Energy 148 (2018) 461–468

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM

Xiangyun Qing*, Yugang Niu

Key Laboratory of Advanced Control and Optimization for Chemical Processes (East China University of Science and Technology), Ministry of Education, East China University of Science and Technology, Shanghai, China

ARTICLE INFO

Article history:
Received 15 September 2017
Received in revised form
3 January 2018
Accepted 30 January 2018
Available online 3 February 2018

Keywords:
Neural networks
Solar irradiance prediction
Structured output prediction
Weather forecasting

ABSTRACT

Prediction of solar irradiance is essential for minimizing energy costs and providing high power quality in electrical power grids with distributed solar photovoltaic generations. However, for residential and small commercial users deploying on-site photovoltaic generations, the historical irradiance data can not be obtained directly because of expensive solar irradiance meters. Thanks to increasingly improved weather forecasting service provided by local meteorological organizations, weather forecasting data such as temperature, dew point, humidity, visibility, wind speed and descriptive weather summary, are becoming readily available through the Internet, while the irradiance forecasting data are often unavailable. This paper proposes a novel solar prediction scheme for hourly day-ahead solar irradiance prediction by using the weather forecasting data. This study formulates the prediction problem as a structured output prediction problem jointly predicting multiple outputs simultaneously. The proposed prediction model is trained by using long short-term memory (LSTM) networks taking into account the dependence between consecutive hours of the same day. We compare persistence algorithm, linear least square regression and multilayered feedforward neural networks using backpropagation algorithm (BPNN) for solar irradiance prediction. The experimental results on a dataset collected in island of Santiago, Cape Verde, demonstrate that the proposed algorithm outperforms these competitive algorithms for single output prediction. The proposed algorithm is %18.34 more accurate than BPNN in terms of root mean square error (RMSE) by using about 2 years training data to predict half-year testing data. Moreover, compared with BPNN, the proposed algorithm also shows less overfitting and better generalization capability. For a case using 10 years of historical data to predict 1 year of irradiance data, the prediction RMSE using the proposed LSTM algorithm decreases by 42.9% against BPNN.

1. Introduction

Distributed photovoltaic (PV) generation has a number of potential benefits such as reduced distribution losses and increased grid integration ability because self-consumption of distributed generation can partly match the consumption with the intermittent generation. Furthermore, the intermittency of PV generation can be compensated by using energy storage. The energy storage can be used as a means to increase the self-consumption ratio. With the introduction of the Tesla Powerwall announced in May 2015, the combined PV-storage systems are starting to rise in popularity for residential and small commercial users to reduce electricity costs. In Germany, the PV-storage systems for households have received funding through to 2018. As a rival to distributed PV-storage leader Germany, Australia has also been on a residential PV-storage boom in recent years.

However, with the decreasing of feed-in tariffs for solar PV power in many countries, seeking to maximize the benefits of PV-storage system has accelerated the necessity of energy management system (EMS) for the residential and small commercial users. The EMS schedules and coordinates the user's energy consumption, distributed generation and storage to minimize energy costs while providing high power quality. As an example shown in Ref. [1], for rooftop PV and battery storage systems, a one-day ahead energy management strategy performed before the start of each day needs to incorporate the variations in PV power outputs and electrical demands of the day. The prediction error of PV output power can cause a negative effect on the economical profit of the PV-storage

* Corresponding author.
E-mail address: xytsing@ecust.edu.cn (X. Qing).

https://doi.org/10.1016/j.energy.2018.01.177
0360-5442/

systems. Therefore, accurate forecasting of solar irradiance and thus PV power generation can reduce the impact of uncertainty of PV generation, improve the control algorithms of battery storage charge controllers and bring the significant economical benefit of the PV-storage systems.

Since solar irradiance being proportional to solar power harvesting can be a proxy for solar generation, solar irradiance prediction has been studied extensively in the literature. The forecasting methods can be roughly split into three categories: physical, statistical and machine learning methods [2]. In the physical method, the forecast is implemented by numerical weather prediction models, which is more suitable for a long forecast horizons (1 and 2 days). Instead, the statistical method is based on the historical time data series, which is simpler than the physical method. However, the prediction performance of the statistical method is limited since the method is based on the concept of persistence or stochastic time series, while characteristics of the irradiance time series are non-stationary in nature. The machine learning method, as a branch of artificial intelligence, can learn from datasets to construct a nonlinear mapping between input and output data without being explicitly programmed. In this paper, we focus on the use of machine learning algorithms for solar irradiance prediction. Cao et al. [3] developed a method of the day-by-day forecast of solar irradiance from the historical day-by-day records of solar irradiance by combining artificial neural network (ANN) and wavelet analysis. Yand et al. [4] proposed three methods of the next hour solar irradiance forecast by using different types of meteorological data as input parameters, namely, global horizontal irradiance (GHI), diffuse horizontal irradiance (DHI), direct normal irradiance (DNI) and cloud cover. They found that the cloud cover information can improve the forecast accuracy. Voyant et al. [5] used a hybrid ARMA/ANN model and data issued from a numerical weather prediction (NWP) model to forecast the hourly global radiation for five places in Mediterranean area. Ahmad et al. [6] used autoregressive recurrent neural networks with exogenous inputs and weather variables to provide a one-day ahead forecast of hourly global solar irradation in New Zealand. Sharma et al. [7] developed a mixed wavelet neural network for prediction of hour-ahead and 15 min-ahead solar irradiance in the tropical region of Singapore. Based on an assumption that solar radiation data repeats itself in the history, Hocaoglu et al. [8] used a novel Mycielski based model and the recorded hourly solar radiation data for hourly solar radiation forecasting. Monjoly et al. [9] presented a hybrid approach based on multiscale decomposition methods for hourly forecasting of global solar radiation by using the historical radiation data. Obviously, above-mentioned methods are by no means exhaustive lists and many other applications of machine learning algorithms to the solar irradiance prediction can be found in the recent literature [10].

However, for residential and small commercial users using on-site PV generations, the historical irradiation data can not be easily available because expensive solar irradiance meters and advanced data acquisition systems are only deployed in large-scale solar farms that produce multiple megawatts. On the contrary, weather forecast data are becoming more and more available through the Internet. For example, nowadays each mobile phone can inquiry hourly weather data of the next day by using some weather service application softwares. Therefore, Sharma et al. [11] proposed a method for automatically creating site-specific prediction models for solar power generation from National Weather Service weather forecasts using machine learning techniques. The solar intensity was formulated as a function of weather forecast parameters such as day, temperature, dew point, wind speed, sky cover, precipitation, and humidity. The function was determined using linear least squares regression and support vector machines (SVMs). Their experimental results showed that their proposed forecasting method was a promising research area for the power output prediction from distributed generation at many small-scale facilities at smart homes and buildings. Similarly, Bae et al. [12] used various meteorological data and SVM regression for 1-h ahead prediction of solar irradiance. Recently, Ceci et al. [13] proposed an alternative prediction method by viewing the prediction of consecutive hours as a structured output prediction problem which is popular in the fields of natural language processing and text mining.

Inspired by the work of references [11–13], we propose a novel prediction scheme for hourly one-day ahead prediction of solar irradiance based on weather forecasts and long short-term memory (LSTM) networks. We first cast predicting the hourly irradiance values of the same day simultaneously as a structured output prediction problem taking into account their mutual dependence. The input data consist of the weather forecasts, month, day and hour at a given time. Then, the structured output prediction problem is learned from data using the LSTM networks. The LSTM network, as one of the most advanced recurrent neural networks, has shown remarkable results in numerous time series learning tasks such as artificial handwriting generation, language forecasting, and speech recognition [14,15]. The LSTM network introduced by Hochreiter and Schmidhuber in Ref. [16], has shown a superior ability to learn long-term dependencies by maintaining a memory cell to determine which unimportant features should be forgot and which important features should be remembered during the learning process. Therefore, using the LSTM for modeling the hourly irradiance data, not only can the dependence between consecutive hours of the same day be captured, but the long-term (e.g. seasonal) behavior can be learned. Once the LSTM network is trained and tuned, it can be used to predict the houly solar irradiance values using the weather forecasts. Due to the proposed prediction tailored to the scheduling stage of the EMS for PV-storage or microgrid systems, intra-hour solar irradiance variations are assumed to be handled by the energy storage units [17].

The remainder of this paper are organized as follows. Section 2 covers the dataset description used in this study, data analysis and the structured output prediction of the hourly day-ahead irradiance values by using the weather forecasting data. Section 3 provides a brief introduction about the LSTM network and its application to the structured output prediction problem. Section 4 presents the results. Finally, section 5 concludes.

2. Data

We collect the irradiance dataset for 30 months (March 2011 to August 2012 and January 2013 to December 2013) from the solar power plant in island of Santiago, Cape Verde. The Cape Verde archipelago lies on the extreme southwest corner of the western Palaearctic region. Instead of the classic four seasons, the Cape Verde archipelago only have two seasons: time of winds from October to mid-July and rainy season from August to September. The temperatures of Cape Verde range between 21^oC and 29^oC, which are milder than that of the African mainland. Cape Verde has about 350 days of sunshine and thus has excellent solar energy potential. Recently, the government of Cape Verde has received a grant from the World Bank to finance the distributed solar energy system project for self-consumption of the central and regional hospitals [18]. Taking into account expensive diesel fuel prices in the archipelago, energy storage could be an economically attractive solution in order to increase wind and solar energy penetration [19].

The houly GHI data are obtained by averaging the data collected with 1 h. Only the hours between 8:00 AM and 6:00 PM are

considered as the duration of sunshine and true solar time. Due to some missing values raised by the fault of the supervisory control and data acquisition (SCADA) system of the solar plant, in fact, the dataset only covers the hourly irradiance data with 875 days. The corresponding weather data such as temperature, dew point, humidity, visibility, wind speed and descriptive weather summary, are collected from a weather service providing historical and real-time weather information via the Internet (http://www.wunderground.com). The hourly historical weather data at Praia International Airport close to the solar plant are refereed as the weather forecast data. Table 1 summarizes the weather variables and their ranges. Thirteen different types of weather in the collected data give descriptive weather summary, which consist of 1) sunny, 2) mostly sunny, 3) clouds, 4) cloudy, 5) overcast, 6) light drizzle, 7) light rain, 8) moderate rain, 9) thunderstorm, 10) rainstorm, 11) fog, 12) steam fog and 13) heavy rain. Subsequently, the weather types are presented as the weather parameters by using the numbers from 1 to 13. The selection of the input variables for a prediction task is particularly pertinent, but the choice of variables dependents on data availability and their correlation [6]. Therefore, a statistical analysis was first carried out to check the correlation of each available weather variable with solar irradiance as following.

Fig. 1 shows the correlation coefficients of all the 5 weather variables with the solar irradiance values by using all data. It can be observed that the temperature variable is moderately positively correlated with the irradiance, while the humidity variable is moderately negatively correlated with the irradiance. The correlation results show that the hourly solar irradiance is likely to be high as the increased temperature and the decreased humidity. Other weather variables have low correlation values. Thus, only using the individual weather variable and ignoring the time-domain relations of consecutive hourly data, we cannot build an effective relationship between the individual weather variable and irradiance.

Furthermore, a box plot depicted in Fig. 2 presents the distribution of all irradiance data at noon (12:00 AM) for each month in 2013. We can find the season behavior that the variances of the irradiance distributions in August and September are significantly larger than that of other months. It is certainly because of the rain season in Cape Verde. Similarly, a box plot depicted in Fig. 3 presents the distribution of all irradiance data at noon for each day in 2013. However, we cannot find significant differences of irradiance values among the days. Fig. 4 shows a box plot of the distribution of hourly irradiance data throughout 2013. Obviously, the solar irradiance values are highly correlated with the hours for each day. The GHI value is low at the beginning of the day and increases to the peak value at noon and then gradually decreases in the afternoon. The outliers of the box plot in Fig. 4 are yielded by some terrible weather conditions. Thus, only the outliers below the first quartile are found, while the outliers above the third quartile do not appear.

Table 1
Weather variables and value ranges at Praia International Airport.

Weather variables	Value Ranges
Temperature (°C)	18–39
Dew Point (°C)	−2–29
Humidity	12%–100%
Visibility (Kilometer)	0.5–10
Wind Speed (m/s)	0–15.4
Weather Type	13 types of weather

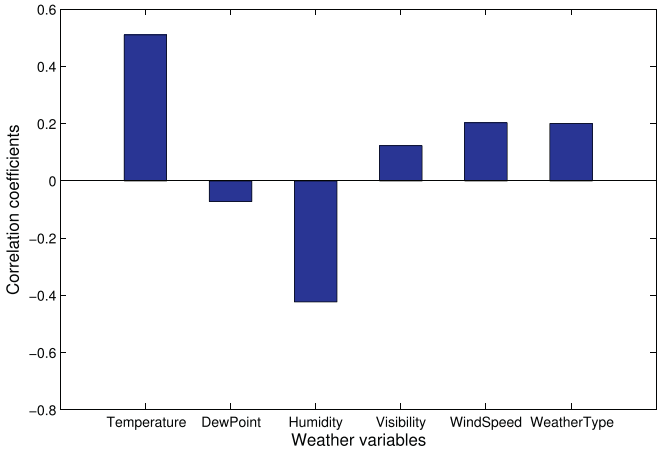

Fig. 1. Correlation of weather variables with solar irradiance.

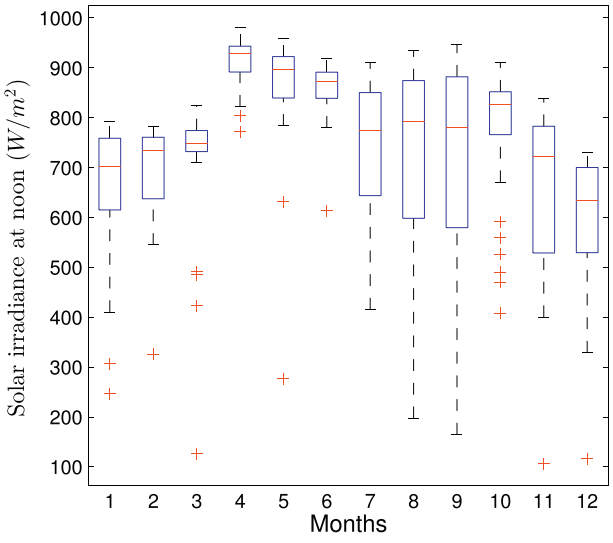

Fig. 2. Distribution of all irradiance data at noon (12:00AM) for each month in 2013.

Based on the data analysis results above, we combine month, day of the month, hour of the day and the hourly weather forecast data provided by the weather service systems as a feature vector $\mathbf{x}_t$ at a specific hour t, where $\mathbf{x}_t$ is denoted by $\mathbf{x}_t = [x_{t,1}, x_{t,2}, \ldots, x_{t,9}]^T$. Elements of the feature vector $\mathbf{x}_t$ are given as:

- $x_{t,1}$: month;
- $x_{t,2}$: day of the month;
- $x_{t,3}$: hour of the day;
- $x_{t,4}$: temperature at the hour t
- $x_{t,5}$: dew point at the hour t
- $x_{t,6}$: humidity at the hour t
- $x_{t,7}$: visibility at the hour t
- $x_{t,8}$: wind speed at the hour t
- $x_{t,9}$: weather type at the hour t

Our goal is to predict the hourly irradiance value $\mathbf{y}_t$ using the given feature vector $\mathbf{x}_t$. Classical methods such as linear regression and feedback neural network, are to learn a predictive model which performs an explicit or implicit function relationship between the individual feature vector and the individual output value at a specific hour, which neglect the dependence between consecutive hours of the same day. In this paper, we formulate prediction of the hourly day-ahead irradiance values of the same day as a structured output prediction, i.e, given an input sequence consisting of the hourly feature vectors as $\mathbf{X}_d = \{\mathbf{x}_8, \mathbf{x}_9, \ldots, \mathbf{x}_{18}\}$, produces a output

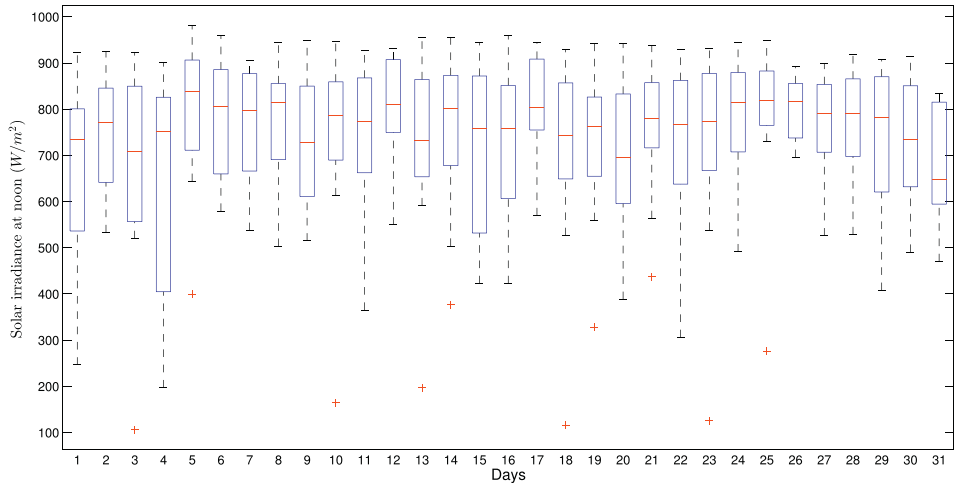

Fig. 3. Distribution of all irradiance data at noon for each day in 2013.

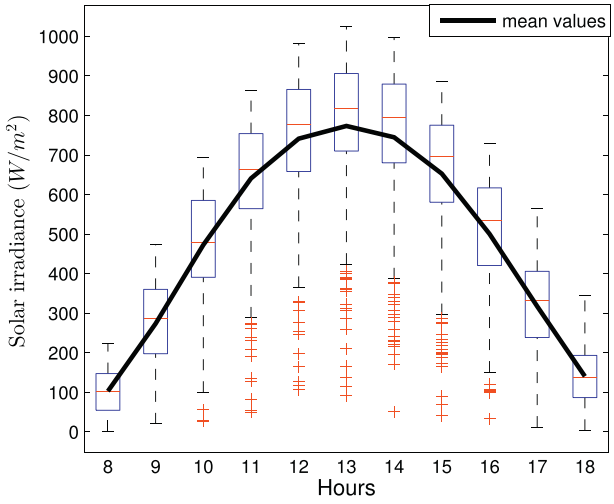

Fig. 4. Distribution of hourly irradiance data throughout 2013.

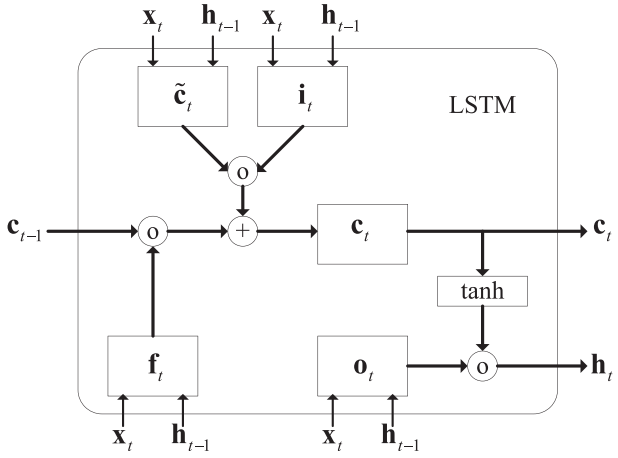

Fig. 5. Structure of a LSTM cell.

sequence consisting of the hourly irradiance values at a specific day as $\mathbf{Y}_d = \{\mathbf{y}_8, \mathbf{y}_9, \ldots, \mathbf{y}_{18}\}$.

Recurrent neural networks has been successfully applied in numerous learning problems when the data is of sequential nature, while it has received little attention for solar irradiance forecasting. The concept of time is introduced into the recurrent neural networks. The LSTM, as mentioned in the introduction section, is one of the recurrent neural networks. Compared with an usual recurrent network, the LSTM can capture autoregressive structures of arbitrary lengths. Recently, some LSTM based deep learning forecasting approaches have been presented for short-term traffic forecasting and short-term residential load forecasting [20,21]. The excellent performance of the LSTM networks and the complex relationships between weather metrics and solar irradiance values motivate our study of applying the LSTM networks to the hourly day-ahead solar irradiance prediction problem.

3. Methodology

The structure of a LSTM cell is shown in Fig. 5. In this figure, at each time t, $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ and $\tilde{\mathbf{c}}_t$ are input gate, forget gate, output gate and candidate value [22], which can be described as following equations:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{i,x}\mathbf{x}_t + \mathbf{W}_{i,h}\mathbf{h}_{t-1} + b_i) \tag{1}$$

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_{f,x}\mathbf{x}_t + \mathbf{W}_{f,h}\mathbf{h}_{t-1} + b_f\right) \tag{2}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{o,x}\mathbf{x}_t + \mathbf{W}_{o,h}\mathbf{h}_{t-1} + b_o) \tag{3}$$

$$\tilde{\mathbf{c}}_t = tanh(\mathbf{W}_{\tilde{c},x}\mathbf{x}_t + \mathbf{W}_{\tilde{c},h}\mathbf{h}_{t-1} + b_{\tilde{c}}) \tag{4}$$

where $\mathbf{W}_{i,x}$, $\mathbf{W}_{i,h}$, $\mathbf{W}_{f,x}$, $\mathbf{W}_{f,h}$, $\mathbf{W}_{o,x}$, $\mathbf{W}_{o,h}$, $\mathbf{W}_{\tilde{c},x}$ and $\mathbf{W}_{\tilde{c},h}$ are weight matrices, b_i, b_f, b_o and $b_{\tilde{c}}$ are bias vectors, $\mathbf{x}_t$ is the current input, $\mathbf{h}_{t-1}$ is the output of the LSTM at the previous time $t-1$, and $\sigma()$ is the Sigmoid activation function. The forget gate determines how much of prior memory value should be removed from the cell state. Similarly, the input gate specifies new input to the cell state. Then, the cell state $\mathbf{c}_t$ is calculated as:

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \tag{5}$$

where $\circ$ denotes the Hadamard product. The output $\mathbf{h}_t$ of the LSTM at the time t is derived as:

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t) \tag{6}$$

Finally, we project the output $\mathbf{h_t}$ to the predicted output $\tilde{\mathbf{y}}_t$ as:

$$\tilde{\mathbf{y}}_t = \mathbf{W_y}\mathbf{h}_t \tag{7}$$

where $\mathbf{W_y}$ is a projection matrix to reduce dimension of $\mathbf{h_t}$. Fig. 6 shows a structure of the LSTM networks unfolded in time. In this structure, an input feature vector $\mathbf{x_t}$ is fed into the networks at the time t. The LSTM cell at current state receives a feedback $\mathbf{h}_{t-1}$ from the previous LSTM cell to capture the time dependencies. The network training aims at minimizing the usual squared error objection function f based on targets $\mathbf{y_t}$ as

$$f = \sum_t ||\mathbf{y}_t - \tilde{\mathbf{y}}_t||^2 \tag{8}$$

by utilizing backpropagation with gradient descent. During training, the weights and bias are adjusted by using the their gradients. When one batch of the training dataset fed into the network has been learned by using the backpropagation optimization algorithm, one epoch is completed.

Since LSTM networks training is an offline task, the computation time for training is not critical for the application. However, prediction using the learned LSTM networks is very fast.

4. Experiments

4.1. Experiment 1

We performed the first experiment using historical data to predict future hourly solar irradiance. In the experiment, the data from March 2011 to June 2013 were set to training dataset. The half-year data from July 2013 to December 2013 were employed for prediction purposes. Thus, the training dataset and testing dataset are composed of hourly weather variables and irradiance values from 8:00 AM to 18:00 PM of 691 days and 184 days, respectively. We compared the prediction performance of the proposed LSTM networks algorithm with that of three benchmarking algorithms: the persistence method, the linear least squares regression method (LR) and the ANN using the classical backpropagation algorithm (BPNN). These algorithms and their parameters setting are described as follows:

- The persistence algorithm simply sets the hourly irradiance value $y_{d-1,t}$ at hour t in the previous day $d-1$ to be the day-ahead prediction value $\tilde{y}_{d,t}$ in the day d. Thus, this algorithm is free of training procedure and parameters setting. The persistence algorithm is frequently used as a baseline algorithm.

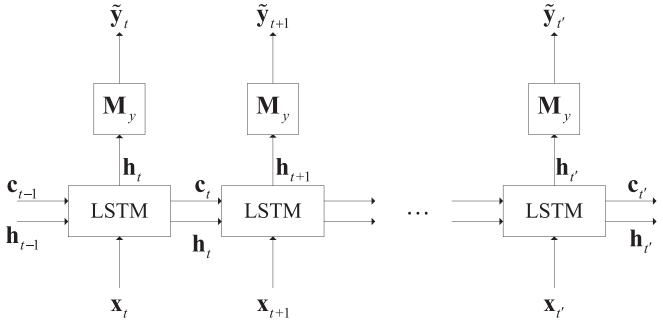

Fig. 6. Structure of LSTM networks.

- The LR algorithm minimizes the sum of the squared differences between the given hourly solar irradiance values and the values predicted by a linear approximation using the nine feature variables in the training dataset. Then, the predict accuracy of the testing dataset is verified by using the learned linear regression model. This algorithm is also free of parameters setting.
- The BPNN algorithm is available in MATLAB. The used BPNN consists of one input layer, one hidden layer and one output layer. The hidden layer neurons were set to be 50 after we made a number of experiments for an optimal choice of the hidden layer neurons. The hourly feature vectors of the training dataset were fed into the input layer, while the output layer provided the predicted hourly solar irradiance values. The 'transig' transfer function was used for all three layers. The 'traingd' (gradient descent) was selected as the training algorithm. Performance was measured by minimizing mean square error. Maximum epochs were set to be 2500.
- The LSTM networks algorithm was implemented by using the Keras deep learning package [23]. The architecture of the LSTM was given by Table 2. Input layer of our trained LSTM network had 9 features and 11 timesteps; Hidden neurons were set to be 30. Output layer with linear activation function had one neuron. Maximum epochs were set to be 50.

Because the scale of the dataset was small, the validation dataset was not set for the ANN and LSTM algorithms. The RMSE (Root Mean Square Error) was selected as the metric for the performance comparison. The RMSE quantitatively assessing the prediction performance of hourly solar irradiance values in this study is mathematically defined as:

$$RMSE = \sqrt{\left(\sum_{d=1}^{N}\sum_{t=8}^{18}\left(y_{d,t} - \tilde{y}_{d,t}\right)^2\right) \Big/ (N \times 11)} \tag{9}$$

where $\tilde{y}_{d,t}$ is the predicted value at hour t in day d, $y_{d,t}$ is the observed value, and N is the number of days in the testing dataset. Obviously, the lower the RMSE, the more accurate is the prediction. That the RMSE is prone to give more weight to error with larger values than to the prediction errors with smaller prediction errors is encouraging to design an optimal energy scheduling strategy for the EMS.

Before applying these training algorithms, the data were be normalized to [−1,1] by using linear scaling normalization technique. Because we did not do cross-validation, we run the BPNN and LSTM algorithms on the same data for 30 times with randomly initialized weights for fair comparison. Table 3 summarizes the results of four competitive algorithms. The RMSEs during the training phase were also calculated and denoted as 'training RMSE' in Table 3. As we can see, the advanced machine leaning algorithms (BPNN and LSTM) do better than the classical algorithms (Persistence and LR). Moreover, both the mean value and standard deviation obtained by using LSTM are significantly lower than that using BPNN. The mean values of training and prediction RMSEs using LSTM decrease by 24.61% and 18.34% against BPNN, respectively. Very low standard deviations of both training and prediction

Table 2
LSTM forecasting architecture based on Keras.

```
model = Sequential()
model.add(LSTM(30,input_shape=(11,9),return_sequences = True))
model.add(Dense(1,activation = 'linear'))
model.compile(loss = 'mse',optimizer = 'adam')
history = model.fit(train_x,train_t,epochs = 50,batch_size = 50)
```

Table 3
Experimental results of four algorithms. The BPNN and LSTM are run 30 times. The mean and standard deviation obtained by using BPNN and LSTM are reported.

Algorithm	training RMSE (W/m^2)	testing RMSE (W/m^2)
Persistence	/	177.031
LR	200.991	195.875
BPNN	133.1408(±13.7332)	150.2845(±12.4174)
LSTM	100.3744(±1.0398)	122.7174(±1.2029)

RMSEs using LSTM exhibit that LSTM for training and predicting the hourly solar irradiance values is almost insensitive to the initialization setting of weight values, while BPNN is highly sensitive to the initialization setting. Fig. 7 shows the results of running 30 times for BPNN and LSTM. We see that a larger training RMSE does often lead to a larger testing RMSE for both BPNN and LSTM, while a smaller training RMSE does not necessarily lead to a smaller testing RMSE. This has been studied extensively in neural networks theory about overfitting.

As an example for the hourly day-ahead irradiance prediction using different algorithms, solar irradiance values for August 20, 2013 were considered. This day was one day of rain season in Cape Verde, whose hourly weather types from 8:00 AM to 18:00 PM were overcast, cloudy, cloudy, cloudy, moderate rain, cloudy, light rain, cloudy, cloudy,cloudy, and cloudy. Thus, predicting the hourly irradiance values of this day is relatively difficult because of the rapid change in hourly weather types during the predicted day. Fig. 8 shows the true irradiance values and their prediction results for August 20, 2013. The black, red, blue, green, and cyan curves indicate true values, predicted irradiance values using the LSTM, BPNN, Persistence and LR algorithms, respectively. The RMSEs of the predicted irradiance values using the LSTM, BPNN, Persistence, LR algorithms are 126.3143(W/m^2), 147.2605(W/m^2), 253.1909(W/m^2), and 147.5852(W/m^2), respectively. Since the weather types of August 19 were significantly different from that of August 20, the Persistence algorithm obtained the worst prediction performance. The LSTM algorithm outperforms other algorithms mainly because it precisely predicted the irradiance at 12:00 AM although the weather type at this time was moderate rain.

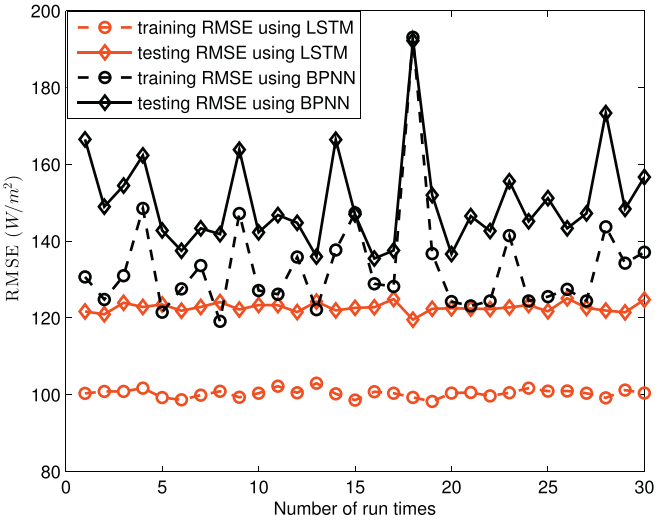

Fig. 7. RMSEs of running 30 times for BPNN and LSTM.

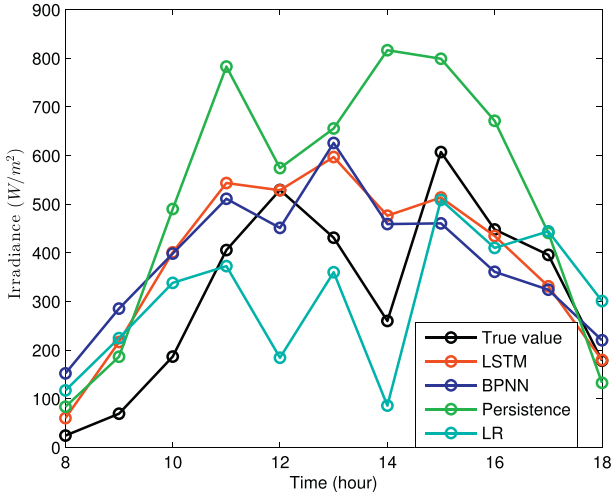

Fig. 8. Irradiance variation for August 20, 2013.

4.2. Experiment 2

We performed the second experiment by randomly splitting the whole dataset into training dataset (85%) and testing dataset (15%). The experiment aims to compare the hourly irradiance fitting generalization performance using LSTM with that using BPNN, although it is not a true prediction task. We randomly sampled 30 pairs of training and testing datasets. Each dataset were run 30 times with randomly initialized weights. A bar plot, as shown in Fig. 9, shows the mean values and stand deviations of BPNN and LSTM on 30 testing datasets. Clearly LSTM significantly outperforms BPNN on 29 sampled datasets in terms of RMSEs of predicted irradiance values, while LSTM sightly underperforms BPNN only on one dataset. Therefore, empirically we conclude that in terms of fitting generalization capability, LSTM is better than BPNN.

4.3. Experiment 3

We performed the third experiment on a dataset covering 11 years hourly data from the Measurement and Instrumentation Data Center (MIDC) (midcdmz.nrel.gov/). Irradiance and Meteorological data from NREL solar radiation research laboratory (BMS) station were used in the experiment, which can be publicly obtained. The data from January 1, 2006 to December 31, 2014 were considered as training set, while the data throughout 2015 were used as validation set for guiding the parameters optimization. The prediction accuracy was then evaluated on the data throughout 2016. Average hourly dew point temperature (Tower), relative humidity (Tower), cloud cover (Total), cloud cover (opaque), wind speed (22′) and east sea-level pressure were selected as weather variables. Maximum epochs were set to be 100 for LSTM. We searched the optimal hidden neurons for LSTM from 30 to 85 with step size 5 by minimizing RMSE of predicted irradiance values on the validation dataset. Consequently, hidden neurons were set to be 30. Similarly, the used BPNN in the experiment consists of two hidden layers with 25 and 15 neurons, respectively. Table 4 summarizes the results. As we can see the RMSE comparisons, our algorithm has significantly smaller RMSE value compared to all other algorithms. The prediction RMSE using LSTM decreases by 42.9% against BPNN. The performance improvement should be partially attributed to the large-scale training dataset and two important meteorological parameters about cloud cover. Generally speaking the larger is the training dataset for LSTM, the more accurate the prediction is.

5. Conclusion

In this work, a LSTM networks based algorithm for predicting

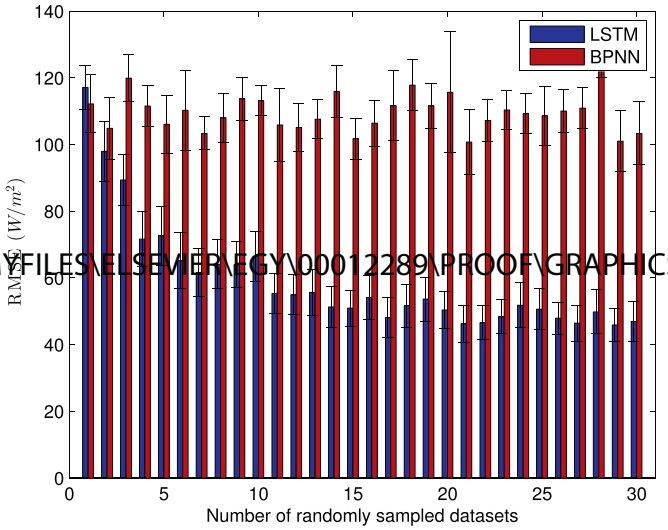

Fig. 9. Testing RMSEs of BPNN and LSTM on all 30 datasets. All datasets are run 30 times with randomly initialized weights for both BPNN and LSTM.

Table 4
Experimental results on the MIDC dataset.

Algorithm	testing RMSE (W/m^2)
Persistence	209.2509
LR	230.9867
BPNN	133.5313
LSTM	76.245

hourly day-ahead solar irradiance has been presented. The proposed algorithm uses the hourly weather forecasts of the same day and data information at the predicted time as the input variables. The outputs are presented by the hourly day-ahead irradiance values of the same predicted day. Therefore, the prediction problem is viewed as a structured output prediction problem with predicting multiple outputs simultaneously. The structured output prediction model is leaned by the LSTM networks. Experimental results on the forecasting problems show that the proposed learning algorithm taking into account the dependencies between different hours of the same day is much more accurate than some competitive algorithms such as persistence, LR, and BPNN, which only predict single output. For the Cape Verde dataset with about two years of training data, the proposed LSTM learning algorithm is %18.34 more accurate than the BPNN learning algorithm in terms of the RMSE. Moreover, experimental results on the overfitting problem using the randomly sampled training and testing datasets show that the proposed learning model overfits less than single output prediction model learned by BPNN. For the MIDC dataset with 9 years training data and 1 year validation data, the proposed LSTM algorithm is able to show a relative improvement of 42.9% on 1 year testing data as compared to BPNN in terms of the RMSE.

Obviously, the prediction accuracy is dependent on the accuracy of hourly day-ahead weather forecasting variables issued from weather service organizations. Future work will focus on evaluating the error in solar irradiance forecasts due to errors in weather forecasts. Anyway, the study provides an alternative novelty way of applying the advanced machine learning algorithms to the solar irradiance prediction.

Acknowledgments

The work was supported by National Natural Science Foundation (NNSF) from China (61673174). The author would like to thank Truewin Renewables Technology (Shanghai) Co., Ltd., for financial backing in field investigating the Santiago solar plant and Electra Company for providing solar irradiance data. Also thanks to Dr. Huanji Xie, at the Truewin Renewables Technology (Shanghai) Co., Ltd. and O.C.Nogueira at the Electra Company in Cape Verde.

References

[1] Keerthisinghe C, Verbic G, Chapman AC. A fast technique for smart home management: ADP with temporal difference learning. In: IEEE Transactions on smart grid; 2017 [online published].
[2] Aggarwal SK, Saini LM. Solar energy prediction using linear and non-linear regularization models: a study on AMS (American Meteorological Society) 2013-14 solar energy prediction contest. Energy 2014;78:247–56.
[3] Cao JC, Cao SH. Study of forecasting solar irradiance using neural networks with preprocessing sample data by wavelet analysis. Energy 2006;31:3435–45.
[4] Yand D, Panida J, Wilfred MW. Hourly solar irradiance time series forecasting using cloud cover index. Sol Energy 2012;86:3531–43.
[5] Voyant C, Muselli M, Paoli C, Nivet M. Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. Energy 2012;39:341–55.
[6] Ahmad A, Anderson TN, Lie TT. Hourly global solar irradiation forecasting for New Zealand. Sol Energy 2015;122:1398–408.
[7] Sharma V, Yang S, Walsh W, Reindl T. Short term solar irradiance forecasting using a mixed wavelet neural network. Renew Energy 2016;90:481–92.
[8] Hocaoglu FO, Serttas F. A novel hybrid (Mycielski-Markov) model for hourly solar radiation forecasting. Renew Energy 2017;108:635–43.
[9] Monjoly S, Andre M, Calif R, Soubdhan T. Hourly forecasting of global solar radiation based on multiscale decomposition methods: a hybrid approach. Energy 2017;119:288–98.
[10] Voyant C, Notton G, Kalogirou S, et al. Machine learning methods for solar radiation forecasting: a review. Renew Energy 2017;105:569–82.
[11] Sharma N, Sharma P, Irwin D, Shenoy P. Predicting solar generation from weather forecasts using machine learning. In: 2011 IEEE International Conference on smart grid Communications (IEEE SmartGridComm); 2011. p. 528–33.
[12] Bae KY, Jang HS, Sung DK. Hourly solar irradiance prediction based on support vector machine and its error analysis. IEEE Trans Power Syst 2017;32(2):935–45.
[13] Ceci M, Corizzo R, Fumarola Malerba D, Rashkovska A. Predictive modeling of PV energy production: how to set up learning task for a better prediction? IEEE Trans Ind Inform 2017;13(3):956–66.
[14] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems; 2014. p. 3104–12.
[15] Schmidhuber J. Deep learning in neural networks: an overview. Neural Network 2015;61:85–117.
[16] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[17] Shakya A, Michael S, et al. Solar irradiance forecasting in remote microgrids using Markov switching model. IEEE Trans Sustain Energy 2017;8(3):

895–905.
[18] https://www.esi-africa.com/tenders/cape-verde-deploy-solar-energy-system-project/.
[19] Qing X. Statistical analysis of wind energy characteristics in Santiago Island, Cape Verde. Renew Energy 2018;115:448–61.
[20] Zhao Z, Chen W, et al. LSTM network: a deep learning approach for short-term traffic forecast. IET Intell Transp Syst 2017;11(2):68–75.
[21] Kong W, Dong ZY, et al. Short-term residential load forecasting based on resident behaviour learning. In: IEEE Transactions on power systems; 2017 [online published].
[22] Fischer T, Krauss C. Deep learning with long-term memory networks for financial market predictions. Eur J Oper Res 2018. FAU Discussion Papers in Economics, No.11/2017.
[23] https://keras.io/.