER FORECAST DATA PREPARATION

The hourly and daily weather forecast data, for the considered city of Arcadia, Florida, is available on the weather forecast channel in [29]. The weather data contain the type of sky, temperature, relative humidity, wind speed and precipitable water. All the weather variables except solar irradiance are, to a great extent, homogeneous over all parts of the city. Hence, the temperature, relative humidity and wind speed forecasts of the city area are considered directly in the synthesized weather forecast of the solar farm area. The approximate numerical solar irradiance of the solar farm location is created for any hour using the saved irradiance clusters for that hour and the categorical type of sky forecast of the city. For each hour of the prediction horizon, the type of sky forecast of the city is associated with the saved cluster that corresponds to that type of sky at that hour of the day. Thereafter, the center for that cluster is considered as the approximate numerical forecast for that hour. The solar irradiance forecast technique is detailed in Algorithm 2. Fig. 3 shows the creation of a 24-hour irradiance forecast profile, where the city categorical type of sky forecast is translated into a numerical irradiance forecast for the PV plant location. The data flow of the proposed forecasting system is summarized in Fig. 4.

III. LSTM NN BASED FORECASTING FRAMEWORK

A. LSTM NEURAL NETWORK STRUCTURE

The LSTM NN was introduced by Hochreiter and Schmidhuber [30] in 1997. A classical LSTM NN is constructed by a sequence of an input layer, hidden layers and an output layer. The hidden layer contains a number of *memory cells* with *input* and *output gates*. The LSTM NN was improved later by Gers *et al.* [31] by introducing a new gate in the *memory cell* named the *forget gate*. The information flow through a *memory cell* is regulated by these gates. The core component of a hidden layer is called a *memory block* where a number of *memory cells* share the same gate units [30], [31]. The architecture of a *memory block* is shown in Fig. 5. An input sequence in the time frame is expressed as $\{x(1), x(2), \ldots, x(M)\} \in \mathbb{R}^{K\times M}$, where $x(\tau) \in \mathbb{R}^K$ is the feature vector at time step τ.

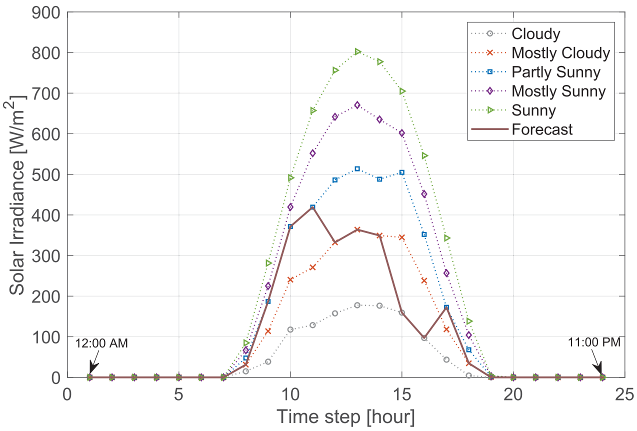

FIGURE 3. Created numerical solar irradiance forecast of Desoto solar farm location on Dec. 6, 2016.

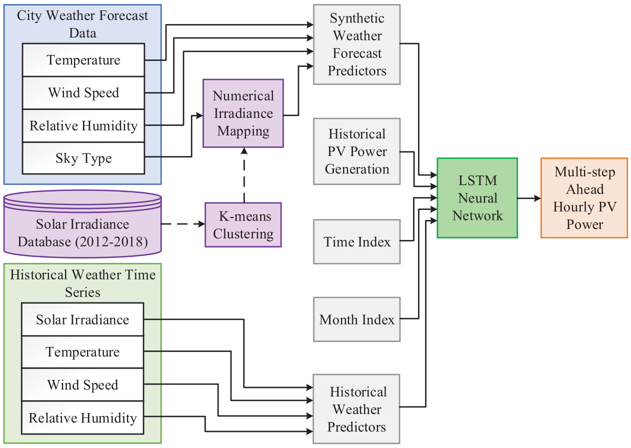
City Weather Forecast Data
Temperature
Wind Speed
Relative Humidity
Sky Type
Numerical Irradiance Mapping
Synthetic Weather Forecast Predictors
Historical PV Power Generation
Solar Irradiance Database (2012-2018)
K-means Clustering
Time Index
LSTM Neural Network
Multi-step Ahead Hourly PV Power
Historical Weather Time Series
Solar Irradiance
Temperature
Wind Speed
Relative Humidity
Month Index
Historical Weather Predictors

FIGURE 4. Simplified system architecture and data flow of the proposed model.

A *memory block*, with J number of *memory cells* and an input feature vector $x(\tau) \in \mathbb{R}^K$, is updated M times, one update for each feature vector in the input sequence. Each update of the *memory block* results in the current *state* vector $c(\tau) \in \mathbb{R}^J$. The cell *state* vector $c(\tau - 1) \in \mathbb{R}^J$ and the cell *output* vector $h(\tau - 1) \in \mathbb{R}^J$, from the previous time step, are utilized to calculate the current *output* vector $h(\tau) \in \mathbb{R}^J$. The *input* activation vector $i(\tau) \in \mathbb{R}^J$, the *forget* activation vector $f(\tau) \in \mathbb{R}^J$, and the *output* activation vector $o(\tau) \in \mathbb{R}^J$ are updated at each time step of the input sequence by utilizing a sigmoid activation function (σ), as shown in Equations (4), (5), (7). A hyperbolic tangent activation function (ϕ) is used to compute intermediate cell states, named as the

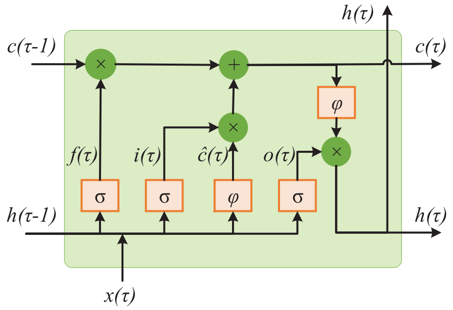

FIGURE 5. Architectural view of an LSTM NN memory block.

cell *candidates* $\widehat{c}(\tau) \in \mathbb{R}^J$, as shown in Equation (6). The corresponding $(J \times K)$ input weight matrices are concatenated as $\{W_i{}^T, W_f{}^T, W_c{}^T, W_o{}^T\}^T$ to give a $(4J \times K)$ matrix. Similarly, the corresponding recurrent $(J \times J)$ weight matrices are concatenated as $\{U_i{}^T, U_f{}^T, U_c{}^T, U_o{}^T\}^T$, which results in a $(4J \times J)$ matrix. The corresponding $(J \times 1)$ biases are concatenated as $\{b_i{}^T, b_f{}^T, b_c{}^T, b_o{}^T\}^T$, which is a $(4J \times 1)$ matrix. The update of a *memory block* at time step τ is formulated as follows:

$$f(\tau) = \sigma[W_f x(\tau) + U_f h(\tau-1) + b_f] \tag{4}$$

$$i(\tau) = \sigma[W_i x(\tau) + U_i h(\tau-1) + b_i] \tag{5}$$

$$\widehat{c}(\tau) = \phi[W_c x(\tau) + U_c h(\tau-1) + b_c] \tag{6}$$

$$o(\tau) = \sigma[W_o x(\tau) + U_o h(\tau-1) + b_o] \tag{7}$$

$$c(\tau) = f(\tau) \odot c(\tau-1) + i(\tau) \odot \widehat{c}(\tau) \tag{8}$$

$$h(\tau) = o(\tau) \odot \phi[c(\tau)] \tag{9}$$

The special sign "$\odot$" is introduced to show the element-wise multiplication.

B. THE PV POWER FORECASTING FRAMEWORK

The LSTM forecasting framework is shown in Fig. 6. The historical PV generation and weather data, the weather forecast variables, the time of the day index, and the month of the season index are considered as the input features of proposed model. The min-max normalization method is adopted to normalize the numerical predictors, whereas the categorical predictors are standardized using the one-hot encoder technique. The input time-sequence length is selected to have the same length (M) as that of the prediction horizon. The input matrix at a time step $(\tau - 1)$ is prepared and processed as follows:

1) The PV power generation sequence of past M hours is set as $P = \{p(\tau - M), p(\tau - M + 1), \ldots, p(\tau - 1)\} \in \mathbb{R}^M$. The historical weather data contains solar irradiance, temperature, wind speed and relative humidity of past M hours. The sequences of the historical weather variables, formulated as $\{w(\tau - M), w(\tau - M + 1), \ldots, w(\tau - 1)\} \in \mathbb{R}^M$, are concatenated to form a $(4 \times M)$ matrix $W \in \mathbb{R}^{4\times M}$. The synthetic weather forecast data comprises solar irradiance, temperature, wind speed and relative humidity of next M hours. The sequences of the weather forecast variables, formulated as $\{f(\tau), f(\tau + 1), \ldots, f(\tau + M - 1)\} \in \mathbb{R}^M$, are concatenated to form a $(4 \times M)$ matrix $F \in \mathbb{R}^{4\times M}$.
2) The sequence of the incremental time of the day indices for past M hours is $\{i(\tau - M), i(\tau - M + 1), \ldots, i(\tau - 1)\} \in \mathbb{R}^M$. The one-hot encoding transform this series into a $(24 \times M)$ matrix, $I \in \mathbb{R}^{24\times M}$, since each hour of the day is encoded using 24 categories. The sequence of month of the season indices of past M hours is $\{s(\tau - M), s(\tau - M + 1), \ldots, s(\tau - 1)\} \in \mathbb{R}^M$. After one-hot encoding, it forms a $(3 \times M)$ matrix $S \in \mathbb{R}^{3\times M}$ as the month of the season has only three categories, i.e. three months.
3) Finally, the predictors are concatenated to create the $(36 \times M)$ input matrix $X(\tau - 1) = \{P^T, W^T, F^T, I^T, S^T\}^T$.

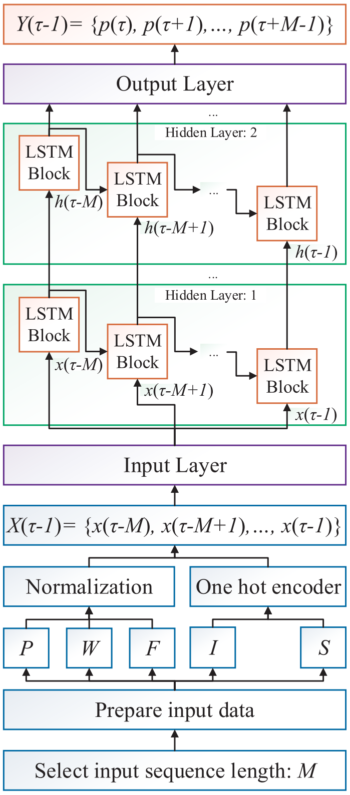

FIGURE 6. LSTM NN forecasting framework unfolded in time.

The input matrix is fed to an LSTM network with two hidden layers, with J_1 number of cells in the first hidden layer, and J_2 number of cells in the second one. The input matrix is fed sequentially, a single feature vector, $x \in \mathbb{R}^{36}$, at each time step. Hence, the *memory block* is updated M times, one for each time step of the input sequence. The *memory block* of the second hidden layer is updated synchronously at each time step by accepting the output vector, $h \in \mathbb{R}^{J_1}$, from the first layer. The cell output vector, $h \in \mathbb{R}^{J_2}$, from each update is sent to the output layer to calculate the output sequence, $Y(\tau - 1) = \{p(\tau), p(\tau + 1), \ldots, p(\tau + M - 1)\} \in \mathbb{R}^M$. Finally, the output sequence is denormalized to produce the predicted PV generation sequence.

IV. PERFORMANCE EVALUATION

The forecasting performance of the proposed model is evaluated using five statistical metrics, namely the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), the mean relative error (MRE), and the mean bias error (MBE) [32]–[34]. These metrics are defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_{predicted}[i] - x_{actual}[i]| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{predicted}[i] - x_{actual}[i])^2} \tag{11}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{x_{predicted}[i] - x_{actual}[i]}{x_{actual}[i]}| \times 100\% \tag{12}$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_{predicted}[i] - x_{actual}[i]|}{P_{PV}^{capacity}} \times 100\% \tag{13}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(x_{predicted}[i] - x_{actual}[i]) \tag{14}$$

where $P_{PV}^{capacity}$ is the capacity of the PV power plant.

V. RESULTS AND DISCUSSION

The simulation is carried out over 3 years (2016-2018) of data. The data set of each season is divided into two subsets using a 9:1 training to testing ratio. In the deep LSTM NN, the first hidden layer contains 75 *memory cells*, while the second layer contains 70 *memory cells*. The algorithm is implemented in MATLAB 2019b, and executed on Intel(R) Core (TM) i3 CPU @ 2.1 GHz and 8GB of memory.

A. FORECASTING PERFORMANCE OF THE PROPOSED ALGORITHM

The forecasting performance of the proposed algorithm is evaluated using different prediction horizon lengths in different seasons. The raw weather forecast categorical type of sky is also used directly in the proposed algorithm to verify the significance of the proposed synthetic weather forecast data. To achieve this, the proposed model is tested using two additional data sets, named as direct-1 and direct-2 versions. In the direct-1 version, the weather forecast data contains the hourly type of sky category of the city area, while the daily type of sky category is used in the direct-2 version.

The forecasting accuracy of the proposed model is compared in Table 2 and Table 3, which show that the algorithm using synthetic weather forecast data performs significantly better than the other two versions. The seasonal effect on the forecasting accuracy is shown in Table 2 for 24-hour prediction horizons. As can be seen, the proposed model can achieve up to 33% improvement in accuracy in comparison to the direct-1 version, and up to 44.6% in comparison to the direct-2 version, in the autumn season. The horizon length impact on the prediction accuracy is shown for the summer season in Table 3.

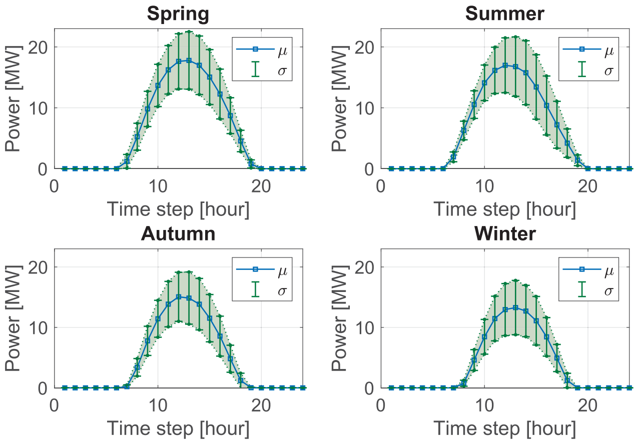

FIGURE 7. Volatility analysis of PV power generation in different seasons.

Season wise, the accuracy is lower in the spring and summer seasons, yet the proposed model still achieves higher accuracy than the other two versions. A volatility analysis of PV power generation in different seasons is performed using seven years (2012-2018) worth of data to investigate the lower prediction accuracy in the spring and the summer. As shown in Fig. 7, PV power generation is more volatile in the spring and the summer in comparison to the autumn and winter seasons. To determine probability density functions of the averaged forecasting error, the prediction horizon is rolled over one hour at a time for a testing period of one year (2018), and prediction horizons are picked randomly. Thereafter, the forecasting errors from each set of samples are averaged. The probability distributions of the averaged forecasting error are shown in Fig. 8 as Gaussian distributions, according to the central limit theorem [35].

Though the model gives the highest error for 24-hour prediction horizons, it shows more consistency as depicted in Fig. 8. Season wise, the model achieves the highest accuracy and consistency in the autumn. The predictions of two 12-hour rolling horizons are shown in Fig. 9. To assess the convergence of the NN training, the training process is repeated for 15 times with the same number of iterations, but with randomly chosen initial weights. The prediction average and standard deviation envelops are shown in Fig. 10, for 24-hour horizons. The proposed model performance under different day conditions is shown in Fig. 11.

B. FORECASTING ENGINE PERFORMANCE EVALUATION

The prediction capability of the LSTM NN, with the proposed features, is evaluated by implementing the proposed algorithm using a recurrent neural network (RNN), a generalized regression neural network (GRNN), and an extreme learning machine (ELM). The hourly data of the autumn season is used to compare the performance of the forecasting

TABLE 2. Prediction accuracy for different seasons.

Season	Synthetic				Direct-1				Direct-2			
	MAE (MW)	RMSE (MW)	MAPE (%)	MRE (%)	MAE (MW)	RMSE (MW)	MAPE (%)	MRE (%)	MAE (MW)	RMSE (MW)	MAPE (%)	MRE (%)
Spring	0.93	1.73	48.18	3.72	1.22	2.27	69.55	4.88	1.49	2.76	79.26	5.96
Summer	0.82	1.68	37.81	3.28	0.99	2.03	41.81	3.96	1.21	2.41	66.61	4.84
Autumn	0.36	0.71	22.31	1.44	0.54	1.04	28.11	2.16	0.65	1.23	35.71	2.6
Winter	0.41	0.83	36.84	1.64	0.55	1.07	55.91	2.2	0.72	1.41	69.65	2.88

TABLE 3. Prediction accuracy for different horizon lengths - (summer).

Hour Ahead	Synthetic				Direct-1				Direct-2			
	MAE (MW)	RMSE (MW)	MAPE (%)	MRE (%)	MAE (MW)	RMSE (MW)	MAPE (%)	MRE (%)	MAE (MW)	RMSE (MW)	MAPE (%)	MRE (%)
6	0.69	0.94	28.59	2.76	0.76	1.02	32.97	3.04	1.03	1.35	54.99	4.12
12	0.75	1.28	38.53	3	0.82	1.37	41.58	3.28	1.06	1.74	63.85	4.24
24	0.82	1.68	37.81	3.28	0.99	2.03	41.81	3.96	1.21	2.41	66.61	4.84

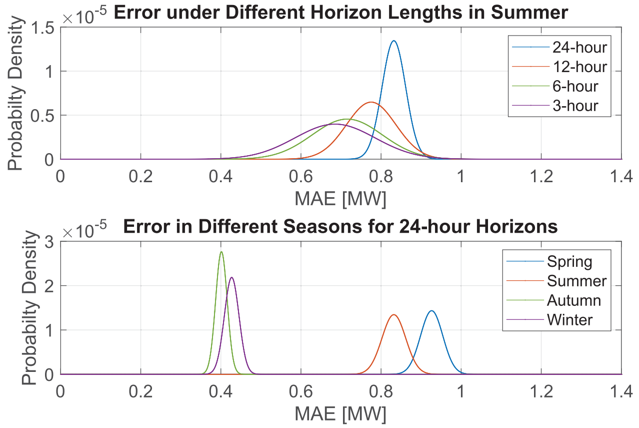

FIGURE 8. Averaged error probability distribution for different prediction horizons of a day, and different seasons of a year.

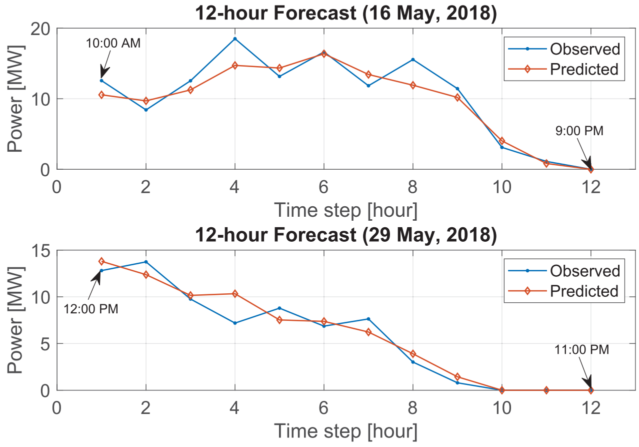

FIGURE 9. 12-hour ahead rolling horizons, with different horizon starting hours, in spring 2018.

engines. The prediction accuracy for 24-hour horizons is compared in Table 4. Performance samples from the autumn of 2018 data are shown in Fig. 12. The comparison reveals the superiority of the LSTM NN with the proposed algorithm, over the RNN, the GRNN, and the ELM. It is worth noticing that the RNN can still achieve higher accuracy than those of the GRNN and the ELM machines due to its recurrent nature. However, the LSTM NN shows higher accuracy due

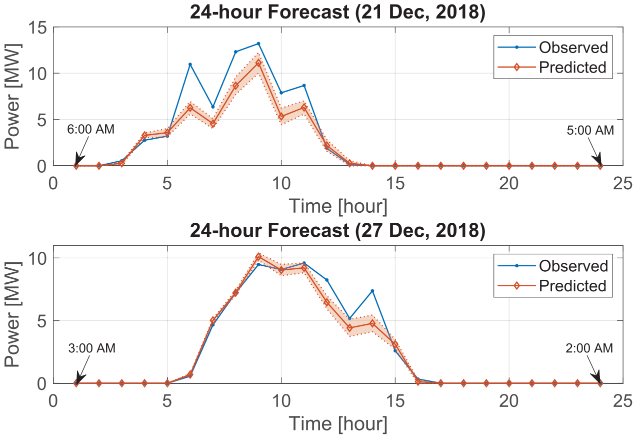

FIGURE 10. 24-hour rolling horizons, with different horizon starting hours, in winter 2018. The prediction average and standard deviation envelopes are shown as a results of 15 training trials, with same number of iteration, but with different initial weights.

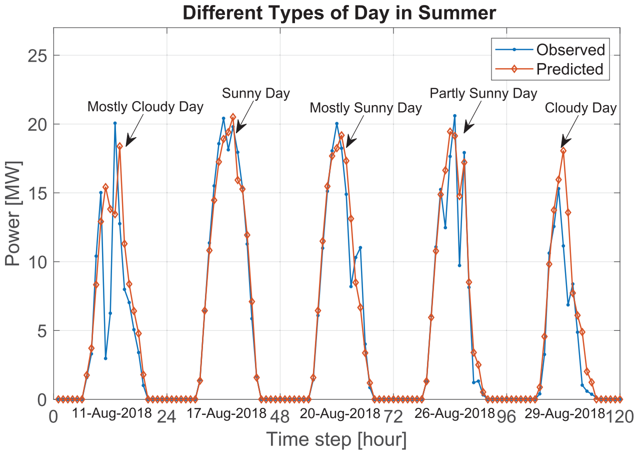

FIGURE 11. 24-hour ahead PV power forecasting for different types of days in summer, 2018.

TABLE 4. Forecasting accuracy comparison among different forecasting engines- autumn.

Engine	MAE (MW)	RMSE (MW)	MAPE (%)	MRE (%)	MBE (MW)
LSTM NN	0.36	0.71	22.31	1.44	0.01
RNN	0.47	0.92	28.79	1.88	-0.03
GRNN	0.67	1.31	34.06	2.68	0.27
ELM	1.38	2.41	284.85	5.52	0.51

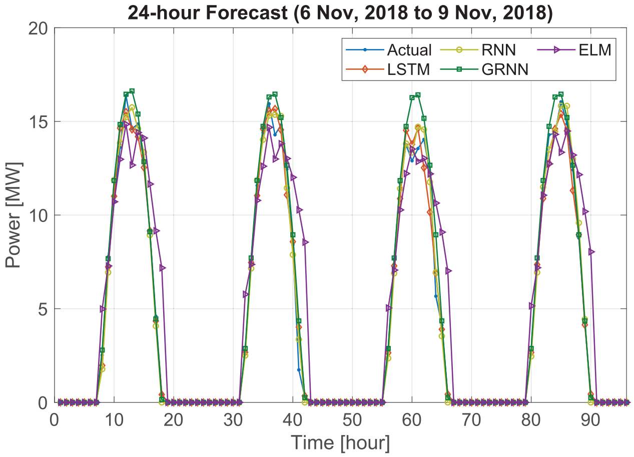

FIGURE 12. Sample performance for 24-hour prediction horizons in autumn, 2018.

to the additional input, forget, and output gates in the memory cells, which equip the LSTM NN with the ability to preserve long-range temporal dependencies more than the RNN.

VI. CONCLUSION

In this paper, an algorithm is proposed to exploit the time series processing qualities of LSTM NNs along with the proposed synthetic irradiance forecast to predict PV power generation. Instead of employing the categorical hourly or daily type of sky forecast, of the city area, the types of sky are defined for each hour uniquely, using different levels of irradiance, based on the hour of the day and the season. It is shown that using this synthetic irradiance forecast can achieve up to 33% improvement in accuracy in comparison to that when the hourly type of sky forecast is used, and up to 44.6% in comparison to that when the daily type of sky forecast is used. Moreover, the superiority of the LSTM NN with the proposed input features, is verified in this paper by implementing the proposed algorithm using other forecasting engines, namely the ELM, the GRNN, and the RNN.

REFERENCES

[1] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and Metaheuristic techniques," *IET Renew. Power Gener.*, vol. 13, no. 7, pp. 1009–1023, May 2019.

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

[3] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, and D. C. Petković, and Sudheer, "A support vector machine–firefly algorithm-based model for global solar radiation prediction," *Sol. Energy*, vol. 115, pp. 632–644, May 2015.

[4] J. Wang, H. Zhong, X. Lai, Q. Xia, Y. Wang, and C. Kang, "Exploring key weather factors from analytical modeling toward improved solar power forecasting," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1417–1427, Mar. 2019.

[5] 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.

[6] W. VanDeventer, E. Jamei, G. S. Thirunavukkarasu, M. Seyedmahmoudian, T. K. Soon, B. Horan, S. Mekhilef, and A. Stojcevski, "Short-term PV power forecasting using hybrid GASVM technique," *Renew. Energy*, vol. 140, pp. 367–379, Sep. 2019.

[7] A. Asrari, T. X. Wu, and B. Ramos, "A hybrid algorithm for short-term solar power prediction—Sunshine state case study," *IEEE Trans. Sustain. Energy*, vol. 8, no. 2, pp. 582–591, Apr. 2017.

[8] C. Wan, J. Lin, Y. Song, Z. Xu, and G. Yang, "Probabilistic forecasting of photovoltaic generation: An efficient statistical approach," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2471–2472, May 2017.

[9] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, and Y. Du, "Short-term photovoltaic power forecasting based on long short-term memory neural network and attention mechanism," *IEEE Access*, vol. 7, pp. 78063–78074, 2019.

[10] H. Zang, L. Cheng, T. Ding, K. W. Cheung, Z. Liang, Z. Wei, and G. Sun, "Hybrid method for short-term photovoltaic power forecasting based on deep convolutional neural network," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 20, pp. 4557–4567, Nov. 2018.

[11] L. Du, L. Zhang, X. Tian, and Q. Cheng, "Short-term photovoltaic power forecasting using deep convolutional networks," in *Proc. 2nd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2019, pp. 149–153.

[12] W. Lee, K. Kim, J. Park, J. Kim, and Y. Kim, "Forecasting solar power using long-short term memory and convolutional neural networks," *IEEE Access*, vol. 6, pp. 73068–73080, 2018.

[13] K. Maciejowska, W. Nitka, and T. Weron, "Day-ahead vs. Intraday—Forecasting the price spread to maximize economic benefits," *Energies*, vol. 12, no. 4, p. 631, 2019.

[14] J. Ospina, A. Newaz, and M. O. Faruque, "Forecasting of PV plant output using hybrid wavelet-based LSTM-DNN structure model," *IET Renew. Power Gener.*, vol. 13, no. 7, pp. 1087–1095, May 2019.

[15] A. F. Atiya, S. M. El-Shoura, S. I. Shaheen, and M. S. El-Sherif, "A comparison between neural-network forecasting techniques-case study: River flow forecasting," *IEEE Trans. Neural Netw.*, vol. 10, no. 2, pp. 402–409, Mar. 1999.

[16] M. Q. Raza, N. Mithulananthan, J. Li, K. Y. Lee, and H. B. Gooi, "An ensemble framework for day-ahead forecast of PV output power in smart grids," *IEEE Trans. Ind. Informat.*, vol. 15, no. 8, pp. 4624–4634, Aug. 2019.

[17] M. Seyedmahmoudian, E. Jamei, G. Thirunavukkarasu, T. Soon, M. Mortimer, B. Horan, A. Stojcevski, and S. Mekhilef, "Short-term forecasting of the output power of a building-integrated photovoltaic system using a Metaheuristic approach," *Energies*, vol. 11, no. 5, p. 1260, May 2018.

[18] 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.

[19] M. S. Hossain and H. Mahmood, "Short-term photovoltaic power forecasting using an LSTM neural network," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2020, pp. 1–5.

[20] C.-J. Huang and P.-H. Kuo, "Multiple-input deep convolutional neural network model for short-term photovoltaic power forecasting," *IEEE Access*, vol. 7, pp. 74822–74834, 2019.

[21] J. Liu, W. Fang, X. Zhang, and C. Yang, "An improved photovoltaic power forecasting model with the assistance of aerosol index data," *IEEE Trans. Sustain. Energy*, vol. 6, no. 2, pp. 434–442, Apr. 2015.

[22] H.-T. Yang, C.-M. Huang, Y.-C. Huang, and Y.-S. Pai, "A weather-based hybrid method for 1-Day ahead hourly forecasting of PV power output," *IEEE Trans. Sustain. Energy*, vol. 5, no. 3, pp. 917–926, Jul. 2014.

[23] C.-M. Huang, C.-J. Kuo, S.-J. Chen, and S.-P. Yang, "One-day-ahead hourly forecasting for photovoltaic power generation using an intelligent method with weather-based forecasting models," *IET Gener., Transmiss. Distrib.*, vol. 9, no. 14, pp. 1874–1882, Nov. 2015.

[24] 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.

[25] W. Wu and M. Peng, "A data mining approach combining K-means clustering with bagging neural network for short-term wind power forecasting," *IEEE Internet Things J.*, vol. 4, no. 4, pp. 979–986, Aug. 2017.

[26] *National Renewable Energy Laboratory (NREL)*. Accessed: Jan. 6, 2020. [Online]. Available: https://www.nrel.gov/research/data-tools.html

[27] *National Oceanic and Atmospheric Administration (NOAA)*. Accessed: Jul. 12, 2019. [Online]. Available: https://www.weather.gov/bgm/forecast terms

[28] C. Keerthisinghe, G. Verbic, and A. C. Chapman, "A fast technique for smart home management: ADP with temporal difference learning," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3291–3303, Jul. 2018.

[29] *The Weather Channel*. Accessed: Jul. 16, 2019. [Online]. Available: https://www.weather.com

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Networks: ICANN*, vol. 2. Edinburgh, U.K., 1999, pp. 850–855.

[32] R. Jiao, T. Zhang, Y. Jiang, and H. He, "Short-term non-residential load forecasting based on multiple sequences LSTM recurrent neural network," *IEEE Access*, vol. 6, pp. 59438–59448, 2018.

[33] G. Chicco, V. Cocina, P. Di Leo, F. Spertino, and A. Massi Pavan, "Error assessment of solar irradiance forecasts and AC power from energy conversion model in grid-connected photovoltaic systems," *Energies*, vol. 9, no. 1, p. 8, Dec. 2015.

[34] M. A. E. Bhuiyan, F. Yang, N. K. Biswas, S. H. Rahat, and T. J. Neelam, "Machine learning-based error modeling to improve GPM IMERG precipitation product over the brahmaputra river basin," *Forecasting*, vol. 2, no. 3, pp. 248–266, Jul. 2020.

[35] M. E. Hajiabadi and H. R. Mashhadi, "Analysis of the probability distribution of LMP by central limit theorem," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2862–2871, Aug. 2013.

MOHAMMAD SAFAYET HOSSAIN (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from the Khulna University of Engineering and Technology, Khulna, Bangladesh, in 2012. He is currently pursuing the M.Sc.Eng. degree in electronic, communication and control systems with Florida Polytechnic University, Lakeland, FL, USA. He worked as an Associate Maintenance Professional at Halliburton, from 2013 to 2016. From 2017 to 2018, he worked as a Maintenance Engineer at Walton Hi-Tech Industries Ltd. His research interests include intelligent energy management systems, microgrids, power electronics, and renewable energy integration.

HISHAM MAHMOOD (Member, IEEE) received the M.E.Sc. degree in control engineering from Lakehead University, Thunder Bay, ON, Canada, in 2008, and the Ph.D. degree in electrical engineering from the University of Western Ontario, London, ON, Canada, in 2014. He was a Postdoctoral Research Fellow with the Distributed Generation Laboratory, University of Western Ontario, from 2015 to 2017. From 2017 to 2018, he was a Research Fellow with the Department of Renewable Energy, University of Exeter, Exeter, U.K., where he was a part of the EU Interreg-funded Intelligent Community Energy (ICE) project to develop smart energy solutions for remote communities along the English Channel. He was also a part of a project called Cornwall New Energy which provides support for small and medium-sized enterprises, in Cornwall and the Isles of Scilly, U.K., to enable new energy products and services to be brought to market. He is currently an Assistant Professor with Florida Polytechnic University, Lakeland, FL, USA. His research interests include intelligent energy management systems, renewable energy integration, distributed generation, microgrids, and modeling and control of switching power converters.