IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 19, 2020, accepted July 21, 2020, date of publication July 27, 2020, date of current version August 5, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3011982*

# A New Data Driven Long-Term Solar Yield Analysis Model of Photovoltaic Power Plants

**BIPLOB RAY[1], (Member, IEEE), RAKIBUZZAMAN SHAH[2], (Member, IEEE), MD. RABIUL ISLAM[3], (Senior Member, IEEE), AND SYED ISLAM[2], (Fellow, IEEE)**

[1]Centre of Intelligent System, School of Engineering and Technology, CQUniversity, Rockhampton, QLD 4701, Australia
[2]School of Engineering, Information Technology, and Physical Sciences, Federation University Australia, Ballarat, VIC 3353, Australia
[3]School of Electrical, Computer, and Telecommunication Engineering, University of Wollongong, Wollongong, NSW 2522, Australia

Corresponding author: Rakibuzzaman Shah (m.shah@federation.edu.au)

**ABSTRACT** Historical data offers a wealth of knowledge to the users. However, often restrictively mammoth that the information cannot be fully extracted, synthesized, and analyzed efficiently for an application such as the forecasting of variable generator outputs. Moreover, the accuracy of the prediction method is vital. Therefore, a trade-off between accuracy and efficacy is required for the data-driven energy forecasting method. It has been identified that the hybrid approach may outperform the individual technique in minimizing the error while challenging to synthesize. A hybrid deep learning-based method is proposed for the output prediction of the solar photovoltaic systems (i.e. proposed PV system) in Australia to obtain the trade-off between accuracy and efficacy. The historical dataset from 1990-2013 in Australian locations (e.g. North Queensland) are used to train the model. The model is developed using the combination of multivariate long and short-term memory (LSTM) and convolutional neural network (CNN). The proposed hybrid deep learning (LSTM-CNN) is compared with the existing neural network ensemble (NNE), random forest, statistical analysis, and artificial neural network (ANN) based techniques to assess the performance. The proposed model could be useful for generation planning and reserve estimation in power systems with high penetration of solar photovoltaics (PVs) or other renewable energy sources (RESs).

**INDEX TERMS** Accuracy, convolutional neural network, data-driven model, deep learning, forecasting, multivariate long and short-term memory, reliability, solar photovoltaic power plants.

## I. ABBREVIATION

| | |
|---|---|
| **ANN** | Average Neural Network |
| **APL** | Artificial Pooling Layers |
| **ARENA** | Australian Renewable Energy Agency |
| **ASEFS** | Australian Solar Energy Forecasting System |
| **B** | Bias |
| **BNL** | Batch Normalization Layer |
| **CNN** | Convolutional Neural Network |
| **DL** | Dropout Layer |
| **DNI** | Direct Normal Irradiance |
| **DHI** | Diffuse Horizontal Irradiance |
| **GHI** | Global Horizontal Irradiance |
| **IW** | Input Weights |
| **LSTM** | Long and Short-Term Memory |
| **NN** | Neural Network |
| **NNE** | Neural Network Ensemble |
| **PSO** | Particle Swarm Optimization |
| **RNN** | Recurrent Neural Network |
| **ReLU** | Rectified Linear Unit |
| **RW** | Recurrent Weights |
| **W** | Weights |

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

## II. INTRODUCTION

The penetrations of solar photovoltaic (PV) are increasing in several countries including Australia in multiple straight years. Significant PVs are either connected to medium or low voltage networks in Australia. The growth of both large and small-scale PV penetrations has economic and environmental benefits. However, it poses a range of management and control issues for grid operators due to the variability of PV outputs. The power system has become increasingly volatile and less predictable with PV systems [1]. The PV systems are weather dependent, therefore, hard to predict. Accuracy of the prediction is critically important for secure operation of power systems with high penetrations of

PV systems. It enables the system operator to deal with output power variability and planning engineers to plan and design the power system for future [2]. There are various methods for such forecasting in different time horizons, e.g. short, medium, and long-term. The physical, persistence, statistical, and combined approaches may be used to estimate the output of variable generations [3]. The meteorological data and energy forecasting are the two significant components related to the forecasting of the PV system [4]. Many procedures were proposed in the literature to forecast meteorological information such as wind speed, cloud cover, temperature, and irradiance [5], [6]. Furthermore, physical, meteorological data-driven, and astronomical driven are the common methods reported in the literature to forecast the output power and energy of the PV system. Different parameters such as power rating, azimuth angle, module type, tilt angle, wind speed are used in the physical model for energy forecasting in PV systems [7]. The historical weather data and the previous measurements of PV system outputs are used in the meteorological data-driven method for PV forecasting [8]. The statistical, persistence, auto-regression are the key methods used for this purpose [8]. Recently, machine learning techniques have widely been applied in the meteorological data-driven approach to forecast PV output [9]. In the astronomical and meteorological data-driven approach, the physical factor has been used with the meteorological data [9]. In a data-driven traditional statistical method, the measured historical PV data in the past time is used in forecast [10]. The auto-regression and spatial-temporal are the other two widely used data-driven methods for such application [11], [12]. However, the physical information of PV is often limited or ignored in these methods [10]–[12]. Although different techniques have already recognized for forecasting PV output, there is still an opportunity to improve the reliability and accuracy regarding the long-term forecasting of the PV system to be used in power system planning. A good number of works have been attempted to estimate the short-term solar yield using historical data. Most of the forecasting techniques applied in minutes into day spatial resolution for dispatching and load following, unit commitment, distributed generation operation, building energy management, and transmission scheduling. However, very few studies have investigated the data-driven long-term estimation of solar yield. In this paper, a data-driven model is proposed for reliable estimation of solar yield from historical data.

Three main forecasting algorithms categories, i.e. statistical analysis [13], machine learning [14], and hybrid [15], [16], were reported. The Australian Renewable Energy Agency (ARENA) has reported the Australian Solar Energy Forecasting System (ASEFS), which used the statistical models like decision tree, random forest, and persistence to forecast the hour ahead prediction of solar energy in Australia. The model has a root mean squire error (RMSE) of 15.80. Hence, there is still a prospect to improve in the forecasting approaches. Furthermore, several machine learning techniques were attempted to forecast minutes, hours, and day-ahead energy outputs of large-scale PV systems [7], [8], [14], and [17]. These were mainly used various neural networks (NN) based forecasting techniques with the short length of dataset. Very few studies have exhibited good forecasting performance as reported in [7], which has a normalized root-mean-square deviation or error (nRMSE) of 0.07356. However, the proposed algorithm in [7] are not suitable in generalized forecasting due to the underlying weather classification and certain assumptions applicable to the specific region. Furthermore, the hybrid techniques were attempted to combine the algorithms for better performance as stated in [7]. The proposed method combines the particle swarm optimization (PSO) with the variation of NN to achieve better forecasting performance. However, the performance of the proposed algorithm is almost similar to other NN based algorithms for forecasting. Recently, the recurrent neural network (RNN) and deep learning [16] based forecasting have received a great deal of attention due to better prediction performance compared to traditional techniques i.e. statistical, PSO, NN. But, most of the deep learning-based methods are used for short-term forecasting with the small length of data.

In this paper, a novel hybrid deep learning method is proposed. A number of studies have individually used long and short-term memory (LSTM) and convolutional neural network (CNN) individually in various application including forecasting of PV output [18]. This paper proposed a method that combines LSTM and CNN to obtain a hybrid algorithm for long-term forecasting of PV output. The proposed algorithm is compared with four baseline modelling methods and demonstrates the better performance compared to the other methods. The rest of the paper is organized as follows: Section III briefly describes the key techniques considered in this paper. The methodology is explained in Section IV. Results and discussions are presented in Section V. The conclusions and the contributions of the paper are given in Section VI.

## III. OVERVIEW

The long-term forecasting of solar PV can be used for the planning of power system reserve with high penetration of PV systems. The goal of this research is to find the PV power and energy in the long-term time horizon – a couple of years ahead. The historic Typical Meteorological Year (TMY) dataset used for this study. The TMY dataset are obtained from Energy Partners' [19]. The TMY data is used in the System Advisor Model (SAM) to prepare the required weather data and PV output data for the prediction model. The blending of two deep learning methods has been considered. Fig. 1. shows an overview of the proposed method. The key techniques used in this work are briefly described next in this section.

### A. RECURRENT NEURAL NETWORK (RNN)

The recurrent neural networks (RNNs) consist of recurrent loops of networks that allow persistent information flow [16].

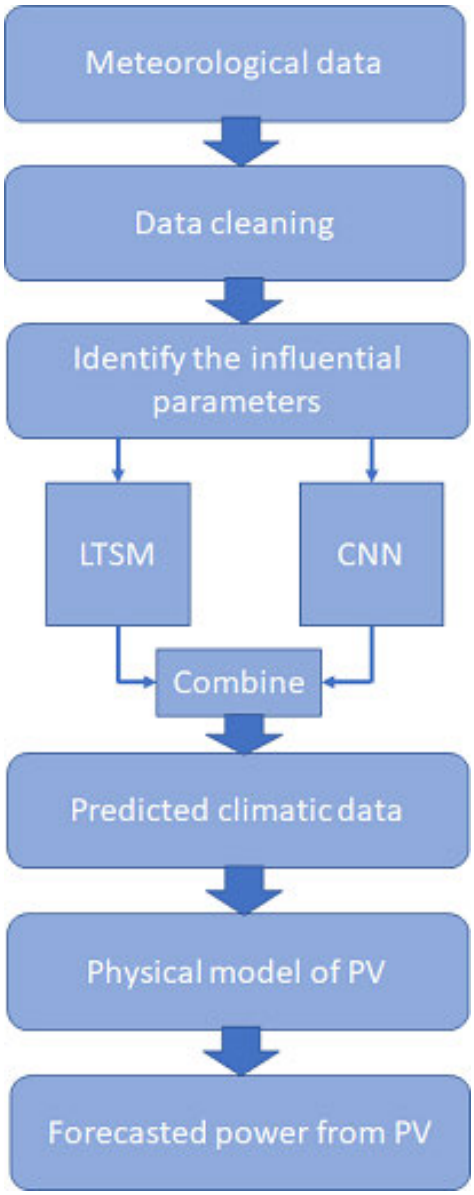


**FIGURE 1.** Overview of the proposed method.

These loops allow the information to flow concurrently from one step of the network to the next using the chain of events within networks which are intimately related to the sequences and lists. The concept of the recurrent neural network is the base of deep learning techniques/algorithms which are inspired by the connection of neurons in human brain [16], [17]. It uses recurrent learning to learn from large and complex dataset. Deep learning is used to solve complex problems that require input from diverse, unstructured, and inter-connected dataset. In this work, two of the most popular deep learning techniques such as long and short-term memory (LSTM) and convolutional neural network (CNN) are utilized. The details about LTSM and CNN are given later in this paper.

### B. LONG- AND SHORT-TERM MEMORY (LSTM)

The LSTM is a deep learning technique explicitly designed to reduce long lasting dependency problem using a chain like structure [20]. The recuring model of LSTM uses concurrent cell update structure. The initial update starts right after the first output of initial LSTM block which uses the initial state of the network and the first-time step of the sequence to compute the output. At time step $t$, the block uses the current state $(c_{t-1}, y_{t-1})$ to update cell state $c_t$, and the following time step of the network to compute the output. Each layer has two states known as the cell and the hidden state (also known as the output state). The output of the LSTM layer at time step $t$ is contained in the hidden state of the same time step [21]. The information erudite from previous steps is confined in the cell state of the current step. The layer adds or removes information from the cell state controlled by gates in each time step. Fig. 2 illustrates a general LSTM block architecture.

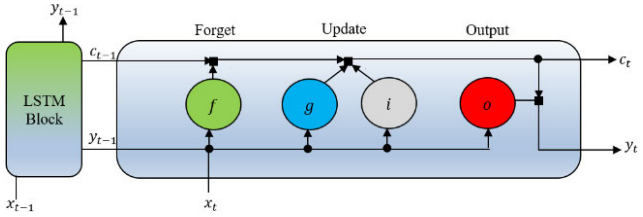


**FIGURE 2.** LSTM general architecture.

From Fig. 2, it is evident that there are four control gates in LSTM: forget $(f)$, cell candidate $(g)$, input $(i)$, and output $(o)$ as illustrated in Fig. 2. When $c\,(t-1)$ points enter to the LSTM unit from LSTM block, it first passed through the forget gate and drop some memory. The new memories are added by update gate. The output is filtered through the output gate. Working mechanisms can be mathematically expressed as in (1) - (4) for timestamp $t$ for each control gate.

$$i_t = \sigma_g\left(W_i x_t + R_i y_{t-1} + b_i\right) \tag{1}$$
$$f_t = \sigma_g\left(W_f x_t + R_f y_{t-1} + b_f\right) \tag{2}$$
$$g_t = \sigma_g\left(W_g x_t + R_g y_{t-1} + b_g\right) \tag{3}$$
$$O_t = \sigma_g\left(W_o x_t + R_t y_{t-1} + b_o\right) \tag{4}$$

In (1)-(4), $\sigma_g$ denotes the gate activation function. The sigmoid function given by $\sigma(x) = (1 + e^{-x})^{-1}$ is used to compute the gate activation function in MATLAB [22]. There are three learnable weights of an LSTM layer: input weights $W$, recurrent weights $R$, and bias $b$. The matrices of $W$, $R$, and $b$ are concatenated as in (5).

$$W = \begin{bmatrix} W_i \\ W_f \\ W_g \\ W_o \end{bmatrix}, \quad R = \begin{bmatrix} R_i \\ R_f \\ R_g \\ R_o \end{bmatrix}, \quad b = \begin{bmatrix} b_i \\ b_f \\ b_g \\ b_o \end{bmatrix} \tag{5}$$

where $i, f, g,$ and $o$ represent the input gate, forget gate, cell candidate, and output gate, respectively.

The cell and hidden state at timestamp $t$ are expressed by (6) and (7), respectively.

$$\boldsymbol{c_t} = f_t \odot \boldsymbol{c_{t-1}} + i_t \odot g_t \tag{6}$$

$$h_t = o_t \odot \sigma_c(c_t) \tag{7}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication of vectors) and $\sigma_c$ denotes the state activation function. The state activation function is compared by using the hyperbolic tangent function (*tanh*) and *lstmLayer* function in MATLAB.

### C. CONVOLUTIONAL NEURAL NETWORK (CNN)

The Convolutional Neural Network (CNN) is one of the most popular deep learning algorithms [21]. It has the advantage of extracting data features effectively. Therefore, the CNN is used widely in image recognition and classification. The CNN networks are like a visual cortex, with arrangements of simple and complex cells [18]. Similar to an RNN neural network, CNN is composed of three main components: the input layer, output layer, and hidden layers in between the input and output layers [23]. A general CNN structure is illustrated in Fig. 3. One or multiple convolutional layers may be involved in CNN as given in Fig. 3. The CNN used in this paper has four 2-D convolutional layers, BNL, ReLU layer, and APL. These are followed by one DL, fully connected layer, and regression output layer, respectively.

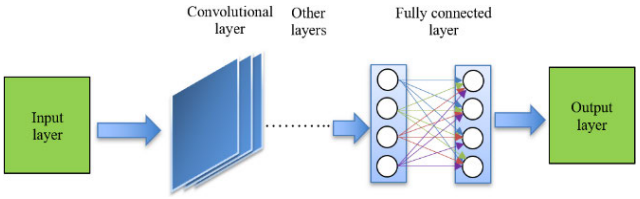


**FIGURE 3.** Generic architecture of CNN.

The influential input parameters of $m \times m \times n$ are used in CNN (where $m \times m$ determine size of each set, and $n$ specifies the total number of dataset). The inputs are passed to the convolutional 2D network consists of neurons that connect to sub-regions of input dataset or the output of the previous layer. The convolutional 2D network uses the set of weights called filter ($k$) that convolved the input. This has extracted the important features of the input dataset for accurate output prediction. Then, the batch normalization is used to normalize inputs ($m_i$) by calculating the mean ($\mu_B$) , and variance ($\sigma_B^2$) over a mini-batch and each input channel. The normalized activations can be obtained as in (8).

$$\hat{x}_i = \frac{m_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{8}$$

In (8), $\epsilon$ is the property Epsilon that improves the numerical stability when the mini-batch variance is very small. The batch normalization layers are followed by ReLU layer which acts as a threshold operation to the input with the following relationship as given in (9).

$$f(x) = \begin{cases} m, & m \geq 0 \\ 0, & m < 0 \end{cases} \tag{9}$$

The ReLU layer is followed by an APL, which performed down sampling. The input is divided into rectangular pooling regions to compute the average values in that region. If the input ($I$) to the pooling layer is $n \times n$, and the pooling region size ($PS$) is $h \times h$, then, the pooling layer down-sampled the regions by $h$ [23]. The output ($O$) of a pooling layer for overlapping regions can be expressed as in (10).

$$O = (I - PS + 2 * Padding)/(Stride + 1) \tag{10}$$

In the final stage, one DL, fully connected layer, and regression layer work together to prepare the output of the CNN network. The dropout layer randomly sets the input elements to zero given by the dropout mask $rand(size(m)) < Probability$ (where $m$ is the layer input). The fully connected layer multiplies the input by a weight matrix $W$ and adds the bias vector $b$. In this case, the fully connected layer acts independently on each time step with the sequential inputs. At time step $t$, the corresponding entry of $Z$ is $WY_t + b$. The loss function of the regression layer is the half-mean-squared-error for the sequence-to-one regression networks of the predicted responses as in (11). This can be computed by a regression layer as given in (11).

$$Loss = \frac{1}{2} \sum_{i=1}^{n} (t_i - y_i)^2 \tag{11}$$

where $n$ is the number of responses, $t_i$ is the target output, and $Y_i$ is the network's prediction for response $i$.

## IV. METHODOLOGY

The step-by-step methodology used in this paper is given in Fig. 4. The monthly dataset from 1990 to 2013 with one-hour time interval have been used here for the forecasting. Solar dataset for four locations in Queensland, e.g. Cairns, Gladstone, Rockhampton, and Townsville are considered to validate the proposed method.

***Step 1: Prepare the initial dataset***-The historic Typical Meteorological Year (TMY) dataset from 1990-2013 with *.tm2* file extension are used to generate the weather data for the proposed algorithm. The System Advisor Model (SAM) is used to generate the energy output of the PV system [24]. The SAM is developed by the National Renewable Energy Laboratory (NREL) to estimate the energy output of renewable energy systems including PV generators by the physical model of the system. The PV system in SAM has been tuned using the manufacturer data of PV cell, inverters, AC lines, derating factors, and others. Using the specification of the physical model of PV plant and relevant TMY dataset, the SAM presents influential weather parameters like global horizontal irradiance (GHI), direct normal irradiance (DNI), diffuse horizontal irradiance (DHI), wet bulb, and dew point temperature in hourly and monthly duration. Fig. 5 illustrates

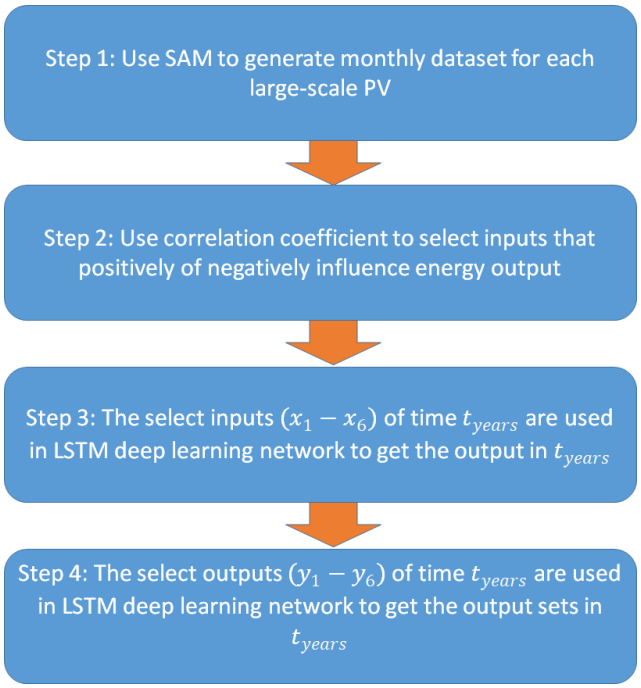


**FIGURE 4.** Process flow.

the output of a PV plant estimated by SAM for a representative year in Cairns. The obtained output can be exported to a.csv format to use this as an input to the deep learning algorithm. The hourly and monthly energy outputs of the PV plants are also calculated using TMY and physical model specification in the SAM.

***Step 2: Input selection***-Initially the generated weather data were analyzed using the correlation coefficient to find positive and negative Correlation Index (CI) for parameters associated with energy production. The CI values in this paper are calculated based on the Pearson product-moment correlation coefficient as given in (12) [25]:

$$CI = \frac{\sum (x - \bar{X})(y - \bar{Y})}{\sqrt{\sum (x - \bar{X})(y - \bar{Y})}} \tag{12}$$

The influential parameters are given in Table 1. As presented in Table 1, the five major influential input values with $CI > \pm 0.5$ are employed. As can be seen from Table 1, GHI and DNI are positively correlated while DHI, wet bulb, and dew point temperature are negatively correlated with energy the production.

**TABLE 1.** Influential weather parameters.

| Influential parameters (Input) | Correlation | Influential parameters (output) |
|---|---|---|
| GHI (W/m$^2$) | Positively correlated | PV output |
| DNI (W/m$^2$) | | |
| DHI (W/m$^2$) | Negatively correlated | |
| Wet bulb temp ($^o$C) | | |
| Dew point temp ($^o$C) | | |

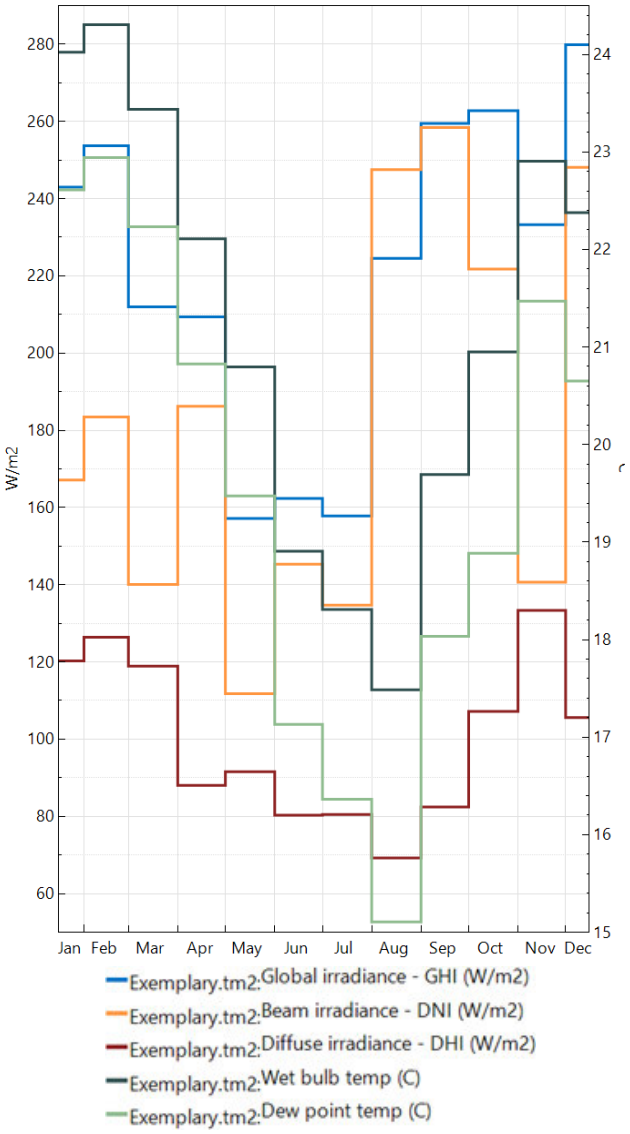


**FIGURE 5.** One year meteorological output.

***Step 3:*** In this step, the dataset are prepared for the training and testing of the hybrid deep learning structure. The LSTM part of the hybrid deep learning technique has been used to predict inputs in $t_{years}$ (this will be used in Step 4 to calculate PV output for $t_{nyears}$ using CNN part of the hybrid structure). The brief overview of dataset preparation and hybrid deep learning is given next.

### A. PREPARATION OF DATASET

From the available solar data, 1990–2014 (25 years), solar data from 1990 – 2013 are used for training and testing. Solar data of 2013 and 2014 are used for the validation. Fig. 7 shows the process flows which used to prepare dataset for training and testing.

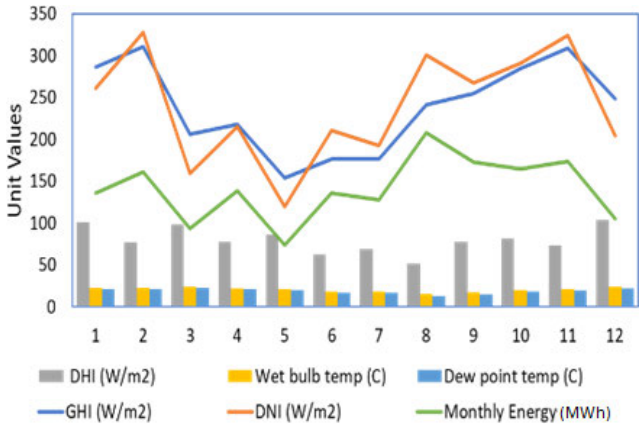


**FIGURE 6.** Influential input parameters.

### B. DATASET STANDARDIZATION

The standardization process is used to prepare the dataset to better fit and keep the deviation minimum. For the dataset matrix $\mathcal{M}_{ij}$, the mean and standard deviation are estimated to get standardizing dataset of $\mathcal{S}$.

$$\mu = \int_{ij=1}^{n} \mathcal{M}_{ij} \tag{13}$$

$$\sigma = \sqrt{\frac{\sum_{ij=1}^{n} \left(\mathcal{M}_{ij} - \bar{\mathcal{M}}\right)}{n-1}} \tag{14}$$

$$\delta = \frac{\left(\mathcal{M}_{ij} - \mu\right)}{\sigma} \tag{15}$$

In (13)-(15), $\mathcal{M}_{ij}$ is the dataset matrix, $\mathcal{S}$ is the standardized dataset, $\mu$ is the mean, and $\sigma$ is the standard deviation of the dataset.

### C. PARTITION OF TRAINING AND TEST DATA

The 90% of the available data are used for training, while, the other 10% are used for testing. The training data size can be estimated as in (16):

$$\mathbb{T} = 0.9 \times \sin(\delta) \tag{16}$$

### D. PREPARE PREDICTORS AND RESPONSES

The training sequences are shifted by $n$ time steps to forecast the value in future time. This has been done to make sure that the proposed method could learn to predict ahead of input sequences. The predictor and responses for the proposed algorithm can be obtained as in (17) and (18):

$$Xtrain = \delta(1 : \mathbb{T} - n)) \tag{17}$$

$$Ytrain = \delta(2 : \mathbb{T}) \tag{18}$$

### E. HYBRID DEEP LEARNING ARCHITECTURE

The hybrid deep learning (LSTM-CNN) architecture has been designed using LSTM and CNN deep learning techniques. Due to the weather variability, it is difficult to predict PV output accurately in longer time horizon. The CNN has intelligently adapting mechanism to understand complex relationships of properties in variable nature which motivated us to choose CNN over other deep learning methods to predict yearly PV output. As illustrated in Fig. 8, a deep learning network using two LSTM layers denoted as $LSTM_1$ and $LSTM_1$ with 500 and 1000 hidden layers are initially considered. These LSTM layers then combined with input data $I_{years}$ which is 5 by 12 matrix as presented in (19).


**FIGURE 7.** Steps to prepare data for machine learning.

$$I_{years} = \begin{bmatrix} x_{11} & - & x_{1i} \\ - & - & - \\ x_{51} & - & x_{5i} \end{bmatrix}_{years}^{i=1\ldots12} \tag{19}$$

The LSTM network is designed with fully connected layer and regression output layer to get $O_{nyears}$ outputs. The LSTM network was set with training option properties as given in Table 2.

The LSTM network is designed to predict input values in future time of *nyears* where $(nyears = years + n)$. The value

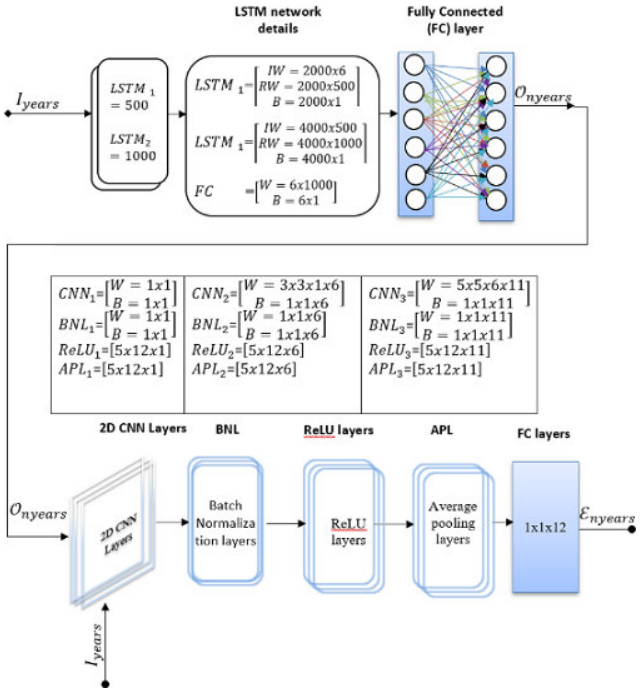


**FIGURE 8.** Hybrid deep learning structure for long term PV forecasting.

**TABLE 2.** LSTM training values.

| Options | Values |
|---|---|
| **Squared Gradient Decay Factor:** | 0.9000 |
| **Epsilon** | 1.0000e-08 |
| **Initial Learn Rate** | 0.0050 |
| **Learn Rate Schedule Settings** | [1×1 struct] |
| **L2Regularization** | 1.0000e-04 |
| **Max Epochs** | 100 |
| **Mini Batch Size** | 128 |

of $n$ can be replaced by any number of years. The output $O_{nyears}$ of LSTM network is 5 *by* 12 matrix which gives all influential input values of $n$ years as presented in (20).

$$O_{nyears}^{(5\times12\times1)} = \begin{bmatrix} x_{11} & - & x_{1n} \\ - & - & - \\ x_{i1} & - & x_{in} \end{bmatrix}_{nyears}^{i=1..5} \tag{20}$$

***Step 4:*** The output from LSTM network used as the inputs into the CNN network as presented in Fig. 8. The CNN network is designed with three 2D CNN deep learning networks followed by equal numbers of BNL, ReLU layers, and APL. The CNN network also added with DL to handle overfitting. Finally, it has fully connected layers of 12 outputs for each year which is followed by the regression layer. The network was set with training option properties as given in Table 3. The CNN network is then trained using $I_{years}$ where $Iyears = 24$ from 1990 to 2013. It was then used to predict output energy $\mathcal{E}_{nyears}$ for *nyears* time as given in (21) (where value of $n_{LSTM} = n_{CNN}$).

**TABLE 3.** CNN training values.

| Options | Values |
|---|---|
| **Initial Learn Rate** | 1.0000e-03 |
| **L2 Regularization** | 1.0000e-04 |
| **Momentum** | 0.9000 |
| **Learn Rate Schedule Settings** | [1×1 struct] |
| **L2 Regularization** | 1.0000e-04 |
| **Max Epochs** | 300 |
| **Mini Batch Size** | 128 |

$$\mathcal{E}_{nyears} = \begin{bmatrix} x_{11} & - & x_{in} \end{bmatrix}_{nyears}^{i=1...12} \tag{21}$$

## V. RESULTS AND DISCUSSIONS

### A. PREDICTION RESULTS

The forecasted performance is tested in North Queensland locations, e.g. Cairns, Gladstone, Rockhampton, and Townsville. However, due to the brevity, only the results related to Cairns are presented in this section. Historical meteorological data from 1990 to 2013 in Cairns has been used for the training of the model. The downloaded data files have some low quality, missing data, and format compliance to SAM and the proposed prediction model. To resolve these problems, data cleaning has been carried out based on the physical model. The SAM has also been used for data cleaning in this paper. For example, if the PV output obtained more than the capacity value for very low irradiance or output of PV at night, flagged as bad data. Sometimes the PV output could be obtained due to missing data. This is also flagged as bad data. Similar to [7], 5271 hours out of 5461 daytime are considered as good data in this work. The bad data are excluded from the training of the proposed method. Often the missing data have been filled based on the previous hour of measurements. The SAM model is later used to compare the forecast model with the baseline PV model.

After processing the monthly weather input parameters and energy output estimated by step 1 and step 2 given in Section IV. The historical dataset of 24 years with a list of input values are established. These have been processed later to pepare a input matrix $I_{years=24}$ as in (18) for LSTM (see step 3). The proposed LSTM algorithm predict output matix $\mathcal{O}_{nyears=6}$ for 6 years as illustrated in Fig. 9.

In Fig. 9, GHIs from 2014 to 2020 are presented. The predicted value of GHIs is compared with the actual measured values to validate the performance of the proposed method. From the results, it is evident that the forecasted values are well-matched with the actual measurements. The rest of the GHI values from 2015 to 2020 are predicted using the proposed hybrid algorithm. These predicted values then used with other predicted input $\mathcal{O}_{nyears} = 6$ as given in (19) to get the energy outputs from 2015 to 2020 in Cairns (Latitude −16.8833 and longitude 145.75).

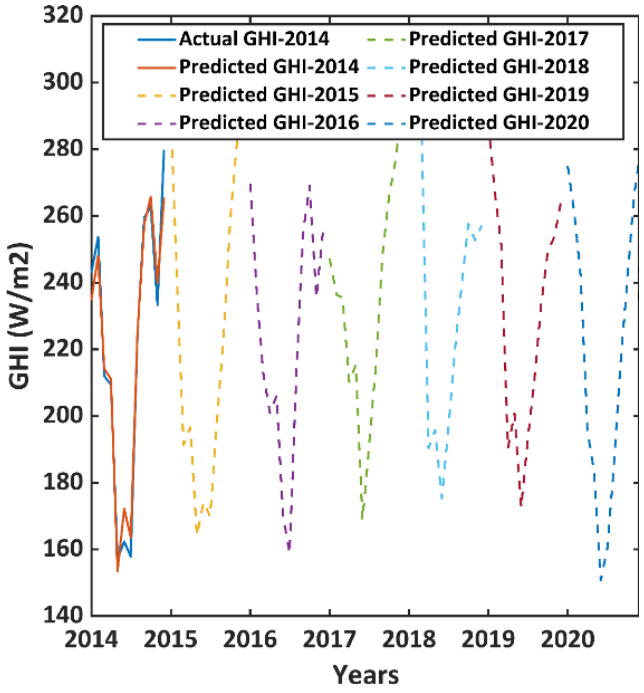


**FIGURE 9.** Yearly forecasted GHI in Cairns.

Fig. 10 shows the actual and predicted energy output for the PV system in Cairns for 2014 to validate the accuracy of the model. The physical model of the PV system and the actual meteorological data are used to get the actual value, whereas, the predicted meteorological data have been used to get the PV output for 2014 using the physical model of PV. From the results given in Fig. 10, it is evident that the predicted energy output almost accurately matched with the actual energy output of 2014 with errors less than 3.3. It should be worth noting that energy prediction in May and September showed the highest positive errors, while August has the highest negative error. From the results in Fig. 10, it is evident that the forecasted energy value closely matched with the actual values of energy in 2014. Thereby, it is evident that the proposed method is able to forecast the long-term energy from PV systems.

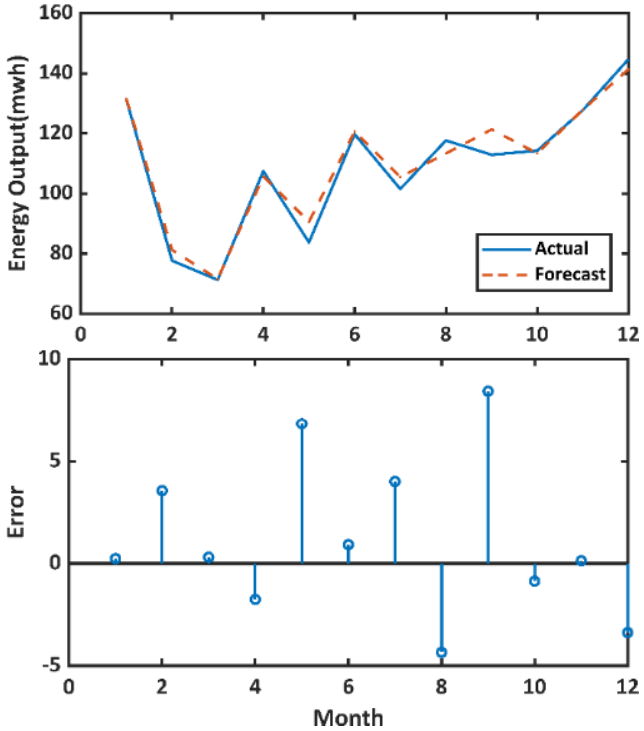


**FIGURE 10.** Monthly energy prediction at Cairns in 2014.

The proposed method is further used to estimate the energy output of a PV system at Cairns. Yearly predicted energy outputs are given in Fig. 11 for 2015 to 2020. From the yearly predicted results, it is evident that the energy production would be high from September to January and low from February to August – which are the general trends for the PV systems in North Queensland.

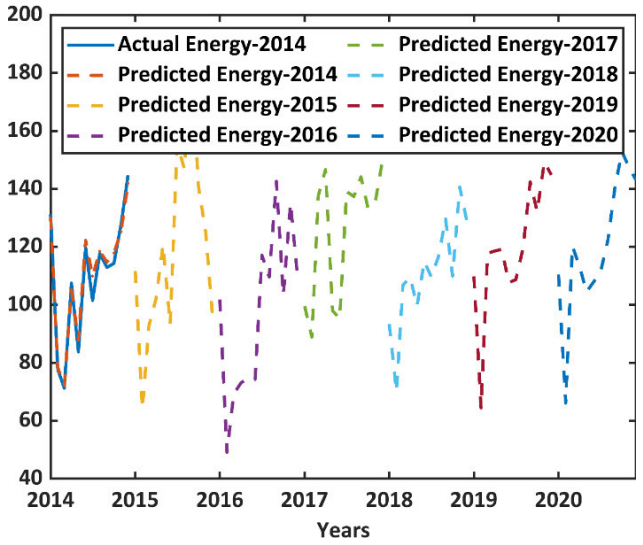


**FIGURE 11.** Yearly predicted energy value.

### B. COMPARATIVE ANALYSIS

There are no standard sets of performance comparison parameters to be used in the existing forecasting techniques. Hence, it is important to cover a reasonable range of performance parameters for benchmarking the proposed method. Four well-known forecasting performance parameters such as RMSE, nRMSE, mean absolute percent error (MAPE), and Rvalue are used to benchmark the proposed algorithm.

The RMSE is more sensitive to forecast errors [14], [26]. Hence, it is suitable where the small errors are more tolerable than the larger ones. The RMSE can be expressed as in (22) [14]:

$$RMSE = \sqrt{\frac{1}{N} \times \int_{i=1}^{N} \left(PV_i^{\alpha} - PV_i^{f}\right)^2} \qquad (22)$$

In (22), $PV_i^a$ is actual $PV$ output power, $PV_i^f$ is forecasted power, and $N$ the number of observations. The RMSE error is nomalized with respect to maximum and minimum value of $PV_i^f$ to get nRMSE as given in (23). It should be noted that the lower the RMSE and nRMSE values, better the performance of the algorithm for forecasting.

$$NRMSE = \frac{RMSE}{PV^f_{max} - PV^f_{min}} \tag{23}$$

The MAPE is widely used index to determine the forecast accuracy with respect of scale-independency and interpretability. The MAPE and error variance can be calculated as in (24) [17], [26], and [27]:

$$MAPE\ (\%) = \frac{1}{FH} \int_{t=1}^{FH} \frac{PV_i^\alpha - PV_i^f}{PV_t^\rho} \tag{24}$$

where $FH$ is the forecast horizon and $PV_t^p$ is the peak output power at time $t$. A higher $MAPE$ value means lower accuracy of forecasting algorithms whereas lower MAPE value means high accuracy of the forecasting algorithms.

The Rvalue is the correlation between the predicted values and the observed values [27]–[29]. It gives an idea about the model generalization. An Rvalue close to one means, the forecasting values are highly close to the fitted regression line and it can be used in more generalized cases. The Rvalue can be calculated as in (25) [27]–[29]:

$$Rvalue = \left(1 - \frac{\sum \left(PV_i^\alpha - PV_i^f\right)^2}{\sum \left(PV_i^f\right)^2}\right) \tag{25}$$

Table 4 shows the baseline comparison of the proposed method against the four well-established methods given in the literature to forecast the long-term energy from the PV system. All four benchmarking performance indices mentioned earlier are used for the comparison. From the results given in Table 4, it is evident that the proposed method has the RMSE of 3.89 which is very low compared to the other methods. The nRMSE value for the proposed method is 0.0529, considerably low with compared to others. However, this can be further improved with training. The given algorithm outperforms all other existing algorithms in MAPE which is 2.83 for the studied location. Furthermore, the Rvalue of the proposed method is 0.9. This indicates that the proposed method is very close to the fitted regression line. Moreover, it is worth noting that the Rvalue of the proposed method is higher with compared to statistical analysis. However, Rvalue of the given method is slightly low with compared to random forest and NNE. From the comparative results, it is evident that the proposed forecasting algorithm is more accurate for forecasting the long-term energy output from PV system.

Fig. 12 illustrated the values of RMSE and MAPE for all four studied locations in the North Queensland, e.g. Cairns, Gladstone, Rockhampton, and Townsville for the proposed method. From the results given in Fig. 12, it is evident that the RMSE values are lower than 15 in all studied locations

**TABLE 4. Benchmarking of the proposed algorithm–carins.**

| Algorithm | RMSE | nRMSE | MAPE | Rvalue |
|---|---|---|---|---|
| ANN | - | 0.2257 | 11.42 | - |
| Random Forest | 27.5 | 0.1356 | 8.84 | 0.94 |
| NNE | 80.62 | 0.1965 | 8.73 | 0.98 |
| Statistical analysis | 15.80 | 0.1171 | 5.67 | 0.69 |
| Proposed deep learning method | 3.89 | 0.0529 | 2.83 | 0.90 |

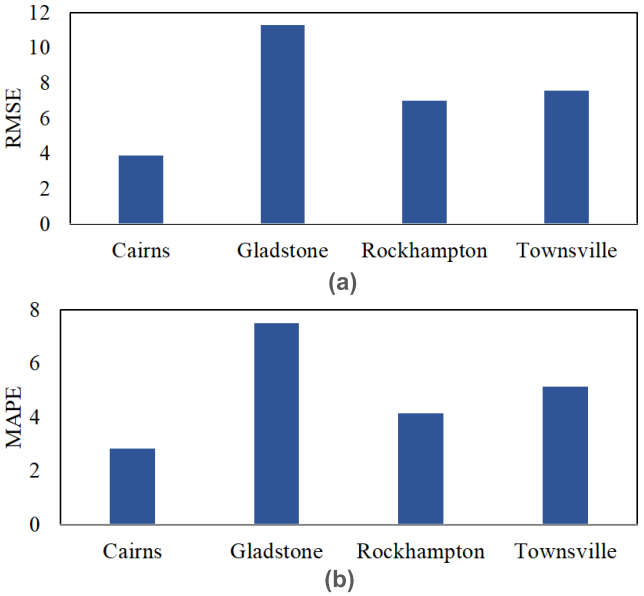


**FIGURE 12. Comparative results of various locations in North Queensland: (a) RMSE; (B) MAPE.**

for the given method. Moreover, the given algorithm has good forecast quality for various locations with RMSE ranging between 3.89 – 11.87. It should be worth to note that the MAPE values of the studied locations are ranging between 2.5 –7.8, which makes the proposed method more reliable in estimating long-term energy output of PV.

Further analyses are conducted to evaluate the reliability of the proposed method for different datasets and layers for LSTM. The RSME, MAPE, nRMSE, and Rvalue are used as the indices to measure the reliability of the proposed method. Table 6 gives the performance of the proposed model under different lengths of training data (i.e. 5 years, 10 years, and 25 years).

Table 7 shows the performance of the proposed method under different LSTM layers and standard deviation of indices in relation to result presented in Table 4. From the results given in Table 6 and 7, it is evident that *the* mean standard deviations of all the indices are lower in relation to actual values presented in Table 4. For example, the average RMSE standard deviations varies in worst case scenarios is $\pm 1.2$ only whereas $\pm 0.2$ best case scenarios. Therefore, the performance indices for the given method are lower under various factors affecting the forecasting performance.

**TABLE 5.** Rvalue of various locations.

| Location | Rvalue |
|---|---|
| Cairns | 0.90 |
| Gladstone | 0.97 |
| Rockhampton | 0.99 |
| Townsville | 0.98 |

**TABLE 6.** Performance of the proposed method for various length of training data.

| No of year | RMSE | nRMSE | MAPE | Rvalue |
|---|---|---|---|---|
| 5 | 6.78(±1.4) | 0.1458(±0.5) | 13.50(±5) | 0.80(±0.4) |
| 10 | 4.69(±0.4) | 0.1342(±0.4) | 4.58(±0.9) | 0.83(±0.4) |
| 25 | 3.89(±0) | 0.0529(±0) | 2.83(±0) | 0.95(±0.06) |

**TABLE 7.** Performance of the proposed method for various layer of LSTM.

| No of layer in LSTM | RMSE | nRMSE | MAPE | Rvalue |
|---|---|---|---|---|
| 1 | 7.15(±1.6) | 0.0946(±0.02) | 5.89(±1.5) | 0.92(±0.01) |
| 2 | 4.89(±0.5) | 0.1229(±0.03) | 3.69(±0.4) | 0.93(±0.01) |
| 3 | 3.69(±0.1) | 0.0963(±0.02) | 2.95(±0.06) | 0.95(±0.03) |

From the results given in Table 6, it should be worth noting that the Rvalue reduced significantly for the smaller training dataset. Moreover, the MAPE value for the lower training datasets is also high. However, the average changes are low which suggests that the proposed method is reliable.

## VI. CONCLUSION

This paper proposed a new and reliable method forforecasting the long-term output of solar PV. The proposed method utilized the multivariate long and short-term memory and convolutional neural network to develop the technique for forecasting the PV output. This paper utilized the twenty-four years of historical data from various locations in North Queensland in Australia to validate the performance of the developed model. Additional meteorological parameters have been used in the proposed algorithm based on their positive and negative influences on the output of the PV system. From the given results and comparisons, it is evident that the proposed method may accurately predict the long-term output of the PV system for planning studies with RMSE lower than 15 for all studied locations. Moreover, the proposed method is robust compared to some well-established methods such as ANN, Random Forest, NNE, and others. The proposed algorithm was run in MATLAB R2018b (9.5) with the computational cost for training and prediction of 203.63 s. Therefore, it could be considered as a low computation cost algorithm compared to others.

In this study, several assumptions had to make for PV output forecasting. Therefore, further sensitivity study around this domain would be performed in the future. This work will be further extended to forecast the long-term generation reserve in power systems with high penetration of wind and solar in Australia.

## REFERENCES

[1] Y. P. Agalgaonkar, B. C. Pal, and R. A. Jabr, "Statistical distribution system operation considering voltage regulation risks in the presence of PV generation," *IEEE Trans. Sustain. Energy*, vol. 6, no. 4, pp. 1315–1324, Oct. 2014.

[2] M. Marinelli, P. Maule, A. N. Hahmann, O. Gehrke, P. B. Norgard, and N. A. Cutululis, "Wind and photovoltaic large-scale regional models for hourly production evaluation," *IEEE Trans. Sustain. Energy*, vol. 6, no. 3, pp. 916–923, Jul. 2015.

[3] A. Tascikaraoglu, B. M. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. G. Paterakis, and J. P. S. Catalao, "Compressive spatio-temporal forecasting of meteorological quantities and photovoltaic power," *IEEE Trans. Sustain. Energy*, vol. 7, no. 3, pp. 1295–1305, Jul. 2016.

[4] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 2, no. 1, pp. 2–10, Mar. 2009.

[5] M. P. Mittermaier and R. Bullock, "Using MODE to explore the spatial and temporal characteristics of cloud cover forecasts from high-resolution NWP models," *Meteorological Appl.*, vol. 20, no. 2, pp. 187–196, Jun. 2013.

[6] F. Bizzarri, M. Bongiorno, A. Brambilla, G. Gruosso, and G. S. Gajani, "Model of photovoltaic power plants for performance analysis and production forecast," *IEEE Trans. Sustain. Energy*, vol. 4, no. 2, pp. 278–285, Apr. 2013.

[7] 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.

[8] D. P. Larson, L. Nonnenmacher, and C. F. M. Coimbra, "Day-ahead forecasting of solar power output from photovoltaic plants in the American southwest," *Renew. Energy*, vol. 91, pp. 11–20, Jun. 2016.

[9] E. Zorita and H. von Storch, "The analog method as a simple statistical downscaling technique: Comparison with more complicated methods," *J. Climate*, vol. 12, no. 8, pp. 2474–2489, Aug. 1999.

[10] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, D. Renné, and T. E. Hoff, "Validation of short and medium term operational solar radiation forecasts in the US," *Sol. Energy*, vol. 84, no. 12, pp. 2161–2172, Dec. 2010.

[11] C. Yang, A. A. Thatte, and L. Xie, "Multitime-scale data-driven spatio-temporal forecast of photovoltaic generation," *IEEE Trans. Sustain. Energy*, vol. 6, no. 1, pp. 104–112, Jan. 2015.

[12] 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.

[13] *Australian Solar Energy Forecasting System Final Report: Project, Results, and Lessons Learnt*, Commonwealth Sci. Ind. Res. Org. (CSIRO), Canberra, ACT, Australia, 2016. [Online]. Available: https://arena.gov.au/assets/2016/07/Aus-Solar-Energy-Forecasting-System-Final-Report.pdf

[14] L. Benali, G. Notton, A. Fouilloy, C. Voyant, and R. Dizene, "Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components," *Renew. Energy*, vol. 132, pp. 871–884, Mar. 2019.

[15] A. Y. Alanis, L. J. Ricalde, C. Simetti, and F. Odone, "Neural model with particle swarm optimization Kalman learning for forecasting in smart grids," *Math. Problems Eng.*, vol. 2013, pp. 1–9, 2013.

[16] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[17] 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.

[18] K. Yan, W. Li, Z. Ji, M. Qi, and Y. Du, "A hybrid LSTM neural network for energy consumption forecasting of individual households," *IEEE Access*, vol. 7, pp. 157633–157642, 2019.

[19] *Exemplary Energy Pty Ltd*. [Online]. Available: http://www.eemplary.com.au
[20] H. D. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *J. Physiol.*, vol. 148, pp. 574–591, Oct. 1959.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[22] The MathWorks, Inc. (2017). *Specify Layers of Convolutional Neural Network*. [Online]. Available: https://au.mathworks.com/help/deeplearning
[23] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Ciresan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Kuala Lumpur, Malaysia, Nov. 2011, pp. 342–347.
[24] N. Blair, A. Dobos, J. Freeman, T. Neises, and M. Wagner. (2014). *System Advisor Model, SAM*. [Online]. Available: https://www.nrel.gov/docs/fy14osti/61019.pdf
[25] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Statistician*, vol. 42, no. 1, pp. 59–66, Feb. 1988.
[26] A. Manthiri. (2017). *PSNR MSE R RMSE NRMSE MAPE Calculation*. [Online]. Available: https://au.mathworks.com/matlabcentral
[27] A. Zagouras, H. T. C. Pedro, and C. F. M. Coimbra, "On the role of lagged exogenous variables and spatio–temporal correlations in improving the accuracy of solar forecasting methods," *Renew. Energy*, vol. 78, pp. 203–218, Jun. 2015.
[28] B. Ray and R. Shah, "Performance assessment of prospective PV systems in queensland and new south wales of australia," in *Proc. IEEE PES GTD Grand Int. Conf. Exposit. Asia (GTD Asia)*, Bangkok, Thailand, Mar. 2019, pp. 200–205.
[29] F. Li, C. Li, J. Shi, J. Zhao, X. Yang, and Z. Chen, "Evaluation index system for photovoltaic systems statistical characteristics under hazy weather conditions in central China," *IET Renew. Power Gener.*, vol. 11, no. 14, pp. 1794–1803, Dec. 2017.

**BIPLOB RAY** (Member, IEEE) received the Ph.D. degree in information technology from Deakin University, Australia. He is working as a Senior Lecturer in information technology with CQUniversity, with a background of a mix of research, academic, and industry experience. His teaching and research are currently focused on networked intelligent systems, big data, security protocols, and the privacy of mHealth. His high-quality research work has been recognized by peers and cited extensively. He has more than 20 international journal and conference publications. He has appeared as a keynote/plenary speaker at a number of international conferences.

**RAKIBUZZAMAN SHAH** (Member, IEEE) received the Ph.D. degree from The University of Queensland, Brisbane, QLD, Australia. He is a Senior Lecturer in smart power systems engineering with the School of Engineering, Information Technology, and Physical Sciences, Federation University Australia (FedUni Australia). Prior to joining FedUni Australia, he has worked with The University of Manchester, The University of Queensland, and Central Queensland University. He has experience working at and consulting with DNOs and TSOs on individual projects and collaborative work on large projects (EPSRC project on multi-terminal HVDC, Scottish, and Southern energy multi-infeed HVDC)-primarily on the dynamic impact of integrating new technologies and power electronics into large systems. He is an active member of the CIGRE. He has more than 60 international journal and conference publications, including 18 journals in the IEEE and IET, and has spoken at leading power system conferences around the world. His research interests include future power grids, such as renewable energy integration and wide-area control, asynchronous grid connection through VSC-HVDC, power system stability and dynamics, the application of data mining in power systems, the application of control theory in power systems, distribution system energy management, and low-carbon energy systems.

**MD. RABIUL ISLAM** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Technology Sydney (UTS), Sydney, Australia, in 2014.

He was appointed as a Lecturer with the RUET, in 2005, and promoted to a Full Professor, in 2017. In early 2018, he joined the School of Electrical, Computer, and Telecommunications Engineering (SECTE), University of Wollongong (UOW), Wollongong, Australia. He has received several funding from government and industries, including the Australian Government ARC Discovery Project 2020 entitled A Next-Generation Smart Solid-State Transformer for Power Grid Applications. He has authored or coauthored 150 articles, including 40 IEEE TRANSACTIONS in international journals and conference proceedings. His research interests are in the fields of power electronic converters, renewable energy technologies, power quality, electrical machines, electric vehicles, and smart grids. He is also a member of the Australian Power Quality and Reliability Center, UOW. He has written or edited four technical books published by Springer. He has served as a Guest Editor for the IEEE TRANSACTIONS ON ENERGY CONVERSION, the IEEE TRANSACTIONS ON APPLIED SUPERCONDUCTIVITY, and *IET Electric Power Applications*. He has been serving as an Editor for the IEEE TRANSACTIONS ON ENERGY CONVERSION and the IEEE POWER ENGINEERING LETTERS, and an Associate Editor for IEEE ACCESS.

**SYED ISLAM** (Fellow, IEEE) received the B.Sc. degree in electrical engineering from the Bangladesh University of Engineering and Technology, Bangladesh, in 1979, and the M.Sc. and Ph.D. degrees in electrical power engineering from the King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 1983 and 1988, respectively. He has been a Visiting Professor with the Shanghai University of Electrical Power and Xi'an Jiatong University, China. He is currently the Dean of the School of Engineering, Information Technology, and Physical Sciences, Federation University Australia, Australia. He received the Dean's Medallion for Research at Curtin University, in 1999. He has published more than 300 technical articles in his area of expertise. His research interests include the condition monitoring of transformers, wind energy conversion, and smart power systems. He has been a keynote speaker and an invited speaker at many international workshops and conferences. He was a member of the Steering Committee of the Australian Power Institute and the WA EESA Board. He is a Fellow of the Engineers Australia and IET, and a Chartered Engineer in U.K. He received the IEEE T. Burke Haye's Faculty Recognition Award, in 2000. He received the Curtin University Inaugural Award for Research Development, in 2012. He received the Sir John Madsen Medal for the Best Electrical Engineering Paper in Australia, in 2011 and 2014. He was the Founding Editor of the IEEE TRANSACTION ON SUSTAINABLE ENERGY and is currently an Associate Editor of *IET Renewable Power Generation*. He was the Guest Editor-in-Chief of the IEEE TRANSACTION ON SUSTAINABLE ENERGY special issue on Variable Power Generation Integration into Grid.