Renewable Energy 66 (2014) 78–89

Contents lists available at ScienceDirect

Renewable Energy

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

An ARMAX model for forecasting the power output of a grid connected photovoltaic system

Yanting Li [a], Yan Su [b], Lianjie Shu [c,*]

[a] Department of Industrial Engineering and Logistics Management, Shanghai Jiao Tong University, Shanghai, China
[b] Department of Electromechanical Engineering, University of Macau, Macau, PR China
[c] Faculty of Business, University of Macau, Macau, PR China

ARTICLE INFO

Article history:
Received 27 June 2013
Accepted 25 November 2013
Available online

Keywords:
Grid connection
Solar irradiance
Photovoltaic system
Efficiency

ABSTRACT

Power forecasting has received a great deal of attention due to its importance for planning the operations of photovoltaic (PV) system. Compared to other forecasting techniques, the ARIMA time series model does not require the meteorological forecast of solar irradiance that is often complicated. Due to its simplicity, the ARIMA model has been widely discussed as a statistical model for forecasting power output from a PV system. However, the ARIMA model is a data-driven model that cannot take the climatic information into account. Intuitively, such information is valuable for improving the forecast accuracy. Motivated by this, this paper suggests a generalized model, the ARMAX model, to allow for exogenous inputs for forecasting power output. The suggested model takes temperature, precipitation amount, insolation duration, and humidity that can be easily accessed from the local observatory as exogenous inputs. As the ARMAX model does not rely forecast on solar irradiance, it maintains simplicity as the conventional ARIMA model. On the other hand, it is more general and flexible for practical use than the ARIMA model. It is shown that the ARMAX model greatly improves the forecast accuracy of power output over the ARIMA model. The results were validated based on a grid-connected 2.1 kW PV system.

1. Introduction

Recently, photovoltaic (PV) technology has been rapidly developed due to the advantages of solar energy being abundant, inexhaustible, and clean [1]. In addition to the wide establishment in remote areas, PV systems are also becoming popular in grid-connected applications with the development and advancement in PV technology [2]. Unlike the traditional power plant that the power output can be easily controlled, the power output of PV system exhibits greatly variability. For PV systems, there are many factors which can influence the power output character such as solar irradiation, system transfer efficiency, installation angle, and temperature. Due to the variability of solar irradiation and environmental factors, the power output of PV system is a stochastic random process.

The variability of power output not only adversely affects the stability of the electrical system being connected but also affects both capital and operational costs [3]. In order to improve the integration stability of output of a solar PV system into electric grid and optimize decision making at management of local storage systems and bidding into electricity markets, the necessity to have forecasting models is increasing. Forecasting energy production can help producers to implement operation strategies in an efficient way and achieve better management. Short term like intra-hour forecasts are relevant for dispatching, regulatory and load following purpose [4] while 1-day ahead forecasts are critical for operational planning of transmission system operator and for trading in electricity markets for PV power system operators [5].

Many researchers have contributed to the development of forecasting tools for accurate prediction of PV power output. Most of the previous research on this problem have employed a two-stage approach. In the first stage, the solar irradiance on different time scales is forecasted, and then the forecasted irradiance and temperature data are used as inputs in commercial PV simulation softwares such as TRNSYSM [6], PVFORM [7], and HOMER [8]. Many attempts to forecast solar irradiance have been presented, which can be generally classified into two categories: time series models and Neural Network (NN) based models. For example, Martín et al. [9] employed time series models to predict global solar irradiance. Reikard [10] compared the time series forecasts of solar radiation at high resolution. Besides, in some other research work, the

* Corresponding author. Tel.: +853 8397 4741; fax: +853 2883 8320.
E-mail address: ljshu@umac.mo (L. Shu).

0960-1481/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.renene.2013.11.067

Nomenclature

t	time in day
Y_t	average daily power in day t
$\widehat{Y}_{t+1}$	one-day ahead forecast of Y_t
$\widehat{Y}_{t+n}$	n-step ahead forecast of Y_t
E_t	estimated base level at time t
T_t	estimated trend level at time t
S_t	estimated seasonality at time t
s	number of seasonal periods per year
$\theta(B)$	MA polynomial
$\phi(B)$	AR polynomial
p	AR order
d	difference order
q	MA order
$d_{1,t}$	daily average temperature (°C)
$d_{2,t}$	daily highest temperature
$d_{3,t}$	daily lowest temperature
$d_{4,t}$	daily dew temperature
$d_{5,t}$	daily wind speed (m/s)
$d_{6,t}$	daily wind direction
$d_{7,t}$	daily precipitation amount (mm)
$d_{8,t}$	daily insolation duration (Hours)
$d_{9,t}$	daily humidity
$d_{10,t}$	daily air pressure
k	number of parameters estimated
N	total number of observations
RMSE	root mean square error
MAD	mean absolute deviation
MAPE	mean absolute percent error

transformed time series data were used to forecast daily global solar irradiance [11,12]. A sample of research on the use of NN-based models to forecast solar irradiance is listed below. Mohandes et al. [13] introduced the artificial NN models to estimate global solar irradiance. Sfetsos and Coonick [14] proposed using various artificial intelligence based techniques for forecasting hourly solar radiation. Hontoria et al. [15] used NN multilayer perception to generate hourly irradiation. Cao and Cao [16] developed a hybrid model that combines artificial NN with wavelet analysis to forecast total daily solar radiation. Hocaoglu et al. [17] used a feed-forward NN model for hourly solar radiation forecasting. Recently, Paoli et al. [18] applied the NN models to forecast the preprocessed daily solar radiation time series. Similarly, Mellit and Pavan [19] applied the NN models to perform a 24 h ahead forecast of solar irradiance based on a grid connected PV plant in Italy.

Instead of using a two-stage approach, another appropriate strategy could be directly forecasting the power output based on some prior information or readily accessed data. Due to the similarity of forecasting solar irradiance and power output, some researchers have extended the time series and NN models for forecasting solar irradiance to the forecast of power output of PV systems. A sample of research on the use of time series models for direct forecasting power output includes [4,20,21] while a sample of research on the use of NN models for this purpose includes [22–25].

The NN models use the historical observed data and some meteorological data to construct the forecast. The design of NN models often involves the design of network architecture and the selection of a good learning algorithm. However, this heavily relies on past experience and is subject to trial and error processes. The time series method is a data-driven method, assuming that the data have an internal structure that can be identified by using simple and partial autocorrelation. Compared to the NN methods, time series forecasting models involve only a few model parameters and have no requirements of reliable knowledge and past experience. Therefore, the time series model is much simpler for forecasting than the NN models. The widely used time series models include the auto-regressive (AR) models, moving average (MA) models, and their generalizations such as the auto-regressive moving average (ARMA) and the auto-regressive integrated moving average (ARIMA) models also known as Box–Jenkins models [26].

Although the ARIMA model has now become the most general class of models for forecasting a time series, it cannot take the process behavior into consideration. To allow for exogenous inputs, the ARMAX model can be used, which has been proved to be a powerful tool in time series forecasting [27]. Parkhurst [28] refers to the ARMAX model as dynamic regression. The ARMAX is a generalization of the ARIMA model and is thus more flexible for practical use. However, the ARMAX model has been less studied for forecasting power output although it has been successfully employed in other applications. The only exception is Bacher et al. [20]. They showed that the ARX model with numerical weather conditions (NWPs) as inputs performs much better than the AR model in forecasting short-term (2-h ahead) power output.

The objective of this paper is to suggest an ARMAX model to forecast the 1-day ahead power output of PV systems for the purpose of better planning and trading in the electricity market. The prediction performance will be compared with the ARIMA model and other time series models. In this work, the daily average data for a 2.1 kW grid connected PV output collected from January 1, 2011 to June 30, 2012 was used to train and validate the time series models.

2. Time series models

2.1. ARIMA models

In the ARIMA model, lags of the differenced series appearing in the forecasting equation are called AR terms, lags of the forecast errors are called MA terms, and a time series which needs to be differenced to be made stationary is said to be an “integrated” version of a stationary series. An ARIMA(p,d,q) model of the nonstationary random process Y_t is expressed as

$$(1-B)^d Y_t = \mu + \frac{\theta(B)}{\phi(B)} a_t \tag{1}$$

where B is the lag operator defined such that $BY_t = Y_{t-1}$; $\{\phi_i\}$ are the AR coefficients; $\{\theta_i\}$ are the MA coefficients; μ denotes the process mean; a_t is a white noise that is generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean.

The value of p denotes the AR order, d is the number of nonseasonal differences, and q denotes the MA order. In the case of $d = 0$, the ARIMA(p,d,q) model is reduced to the an ARMA(p,q) model [26], which is a class of stationary models. The ARMA(p,q) model is further reduced to an AR(p) model when $q = 0$, and an MA(q) model when $p = 0$.

The Box–Jenkins' procedure for constructing ARIMA models involves iterative three steps: identification, estimation, and diagnostic checking. In the identification phase, one can calculate the sample autocorrelation coefficient (ACC) and partial autocorrelation coefficient (PACC) to determine the necessity and degree of differencing, and then identify the orders p and q of the ARIMA models based on the properly transformed time series. The coefficients of the ARIMA model can be estimated by the Yule-Walker estimator, the least squares estimator, or the maximum likelihood estimator [29]. After the model parameters were estimated, the goodness-of-fit of the model is examined. The time series residuals must met the white noise assumptions. There are some diagnostic checks to check these assumptions. If the fitted model passes the diagnostic checking, the model can be used to make forecast. Otherwise, the tentative model is unacceptable and the above three-step procedure should be repeated until an adequate model is obtained.

2.2. ARMAX models

It is intuitively appealing to include useful covariates into the ARIMA model as these external covariates can consider process behaviors and thus improve the forecasting accuracy of the ARIMA models. The ARMAX model with exogenous inputs are expressed as follows [28]:

$$Y_t = \beta_0 + \sum_{i=1}^{m} \frac{\omega_i(B)}{\delta_i(B)} B^{k_i} X_{i,t} + N_t, \tag{2}$$

where Y_t is the output series, X_{it} is the ith input time series or a difference of the ith input series at time t, m is the total number of external covariates, k_i is the pure time delay for the effect of the ith input series, $\omega_i(B)$ is the numerator polynomial of the transfer function for the ith input series, $\delta_i(B)$ the denominator polynomial of the transfer function for the ith input series, and N_t denotes the stochastic disturbance in the form of an ARMA model

$$N_t = \frac{\theta(B)}{\phi(B)(1-B)^d} a_t.$$

The construction of an ARMAX model is an iterative process similar to the construction of ARIMA models, i.e., identification, estimation, and diagnostic checking. After checking both input and output series are stationary, we can start with the linear transfer function (LTF) method to determine the rational form transfer functions, $\omega_i(B)$ and $\delta_i(B)$. First, we specify free-form distributed lag model in which the process variables are included. Then one can follow the Box–Jenkins' procedure to determine the form of the disturbance series N_t based on the plot of ACC and PACC. If the disturbance is not stationary, then it is necessary to difference Y_t and the inputs accordingly. If the disturbance is stationary, then we can go to another stage where we may use the preliminary estimated transfer functions and the tentative ARMA disturbance model. For parameter estimation, there are a variety of estimation methods. Parameters can be estimated based on conditional likelihood function, following Box and Jenkins [26], this involves choosing coefficients that minimize the sum of the squared residuals. But as pointed out by Alan [28] this method can lead to badly biased estimate of MA coefficients, especially when these coefficients are near the invertibility boundary. Therefore, in this paper, we use maximum-likelihood estimates. There are also several diagnostic checks to decide whether the ARMAX model is adequate based on the residuals which should be independent as well as input series. The flow chart in Fig. 1 provides the detailed steps of the LTF method for the construction of the ARMAX models.

In addition to the ARIMA and ARMAX models, some other models based on the decomposition methodology have been also proposed. In general, any time-series data can be decomposed into three basic components: trend, seasonal effect, and irregularity. The trend component tends to make the time series goes upwards or downwards in the long run while the seasonal effect follows the same pattern but repeats in a systematic interval over time. To model different components in time series, different methods have been proposed. In summary, these methods generally fall into the following categories: moving averaging methods and exponential smoothing methods. See Ref. [26] for more details.

2.3. Moving average

2.3.1. Simple moving average

Simple averaging methods are suitable for stationary time series data where the series is in equilibrium around a constant value (the underlying mean) with a constant variance over time. The simple moving average uses the average of the historical data as the forecast. A moving average model with order m predicts the value of Y_{t+1} based on

$$\widehat{Y}_{t+1} = \frac{1}{m} \sum_{i=t-m+1}^{t} Y_i, \tag{3}$$

where $\widehat{Y}_{t+1}$ represents the 1-step ahead forecast. This method is appropriate when there is no noticeable trend or seasonality.

2.3.2. Double moving average

In presence of trend, in order to compensate the drawbacks of simple moving average, double moving average method calculates a second moving average from the simple moving average, and uses these two averages to compute the slope (T_t) and intercept (E_t). The n-step ahead double moving average forecasting function is then given by

$$\widehat{Y}_{t+n} = E_t + nT_t, \tag{4}$$

where

$$\begin{aligned} E_t &= 2M_t^1 - M_t^2 \\ T_t &= \tfrac{2}{k-1}\left(M_t^1 - M_t^2\right) \end{aligned} \tag{5}$$

and

$$\begin{aligned} M_t^1 &= (Y_t + Y_{t-1} + \cdots + Y_{t-k+1})/k \\ M_t^2 &= \left(M_t^1 + M_{t-1}^1 \cdots + M_{t-k+1}^1\right)\Big/k. \end{aligned} \tag{6}$$

The values of E_t and T_t represent the estimated level of the time series and the trend at time t, respectively.

2.4. Exponential smoothing

Exponential smoothing was first suggested by Brown [30], and then expanded by Holt [31] and Winter [32]. Instead of imposing equal weights to the past data, this method assigns unequal set of weights to past data, where the weights decay exponentially from the most recent to the most distant data points.

2.4.1. Simple exponential smoothing

This simple form of exponential smoothing is also known as an exponentially weighted moving average (EWMA). The EWMA model assumes the following form

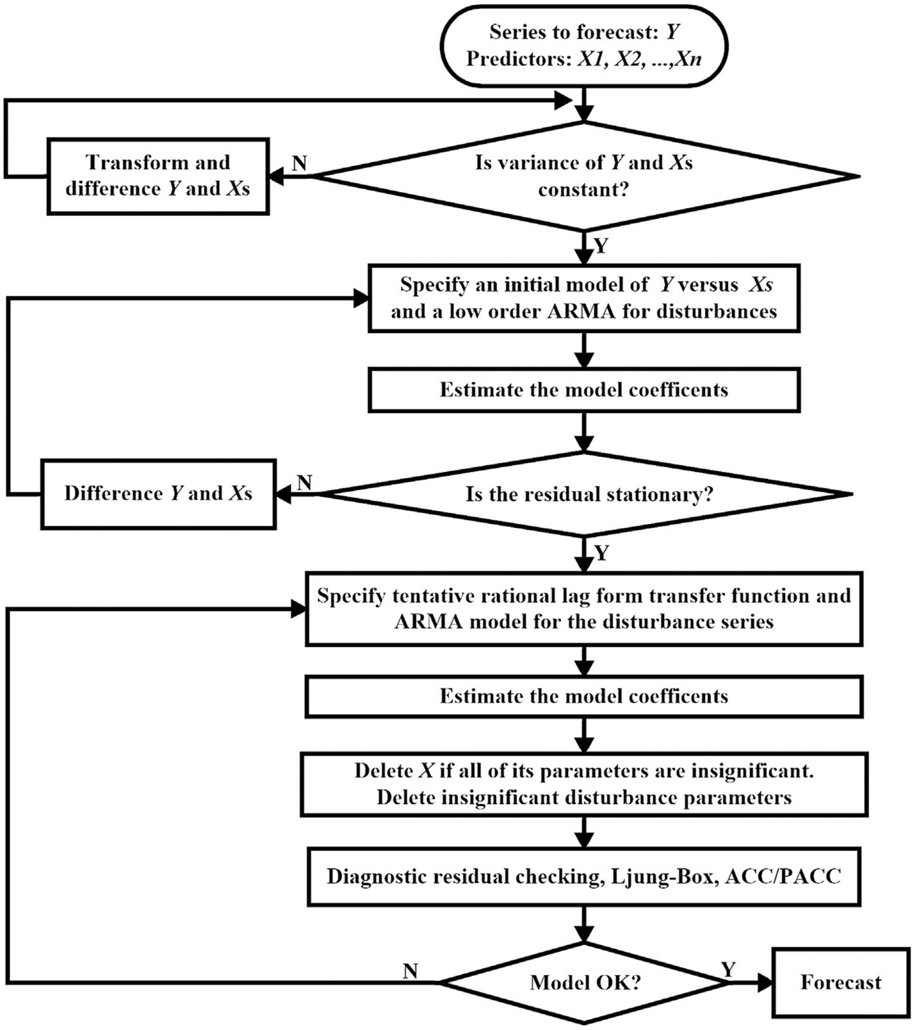
Series to forecast: Y
Predictors: X1, X2, ...,Xn
Is variance of Y and Xs constant?
N
Transform and difference Y and Xs
Y
Specify an initial model of Y versus Xs and a low order ARMA for disturbances
Estimate the model coefficents
Is the residual stationary?
N
Difference Y and Xs
Y
Specify tentative rational lag form transfer function and ARMA model for the disturbance series
Estimate the model coefficents
Delete X if all of its parameters are insignificant. Delete insignificant disturbance parameters
Diagnostic residual checking, Ljung-Box, ACC/PACC
Model OK?
N
Y
Forecast

Fig. 1. The linear transfer function (LTF) method.

$$\widehat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\widehat{Y}_t = \widehat{Y}_t + \alpha\left(Y_t - \widehat{Y}_t\right) \tag{7}$$

where the parameter α can assume any value between 0 and 1. The EWMA forecast indicates that the predicted value for time period $t+1$, $\widehat{Y}_{t+1}$, is equal to the last predicated value plus an adjustment for the error made in predicting $\widehat{Y}_t$, $\alpha(Y_t - \widehat{Y}_t)$. The initial forecast $\widehat{Y}_t$ is often set to $\widehat{Y}_t = Y_1$. Similar to the simple moving average method, the EWMA method is appropriate for a stationary without trend time series. The EWMA forecast would be lagging behind the trend if one exists.

2.4.2. Holt's method

Since simple exponential smoothing does not do well when there is a trend in the data, Holt [31] extended it to two parameter exponential smoothing method by adding a growth factor (or trend factor) to the smoothing equation as a way of adjusting for the trend. The forecasting function in Holt's method is represented by

$$\widehat{Y}_{t+n} = E_t + nT_t, \tag{8}$$

where

$$\begin{aligned} E_t &= \alpha Y_t + (1-\alpha)(E_{t-1} + T_{t-1}) \\ T_t &= \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1}. \end{aligned} \tag{9}$$

Once again, E_t and T_t represent the estimated level of the time series and the expected rate of increase or decrease (the trend) at time t, respectively. The smoothing parameters assume any value between 0 and 1, i.e., $0 \le \alpha \le 1$ and $0 \le \beta \le 1$.

2.4.3. Holt–Winter's model

In addition to having an upward or downward trend, nonstationary data may also exhibit seasonal effects. Holt–Winter's method is a further generalization of Holt's method, which can be applied to time series exhibiting trend and seasonality. The seasonal effects may be additive or multiplicative in nature. In presence of additive seasonality, the time series shows steady seasonal fluctuations, regardless of the overall level of the time series. On the other hand, in presence of multiplicative seasonality, the size of the seasonal fluctuations varies, depending on the overall level of the time series.

Define s as the number of seasonal periods within one year in the time series ($s = 12$ for monthly data, and $s = 4$ for quarterly data). The forecasting function based on Holt–Winter's additive method is given by

$$\widehat{Y}_{t+n} = E_t + nT_t + S_{t+n-s}, \tag{10}$$

where the expected base level E_t, trend T_t, and seasonality S_t of the time series in period t are given by

$$\begin{aligned} E_t &= \alpha(Y_t - S_{t-s}) + (1-\alpha)(E_{t-1} + T_{t-1}) \\ T_t &= \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1} \\ S_t &= \gamma(Y_t - E_t) + (1-\gamma)S_{t-s}. \end{aligned} \tag{11}$$

Holt–Winter's method for multiplicative seasonal effects is very similar to that for additive seasonal effects. The forecasting equation is given by

$$\widehat{Y}_{t+n} = (E_t + nT_t) \times S_{t+n-s}, \tag{12}$$

where the expected base level E_t, trend T_t, and seasonality S_t of the time series in period t are given by

$$\begin{aligned} E_t &= \alpha(Y_t / S_{t-s}) + (1-\alpha)(E_{t-1} + T_{t-1}) \\ T_t &= \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1} \\ S_t &= \gamma(Y_t / E_t) + (1-\gamma)S_{t-s}. \end{aligned} \tag{13}$$

3. Model identification

3.1. Data set

The historical power data were obtained from a grid-connected PV system located in a low density area, the Coloane island of Macau Special Administrative Region (SAR) (latitude = 22°10′0″ N and longitude = 113°33′0″ E). The PV system is mounted on the rooftop of an institutional building, with PV modules installed of an inclination angle of 10° and facing south-east. The underlying PV system consists of twelve Kyocera HTS-175 with an installed capacity of 2.1 kW. Fig. 2 shows the installation of the underlying PV system.

Under standard test conditions (i.e., cell temperature = 25 °C), the maximum output power of the HTS-175 module was 175W ± 5%. A single Xantrex™ grid tie (GT series) solar inverter was used to convert direct current (DC) to alternating current (AC). Under AC output voltage of 240 V, this inverter can provide maximum output current of 11.7 A. The maximum AC power output is 2.8 kW for the electrical input. This inverter has MPPT operating range from 193 V dc to 550 V dc with the maximum input current of 15.4 A dc. The maximum inverter efficiency is 95%. Power generated from the PV system was fed directly into the grid.

Fig. 2. The installation of the underlying PV system.

The recorded data in the underlying PV system include array power output and global horizontal irradiance (GHI). They were monitored at 1-min sampling intervals. The solar data recorded during the period from January 1, 2011 to June 30, 2012 were gathered for the present study. The time series Y_t, average daily output power, is constructed from the minutely readings of array output power. It is inevitable that there were some missing data for various reasons such as instrument malfunction and maintenance. Out of the 546 daily records in the present data set, there are 13 missing records due to maintenance. In order to enable comprehensive analysis and forecasting, we used the simple linear interpolation method to estimate the missing observations while other methods can also be used. Fore more details on the methods estimating missing values in time series, interested readers may refer to Brockwell and Davis [33]. In what follows, we use the data from January 1 to December 31 2011 for model fitting and the remaining data in year 2012 for model validation.

Note that the GHI is not constant, depending on the weather conditions and the temperature changes as well. Thus, it is very difficult to forecast GHI accurately. For this reason, we don't use the GHI measurements for forecasting the daily power output. Rather, we use data from weather forecasting in the next day as inputs of the ARMAX model. In particular, we take the following types of climatic variables into consideration: (1) daily average temperature (d_1), dew temperature (d_2), highest temperature (d_3) and lowest temperature (d_4); (2) wind speed (d_5) and wind direction (d_6); (3) precipitation amount (d_7); (4) insolation duration (d_8); (5) humidity (d_9); and (6) air pressure (d_{10}). The reason for choosing these variables as candidate inputs is due to their easy accessibility. Note that information about these variables can be easily forecasted and can be obtained from the local observatory.

3.2. Developed models

Fig. 3 depicts the mean daily output power Y_t generated from the PV system during the period from January 1 to December 31, 2011. Fig. 4 plots the corresponding histogram, QQ plot, sample ACC and PACC of Y_t. From Fig. 4, the histogram of Y_t is not symmetrical. Also, the data on the QQ-plot does not fit a straight line. Clearly, these observations indicate that the measured time series is not normal. The sample ACC decays slowly. Note that the PACC has a significant spike only at lag 1, suggesting an autoregressive model of order 1.

When fitting the ARIMA or ARMAX models, it is possible to increase the likelihood by adding more parameters or increasing the orders of the transfer function. However, including some unnecessary parameters in the model can cause the overfitting issue. To resolve this issue, the Bayesian information criterion (BIC) can be used as selection criteria for time series models. The BIC of ARIMA family models is defined as [34]

$$\mathrm{BIC}(k) = -2\ \ln L + k \ln(N), \tag{14}$$

where k is the number of parameters to be estimated in the model, L is the maximized value of the likelihood function for the estimated model, and N is the total number of observations. The optimal order of the model is chosen by the value of k to minimize BIC(k).

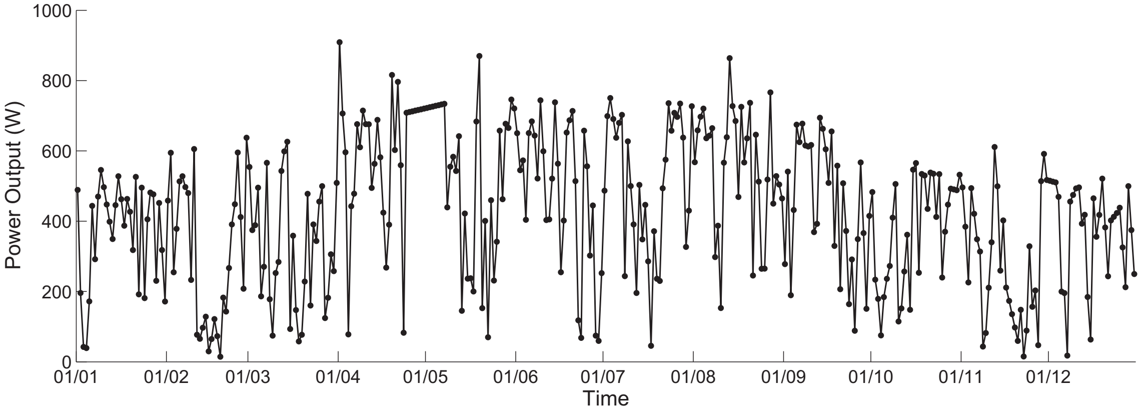

Fig. 3. Mean daily power output in 2011.

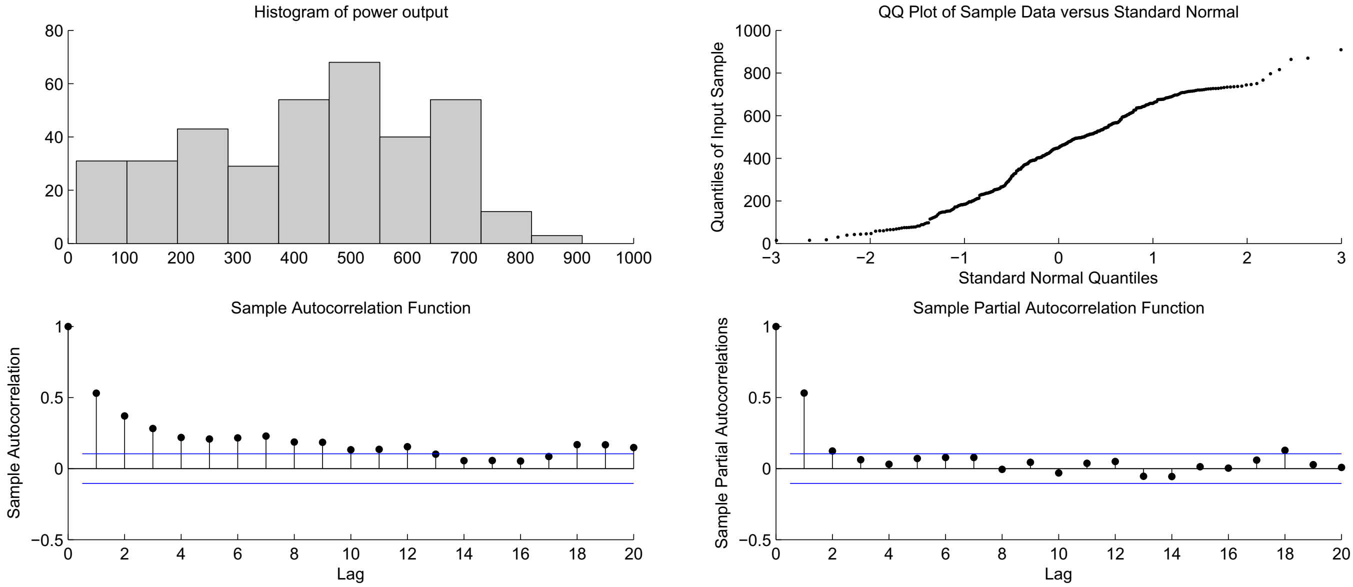

Fig. 4. Histogram, QQ-plot, the sample autocorrelation coefficient (ACC) and partial autocorrelation coefficient (PACC) of the measured mean daily output power in 2011.

Table 1
Parameter estimates of different models.

Model	Parameter	Value
ARMAX	Intercept	237.565
	AR (Lag 1)	0.426
	MA (Lag 1)	−0.153
	Daily Avg. Temp.(d_1)	8.908
	Daily precipitation amount (d_7)	−1.557
	Daily insolation duration (d_8)	31.919
	Daily humidity (d_9)	−2.045
ARIMA(1,1,1)	AR(1)	0.4221
	MA(1)	0.9490
Double moving average	Order	100
Single exponential smoothing	α	0.3705
Double exponential smoothing	α	0.3711
	β	0.0010
Holt–Winter's additive	α	0.3103
	β	0.0010
	γ	0.0825
Holt–Winter's multiplicative	α	0.2516
	β	0.0010
	γ	0.0949
Single moving average	Order	9

Table 1 lists the parameter estimates of the best model in each category. From Table 1, the ARIMA model gives the smallest BIC value is the ARIMA(1,1,1). The best ARMAX model fitted is the ARMAX model given by

$$\begin{aligned} Y_t = {} & 237.565 + 0.426Y_{t-1} + a_t - 0.153a_{t-1} + 8.9087d_{1,t} \\ & - 1.557d_{7,t} + 31.919d_{8,t} - 2.045d_{9,t}, \end{aligned} \tag{15}$$

where a_t represents random noise. For time series models based on moving average exponential smoothing techniques, the maximum likelihood ratio can be used for estimating the optimal parameters. The parameter estimates were obtained based on the software *SPSS 18*.

To compare the forecast accuracy of the above time series models, Table 2 presents the summary measures of performance based on the mean square errors (RMSE), mean absolute deviation (MAD), and mean absolute percent error (MAPE) applied to the training data in 2011. The RMSE, MAD and MAPE are calculated with

Table 2
Performance comparison of different models for training data in 2011.

Methods	Rank	RMSE	MAD	MAPE (%)
ARMAX	1	104.77	77.27	38.88
ARIMA(1,1,1)	2	172.96	140.9	76.66
Double moving average	3	180.25	152.0	88.10
Single exponential smoothing	4	180.95	141.5	72.93
Double exponential smoothing	5	181.04	141.5	72.85
Holt–Winter's additive	6	185.10	144.6	72.36
Holt–Winter's multiplicative	7	185.43	146.5	75.94
Single moving average	8	190.59	153.8	82.09

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{N}\left(Y_t - \widehat{Y}_t\right)^2}{N}},$$

$$\mathrm{MAD} = \frac{\sum_{t=1}^{N}\left|Y_t - \widehat{Y}_t\right|}{N},$$

and

$$\mathrm{MAPE} = \frac{\sum_{t=1}^{N}\left|Y_t - \widehat{Y}_t\right|/Y_t}{N}$$

respectively. The models are ranked in the order of RMSE.

From Table 2, the ARMAX model provides the best prediction performance in the sense that it gives the lowest RMSE, MAD, and MAPE. The ARIMA(1,1,1) performs the second best. The time series models based on moving average, exponentially smoothing and Holt–Winter's models, due to the high volatility of the power output data, perform unsatisfactorily. The RMSE of these models is more than 180. Comparing ARMAX with ARIMA, it is obvious that the prediction accuracy can be greatly improved when the easily accessible weather information was included in the ARIMA model for predicting the output power of PV system. For example, the RMSE of the ARMAX model is 104.77, which is 60% of that of the ARIMA(1,1,1) model. The MAPE of the ARMAX model is only 38.88% while that of the ARIMA(1,1,1) model is as large as 76.66%.

The parameter estimates of the ARMAX model show that the amount of precipitation and relative humidity have negative effect on the power output while the insolation duration and average air temperature have positive correlation with the power output. To help us understand this, we further display the variational features of power output of the PV system relative to these climatic variables. For illustration, the data based on observations in the summer month from August 1 to August 31, 2011 were used in plotting Figs. 5–10.

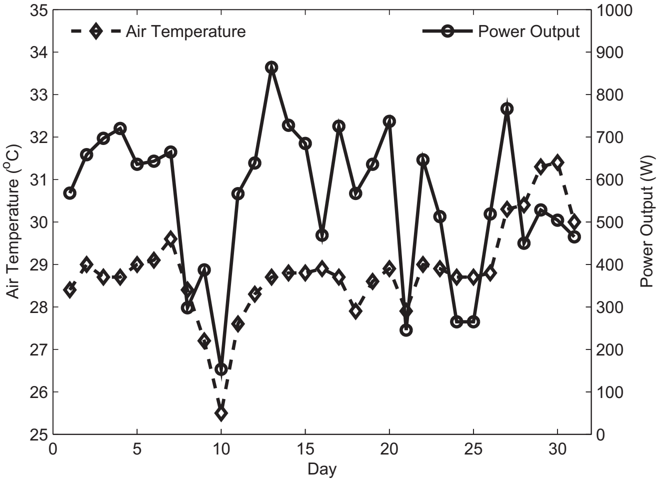

Fig. 6. The curves of mean daily power output of PV system and average daily temperature in August 2011.

Figs. 5 and 6 visually display the variational patterns of the mean daily power output relative to insolation and average air temperature in August 2011, respectively. It can be seen from Fig. 5 that the change pattern of power output is basically similar to that of insolation. The peak (or valley) points on the curve of power output seem to coincide with the peak (or valley) points on the insolation curve. Also, as can be seen from Fig. 6, the changing patterns of power output and average daily air temperature are basically similar in shape. The coincidence implies a positive correlation between the power output and insolation and average air temperature.

Figs. 7 and 8 present the variational patterns of the mean daily power output relative to precipitation and relative humidity, respectively. It is found that the pattern of power output of PV system seems to be changing in an opposite way to the patterns of precipitation and relative humidity. The valley points on the curve

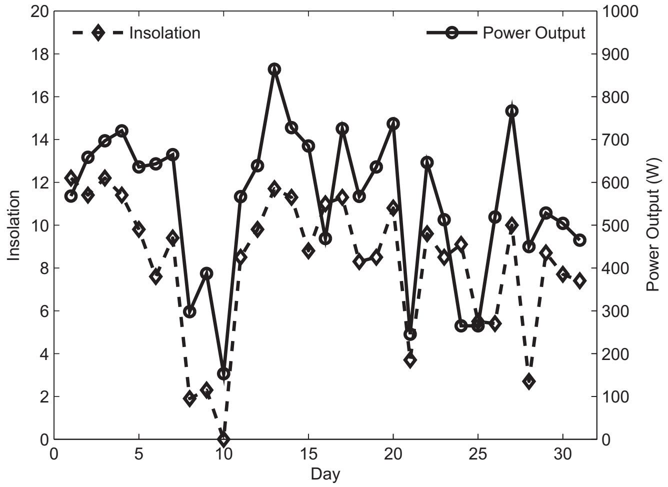

Fig. 5. The curves of mean daily power output of PV system and insolation in August 2011.

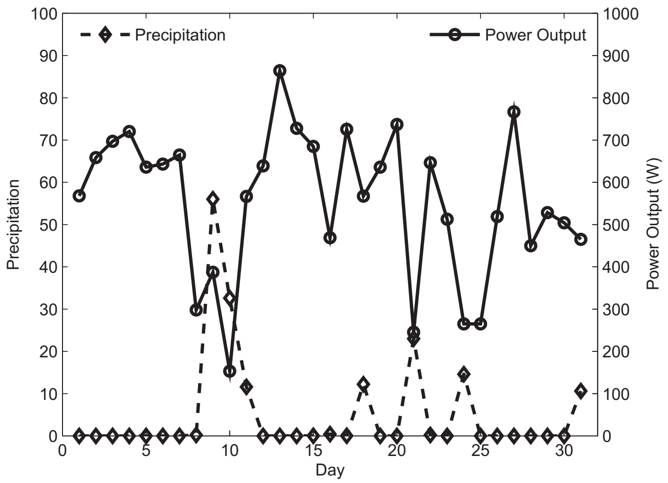

Fig. 7. The curves of mean daily power output of PV system and precipitation in August 2011.

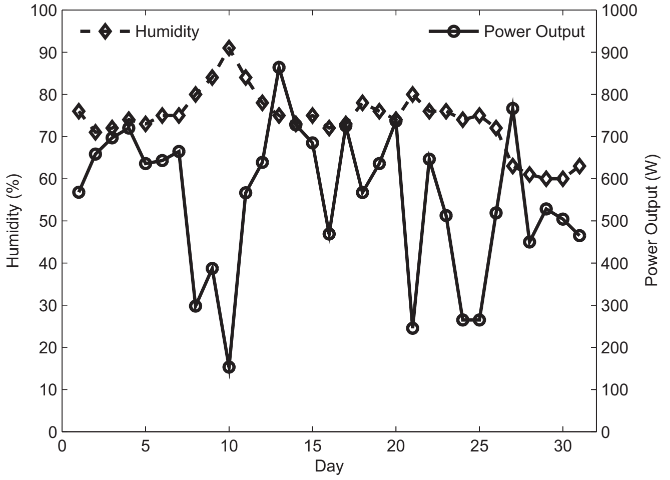

Fig. 8. The curves of mean daily power output of PV system and humidity in August 2011.

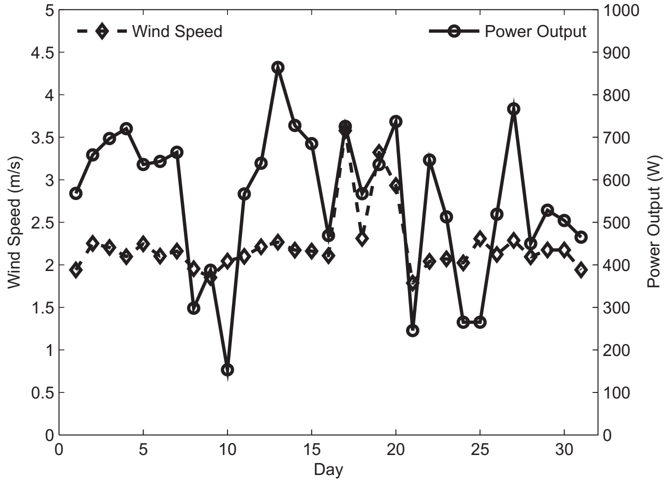

Fig. 10. The curves of mean daily power output of PV system and wind speed in August 2011.

of power output tend to coincide with the peak points on the curves of precipitation and relative humidity. Therefore, a negative correlation between the power output and precipitation and relative humidity is expected.

Fig. 9 further plots the variational features of the mean daily power output and GHI measured in August, 2011. As expected, the power output changes with solar irradiation in nearly the same way. The peak and valley points of both curves are essentially coincident. This observation clearly illustrates that the power output of PV system strongly depends on solar irradiance. However, the meteorological forecast of solar irradiance in practice is often complicated to obtain. For this reason, we don't take it into account in the ARMAX forecasting model. Moreover, Fig. 10 presents the variational pattern of the power output relative to wind speed.

Table 3 also reports the correlation between the power output and the climatic variables excluding wind direction d_6 based on the data in year 2011. From Table 3, the power output has a positive correlation coefficient with daily insolation d_8 (0.74) and average air temperature d_1 (0.4) as expected. Also, the output power has a negative correlation with the amount of precipitation d_7 (-0.26) and relative humidity d_9 (-0.23) as expected. Moreover, the correlation coefficients among the four predictors d_1, d_7, d_8, d_9 have the absolute value less than 0.41, implying that these predictors included in the above ARMAX model are less dependent and multicollinearity is not present.

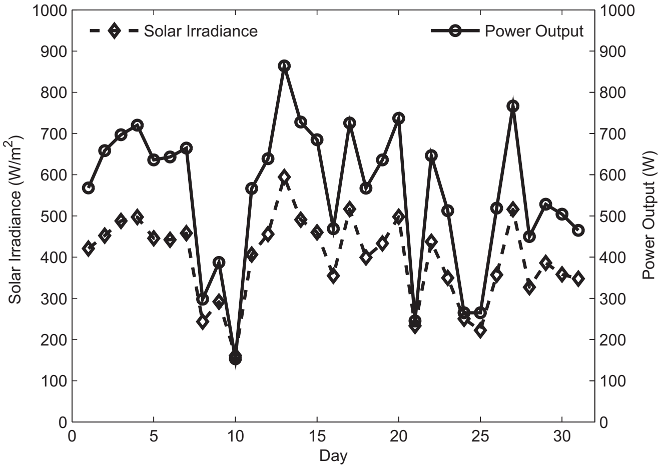

Fig. 9. The curves of mean daily power output of PV system and solar irradiation in August 2011.

To assess the adequacy of time series models, the model assumption that the residuals are white noise with zero mean and constant variance needs to be checked. Fig. 11 displays the histogram, QQ-plot, sample ACC and PACC of the residuals of the ARMAX model. The histogram is uni-modal and symmetric around zero. From the QQ-plot, the residuals approximately fit a straight line and thus can be assumed to be normal. The samples ACC and PACC do not exhibit any obvious patter and are generally within the critical limits. Therefore, the residuals can be assumed to be independent. To sum up, the white noise assumption of the residuals is found to be valid. This implies the ARMAX model fitted in Equation (15) is adequate.

4. Model validation

To further validate the above results, we compare the prediction performance of all the above models applied to the testing data set in terms of RMSE, MAD, and MAPE. The data from January 1 to June 30, 2012 were used for validation. Table 4 summarizes the prediction performance of all the models applied to the validation data set.

As can be seen from Table 4, once again the ARMAX model provides the best prediction performance when applied to the validation data. The RMSE, MAD and MAPE based on the ARMAX model are given by 125.84, 98.61 and 82.69%, respectively, which are much smaller than those based on the ARIMA(1,1,1) model. These observations are similar to those made in the model fitting phase.

Fig. 12 plots the measured and forecasted power output for the first six months in 2012. Due to the similarity between Holt–Winter's additive and Holt–Winter's multiplicative methods, the similarity between single moving average and double moving average methods, and the similarity between single exponential smoothing and double exponential smoothing, we only choose the one with slightly better forecasting performance out of each pair in the comparison. In particular, the following five forecast models were compared with the measured data, including the ARMAX, ARIMA, single moving average, double exponential smoothing, and Holt–Winter's additive models. As can be seen from Fig. 12 that the forecast of power output based on the ARMAX model seems to be the most close to the measured values on average among all the methods considered here, although it may not produce the smallest forecast error at each time point. This observation is especially obvious in June

Table 3
Correlation between the climatic variables and the power output.

	d_1	d_2	d_3	d_4	d_5	d_7	d_8	d_9	d_{10}	Y
d_1	1.00	0.95	0.98	0.99	−0.47	0.14	0.32	0.21	−0.86	0.40
d_2	0.95	1.00	0.90	0.95	−0.51	0.23	0.14	0.51	−0.88	0.28
d_3	0.98	0.90	1.00	0.95	−0.52	0.09	0.40	0.14	−0.82	0.39
d_4	0.99	0.95	0.95	1.00	−0.44	0.14	0.25	0.26	−0.85	0.35
d_5	−0.47	−0.51	−0.52	−0.44	1.00	0.02	−0.30	−0.29	0.41	−0.38
d_7	0.14	0.23	0.09	0.14	0.02	1.00	−0.25	0.34	−0.29	−0.26
d_8	0.32	0.14	0.40	0.25	−0.30	−0.25	1.00	−0.41	−0.14	0.74
d_9	0.21	0.51	0.14	0.26	−0.29	0.34	−0.41	1.00	−0.38	−0.23
d_{10}	−0.86	−0.88	−0.82	−0.85	0.41	−0.29	−0.14	−0.38	1.00	−0.25
Y	**0.40**	0.28	0.47	0.35	−0.38	**−0.26**	**0.74**	**−0.23**	−0.25	1.00

The significance level of the bold values 0.40, −0.26, 0.74, and −0.23 are given by 0.01, 0.047, 0.004, and 0.049, respectively.

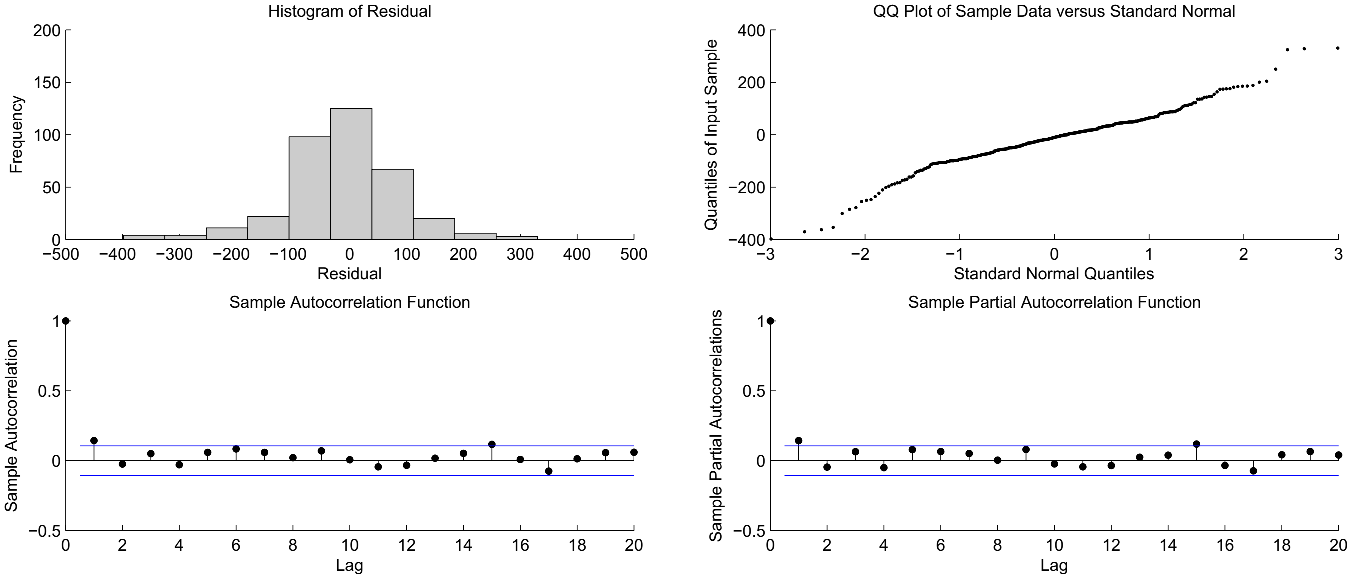

Fig. 11. Histogram, QQ-plot, sample ACC and PACC of the residuals of the ARMAX model.

2012. The 1-day ahead forecast of power output by using the ARMAX model for June of 2012 exhibits a pattern very similar to that of the measured data. The peak and valley points of the forecast curve almost coincide with those of measured power output.

5. Comparison with NN models

As suggested by an anonymous referee, we also compare the ARMAX model with the NN model for forecasting the 1-day ahead power output of the underlying PV system. In particular, we consider the radial basis function (RBF) network that commonly used for output power forecast. The RBF network model can learn a strong non-linear function faster and easily as compared to other NN models. It consists of three functionally distinct layers. The input layer is simply a set of sensory units. The second layer is a hidden layer of sufficient dimension, which performs a nonlinear transformation from the input space to a higher-dimensional hidden-unit space. The third layer performs a linear transformation from the hidden unit space to the output space.

The output of the RBF network is given by

$$\widehat{Y} = \sum_{i=1}^{m} \omega_i \phi(\cdot) \tag{16}$$

where m is the number of neurons in the hidden layer, ω_i is the weight between the hidden and output layer, and $\phi(\cdot)$ is the activation function in the hidden layer. Many functions are be defined as activation function. Among them, Gaussian soft-max networks have generalization ability more than regular RBF networks because of their extrapolation ability. Therefore in this paper, the function $\phi(\cdot)$ is defined as the soft-max function. Variables including daily average temperature, dew temperature, lowest temperature, wind speed and wind direction, precipitation amount, insolation duration, air pressure are treated as continuous input signals. The parameter estimates of NN were also obtained based on the software *SPSS 18*.

Table 5 summarizes the prediction performance based on the RBF network model for forecasting the 1-day ahead power output of the underlying PV system. From Table 5, it is clear that the RBF network model produces larger RMSE, MAD, and MAPE values than the ARMAX model when applied to the training and validation data. Fig. 13 compares the 1-day ahead forecasts based on the ARMAX model and RBF network during the period from January 1

Table 4
Prediction performance of different time series models for the validation data from January 1 to June 30 2012.

Methods	RMSE	MAD	MAPE (%)
ARMAX	125.84	98.61	82.69
ARIMA(1,1,1)	171.73	137.76	104.10
Double moving average	202.67	168.30	119.00
Single exponential smoothing	181.99	143.62	100.02
Double exponential smoothing	182.08	143.61	99.72
Holt–Winter's additive	184.32	144.43	100.69
Holt–Winter's multiplicative	186.57	148.03	106.22
Single moving average	196.22	164.07	127.05

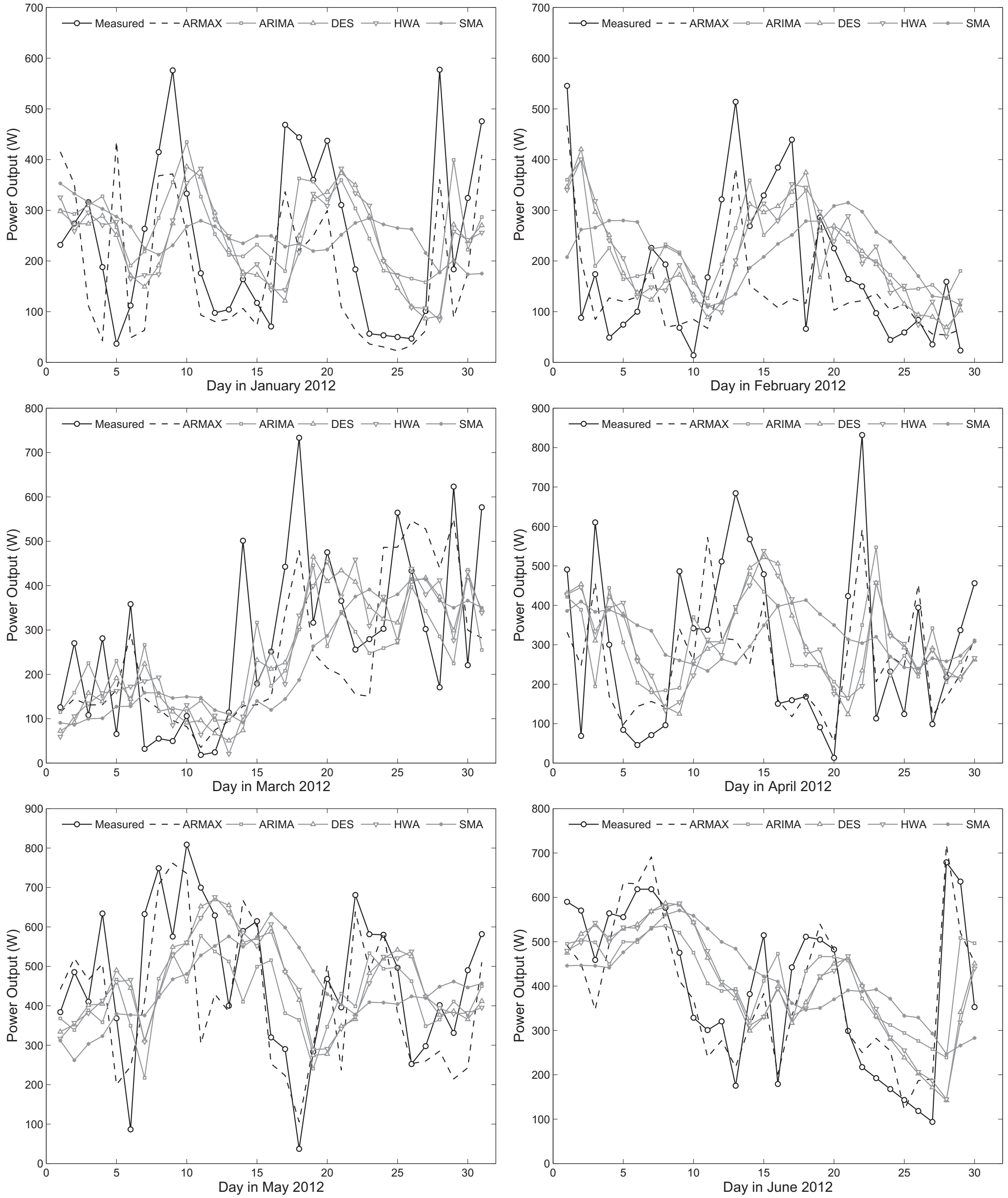

Fig. 12. Comparison between the 1-day ahead forecast and measured values of power output of the PV system during January to June 2012. DES: double exponential smoothing; HWA: Holt–Winter's additive model; SMA: single moving average.

to June 30 2012. As can be seen from Fig. 13, the forecast of power output based on the ARMAX model is closer to the measured values. This is especially obvious in March and June of 2012.

6. Conclusion

This paper assesses a wide variety of time series models for the 1-day ahead forecasting of the mean daily output power of a 2.1 kW grid connected PV system. These models consist of the models based on moving average techniques, models based on exponential smoothing techniques, ARIMA models, and ARMAX models. Compared to the ARIMA model, the ARMAX model can take some climatic variables into consideration in forecasting the output power. This is expected to improve the forecast accuracy of the ARIMA model. This paper considers some easily accessible climatic variables as exogenous inputs in the ARMAX model. In general, information about these climatic variables can be easily downloaded from the local observatory.

Table 5
Prediction performance of the RBF network model for the training and validation data.

Methods	Training data			Validation data		
	RMSE	MAD	MAPE (%)	RMSE	MAD	MAPE (%)
ARMAX	104.77	77.27	38.88	125.84	98.61	82.69
RBF network	127.82	96.02	53.41	163.04	132.73	119.91

The results show that the ARMAX modeling approach generates the best prediction performance and significantly improves the forecast accuracy of the pure ARIMA model. In the best ARMAX model fitted, it is shown that the information about the climatic variables such as daily average temperature, precipitation amount, insolation duration and humidity is valuable in forecasting output power of a PV system. This reveals that some easily accessible climatic information can be used together with ARIMA to enhance the forecasting accuracy of time series models. Moreover, the ARMAX

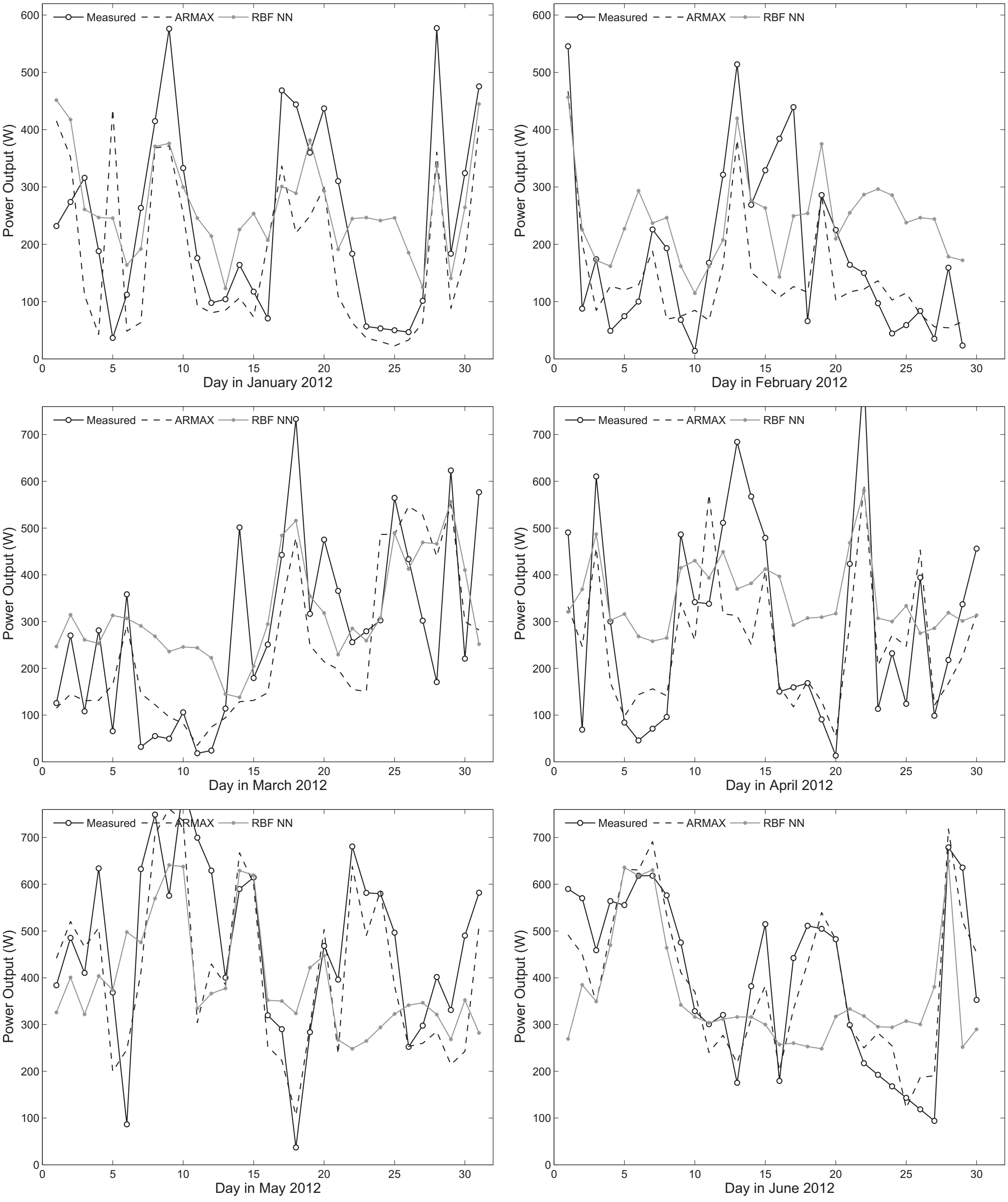

Fig. 13. Comparison of the 1-day ahead forecasts generated based on the RBF network and ARMAX models during January to June 2012.

model is shown to provide better prediction performance then the NN model.

Acknowledgments

The authors would like to thank the editor and referees for their valuable comments. Dr. Y. T. Li's work was supported in part by NSF of China under the grant NSFC 71272003. Dr. Y. Su's work was supported in part by the Research Committee under the grant MYRG080(Y1-L2)-FST13-SY and FDCT/115/2012/A. Dr. L. J. Shu's work was supported in part by the Research Committee under the grant MYRG096(Y2-L2)-FBA12-SLJ, MYRG090(Y1-L2)-FBA13-SLJ, and FDCT/002/2013/A.

References

[1] Mondal MAH, Islam AKMS. Potential and viability of grid-connected solar PV system in Bangladesh. Renew Energy 2011;36:1869–74.
[2] Strzalka A, Alam N, Duminil E, Coors V, Eicker U. Large scale integration of photovoltaics in cities. Appl Energy 2012;93:413–21.
[3] Woyte A, Thong VV, Belmans R, Nijs J. Voltage fluctuations on distribution level introduced by photovoltaic systems. IEEE Trans Energy Convers 2006;21:202–9.
[4] Pedro HTC, Coimbra CFM. Assessment of forecasting techniques for solar power production with no exogenous inputs. Sol Energy 2012;86:2017–28.
[5] Chen CS, Duan SX, Cai T, Liu BY. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Sol Energy 2011;85(11):2856–70.
[6] Alamsyah TMI, Sopian K, Shahrir A. Predicting average energy conversion of photovoltaic system in Malaysia using a simplified method. Renew Energy 2003;29(3):403–11.
[7] Ropp ME, Begovic M, Rohatgi A, Long R. Design considerations for large roof-integrated photovoltaic arrays. Prog Photovolt Res Appl 1997;5(1):55–67.
[8] Dalton GJ, Lockington DA, Baldock TE. Feasibility analysis of renewable energy supply options for a grid-connected large hotel. Renew Energy 2009;34(4):955–64.
[9] Martín L, Zaralejo LF, Polo J, Navarro A, Marchante R, Cony M. Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning. Sol Energy 2010;84:1772–81.
[10] Reikard G. Predicting solar radiation at high resolutions: a comparison of time series forecasts. Sol Energy 2009;83:342–9.
[11] Safi S, Zeroual A, Hassani MM. Prediction of global daily solar radiation using higher order statistics. Renew Energy 2002;27:647–66.
[12] Safi S, Zeroual A, Hassani MM. Modeling solar half-hour data using fourth order cumulants. Int J Sol Energy 2002;22:67–81.
[13] Mohandes M, Rehman S, Halawani TO. Estimation of global solar radiation using artificial neural networks. Renew Energy 1998;14:179–84.
[14] Sfetsos A, Coonick A. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Sol Energy 2000;68(2):169–78.
[15] Hontoria L, Aguilera J, Zuria P. Generation of hourly irradiation synthetic series using the neural network multilayer perception. Sol Energy 2002;72:441–6.
[16] Cao SH, Cao JC. Forecast of solar irradiance using recurrent neural networks combined with wavelet analysis. Appl Therm Eng 2005;25:161–72.
[17] Hocaoglu FO, Gerek ON, Kurban M. Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks. Sol Energy 2008;82(8):714–26.
[18] Paoli C, Voyant C, Muselli M, Nivet M. Forecasting of preprocessed daily solar radiation time series using neural networks. Sol Energy 2010;84:2146–60.
[19] Mellit A, Pavan AM. A 24-h forecast of solar irradiance using artificial neural network: application for performance prediction of a grid-connected PV plant at Trieste, Italy. Sol Energy 2010;84:807–21.
[20] Bacher P, Madsen H, Nielsen HA. On-line short-term solar power forecasting. Sol Energy 2009;83:1772–83.
[21] Fernandez-Jimenez LA, Munoz-Jimenez A, Falces A, Mendoza-Villena M, Garcia-Garrido E, Lara-Santillan PM, et al. Short-term power forecasting system for photovoltaic plants. Renew Energy 2012;44:311–7.
[22] Wong S, Wan KK, Lam TN. Artificial neural networks for energy analysis of office buildings with daylighting. Appl Energy 2010;87(2):551–7.
[23] Sulaiman SI, Abdul Rahman TK, Musirin I, Shaari S. Performance analysis of evolutionary ANN for output prediction of a grid-connected photovoltaic system. Int J Electr Comput Eng 2010;5(4):244–9.
[24] Ding M, Wang L, Bi R. An ANN-based approach for forecasting the power output of photovoltaic system. Procedia Environ Sci 2011;11:1308–15.
[25] Mora-lopez L, Martinez-Marchena I, Piliougine M, Sidrach-deCardona M. Machine learning approach for next day energy production forecasting in grid connected photovoltaic plants. World Renewable Energy Congress; 2011. pp. 2869–74.
[26] Box GEP, Jenkins GM. Time series analysis: forecasting and control. San Francisco: Holden-Day; 1976.
[27] Ljung L. System identification: theory for the user. New York: Prentice Hall; 1987.
[28] Alan P. Forecasting with dynamic regression models. New York: John Wiley & Sons, Inc; 1991.
[29] Wei WWS. Time series analysis: univariate and multivariate methods. Redwood city, CA: Addison-Wesley; 1990.
[30] Brown RG. Exponential smoothing for predicting demand. Cambridge, Massachusetts: Arthur D. Little Inc; 1956.
[31] Holt CC. Forecasting trends and seasonal by exponentially weighted averages. Int J Forecast 1957;20(1):5–10.
[32] Winters PR. Forecasting sales by exponentially weighted moving averages. Manag Sci 1960;6(3):324–42.
[33] Brockwell PJ, Davis RA. Time series: theory and methods. New York, USA: Springer-Verlag; 1991.
[34] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6(2):461–4.