

## Research Article

# Time-Series Regression Model for Prediction of Mean Daily Global Solar Radiation in Al-Ain, UAE

**Hassan A. N. Hejase and Ali H. Assi**

*Department of Electrical Engineering, United Arab Emirates University, P.O. Box 17555, Al-Ain, UAE*

Correspondence should be addressed to Ali H. Assi, ali.assi@uaeu.ac.ae

Received 1 December 2011; Accepted 10 January 2012

Academic Editors: A. Hasan and S. Rehman

Copyright © 2012 H. A. N. Hejase and A. H. Assi. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The availability of short-term forecast weather model for a particular country or region is essential for operation planning of energy systems. This paper presents the first step by a group of researchers at UAE University to establish a weather model for the UAE using the weather data for at least 10 years and employing various models such as classical empirical models, artificial neural network (ANN) models, and time-series regression models with autoregressive integrated moving-average (ARIMA). This work uses time-series regression with ARIMA modeling to establish a model for the mean daily and monthly global solar radiation (GSR) for the city of Al-Ain, United Arab Emirates. Time-series analysis of solar radiation has shown to yield accurate average long-term prediction performance of solar radiation in Al-Ain. The model was built using data for 10 years (1995–2004) and was validated using data of three years (2005–2007), yielding deterministic coefficients ( $R^2$ ) of 92.6% and 99.98% for mean daily and monthly GSR data, respectively. The low corresponding values of mean bias error (MBE), mean absolute bias error (MABE), mean absolute percentage error (MAPE), and root-mean-square error (RMSE) confirm the adequacy of the obtained model for long-term prediction of GSR data in Al-Ain, UAE.

## 1. Introduction

Gulf Corporation Countries (GCC) that include the United Arab Emirates (UAE) are seeking to make better use of their abundant sustainable energy sources and specifically solar energy. New sources of clean energy are essential to help reduce carbon dioxide levels that are increasing at an alarming level in order to protect the endangered ozone layer and avoid future climate changes. The potential of solar energy harvesting in the UAE is significant, with an average annual sunshine hours of 3568 h (i.e., 9.7 h/day), which corresponds to an average annual solar radiation of approximately 2285 kWh/m<sup>2</sup>, that is, 6.3 kWh/m<sup>2</sup> per day [1].

Many researchers worldwide have developed models to predict long-term daily and monthly average solar radiation in their region using different combinations of measured weather parameters. The availability of a solar radiation model in a particular region is very useful in estimating the amount of power that can be generated from a particular solar energy system. Short-term forecasts of solar radiation are also useful in planning the operation of energy systems as

well as in agricultural studies, for example, for crop modeling and for hydrology, meteorology, and soil physics.

Techniques used for modeling the mean daily global solar radiation include empirical regression, neural network, and time-series ARIMA models. The use of neural network (NN) or hybrid regression-NN models was mainly due to the nonlinearity in weather data.

Numerous authors [2–24] developed empirical regression models to predict the monthly average daily GSR in their region using various parameters. The mean daily sunshine duration was the most commonly used and available parameter. The most popular model was the linear model by Angström-Preseott [8, 19], which establishes a linear relationship between global radiation and sunshine duration with knowledge of extraterrestrial solar radiation and the theoretical maximum daily solar hours. Many studies with empirical regression models were done for diverse regions around the world. Menges et al. [17] reviewed 50 global radiation empirical models available in the literature for computing the monthly average daily global radiation on a horizontal surface. They tested the models on data recorded

in Konya, Turkey, for comparison of model accuracy. The number of weather parameters varied between models. The diverse regression models used include linear, logarithmic, quadratic, third-order polynomial, and logarithmic-linear and exponential and power models relating the normalized GSR to normalized sunshine hours. Other models included in Menges work used direct regression models involving various weather parameters in addition to geographical data (altitude, latitude) and other weather parameters such as precipitation and cloud cover. Şahin [20] presented a novel method for estimating the solar irradiation and sunshine duration by incorporating the atmospheric effects due to extraterrestrial solar irradiation and length of day. In this work, the author compares the model with Angström's equation with favourable advantages as his method does not use Least Square Method in addition to having no procedural restrictions or assumptions. Ulgen and Hepbasli [24] developed two empirical correlations to estimate the monthly average daily GSR on a horizontal surface for Izmir, Turkey. The developed models resemble Angström-type equations and were compared with 25 models previously reported in the literature on the basis of statistical error test (MBE, RMSE, MPE, and  $R^2$ ) with favourable results.

The main limitation of using many weather parameters is the difficulty of obtaining this data due to the high expense and availability of recording equipment. Some of the regression models are more accurate for monthly data than daily data and most published work only shows monthly data comparison results as the daily mean data are less accurate.

Other authors worked on prediction models based on ANN techniques and most specifically multilayer perceptron (MLP) and radial-basis function (RBF) methods [25–37]. The advantage of the ANN models is their ability to handle large amounts of data as well as the ability to handle random data without worry of incomplete, inaccurate, or noise-contaminated data. Behrang et al. [28] compared MLP and RBF ANN techniques for daily GSR modeling based on 6 proposed combinations. Their MLP-ANN models outperformed the best conventional GSR prediction (CGSRP) model yielding 5.21% MAPE error compared to a 10.02% MAPE error for the best CGSRP model. Another study by Benghanem et al. [29] uses ANN models for estimating and modeling daily GSR data in Al-Madinah, Saudi Arabia. In this work, the input data included the Mean daily air temperature, relative humidity, sunshine duration, and day of year. The authors compared the obtained models with conventional empirical regression models showing promising improvements. Numerous other researchers have worked on modeling GSR in their region using ANN techniques. Mohandes et al. [33] have applied ANN techniques to predict the mean monthly GSR using weather data from 41 stations in Saudi Arabia. Data from 31 stations was used in training the network, and the remaining data was used for testing. Input variables to the network included latitude, longitude, altitude, and sunshine duration. Their RBF model is compared with MLP ANN technique and other classical empirical models such as Angström-type equations. In this work, data samples were not large enough to obtain a credible comparison.

Another group of investigators [38–43] employed time-series regression modeling for the deterministic component and Box-Jenkins ARIMA modeling [44, 45] for the stochastic residual component. Boland [38] and Zeroual et al. [39] use a combination of regression model with Fourier series for the deterministic part and ARMA modeling for the residual stochastic part. Sulaiman et al. [40] and Zaharim et al. [41] use the ARMA Box-Jenkins method to model GSR data in Malaysia. The authors model the data using nonseasonal autoregressive models where the model adequacy is checked using the Ljung-Box statistic for diagnostic data. However, only short-term data was used for testing the model. Reikard [42] employs a combination of logarithmic regression and ARIMA modeling to predict solar radiation at high resolution and compares his models with other forecast methods such as ANN and considered the 24-hour daily seasonality not taken into account in most modeling approaches.

This work uses examples with classical empirical regression techniques and time-series techniques to predict the monthly average daily GSR data in Al-Ain, UAE. In this time-series regression technique, the deterministic component was modeled by decomposing it into a multiple linear regression component as a function of the available weather variables plus a cyclical component accounting for the annual periodicity and a linear trend. The residual error is studied using ARIMA modeling techniques to resemble a quasinormal noise error with zero mean and constant variance. The stationary form of the resulting time-series and the white-noise-type residual error provide extra-confidence of the long-term prediction accuracy of the estimated model.

The authors are doing extensive work on the analysis of weather data in three UAE cities, namely, Al-Ain, Abu Dhabi, and Sharjah. The prediction techniques employed range from classical one-parameter-based regression techniques (Sunshine data), MLP and RBF ANN techniques, and lately using time-series regression techniques with ARIMA modeling. The current work on solar radiation data in Al-Ain City, UAE, will be correlated with other ongoing studies by authors to come up with solar radiation prediction models for the UAE cities of Abu Dhabi and Sharjah. The final objective is to come up with a good national weather model capable of predicting the mean monthly GSR for the whole UAE within an acceptable prediction error.

## 2. Methodology

This paper discusses an ARIMA-based time-series regression technique used for predicting the monthly average GSR in the city of Al-Ain, UAE.

The weather database is provided by the National center of Meteorology and Seismology (NCMS) in Abu Dhabi, UAE. Weather data was provided for years 1995–2007 include the mean air temperature ( $T$  in °C), mean wind speed ( $W$  in knots), daily sunshine hours (SSH), and percent relative humidity (RH) in addition to the mean daily GSR in kWh/m<sup>2</sup>. The data is divided into two groups: one data group from 1995–2004 for the prediction model and the second group (2005–2007) for testing and validating the resulting

regression model. The data is first examined for any missing values and outliers. The missing values are replaced by the average of values from same week. Long arrays of missing data (a month for example) are replaced with the corresponding average of same days over the remaining model years. The total number of missing data in SSH or GSR columns does not exceed 3% of the total data points. Three leap-year days were also removed (February 29 for years 1996, 2000, 2004) to ensure uniformity of month comparison over the ten-year model period.

The time-series regression model to predict the mean daily global solar radiation (GSR) makes use of the four weather parameters (T, W, SSH, RH) measured daily over a period of 10 years, that is, 1995–2004. The highest mean monthly temperatures occur during the period of May–September with monthly temperatures around 45°C. The highest sunshine hours result during the months, of May and June with an average of 11.5 sunshine hours daily. During these two months the mean daily GSR values are the highest at 8 kWh/m<sup>2</sup>. March and August experience the highest wind speeds, while the humidity rates reach as high as 65% in December-January and as low as 30% during May.

The regression model relating the GSR data to the four predictor weather variables is obtained using the data of years 1995–2004. The partial least square multivariate regression technique is first used to correlate GSR with the four aforementioned dependent weather variables. The use of all the four predictor variables yields the highest correlation with a deterministic coefficient  $R^2 = 0.782$ .

The model equation obtained using SPSS [46] is

$$\begin{aligned} \text{GSR}_{\text{regression}}(t) = & -1.197 + 0.097 T(t) + 0.051 W(t) \\ & + 0.385 \text{SSH}(t) - 5.213 \times 10^{-3} \text{RH}(t). \end{aligned} \quad (1)$$

The regression model of (1) yields  $R^2 = 0.782$  and  $\text{MSE} = 0.425$ .

The next task is to examine the GSR residual component obtained after subtracting the regression component, that is,  $\text{GSR}_{\text{residue1}}(t)$ , for trends and/or seasonality and decompose it into the following:

$$\text{GSR}_{\text{residue1}}(t) = \text{GSR}_{\text{trend}}(t) + \text{GSR}_{\text{seasonal}}(t) + \text{GSR}_{\text{residue2}}(t). \quad (2)$$

The trend equation is found by performing a linear polynomial regression fit to yield

$$\text{GSR}_{\text{trend}}(t) = 0.077 - 4.221 \times 10^{-5} t. \quad (3)$$

The mathematical model for the seasonal component ( $\text{GSR}_{\text{seasonal}}(t)$ ) is found using the FFT algorithm ( $N = 3650$  data points) and is expressed in the Fourier series form:

$$\text{GSR}_{\text{seasonal}}(t) = \sum_{k=1}^{(\text{NP}-1)/2} \left[ a_k \cos\left(\frac{2\pi}{\text{NP}} kt\right) + b_k \sin\left(\frac{2\pi}{\text{NP}} kt\right) \right], \quad (4)$$

where the Fourier coefficients are computed from the FFT of  $\text{GSR}_{\text{seasonal}}(t)$ , that is,  $y[k]$ ,  $k = 1, 2, \dots, N$ , using the relations:

$$\begin{aligned} a_k &= 2 \text{Real}(y[k]), \\ b_k &= -2 \text{Imag}(y[k]), \end{aligned} \quad (5)$$

and where  $\text{NP} = 365$  days is the period of the data set. It is worth pointing out that nonzero values of  $a_k$  and  $b_k$  coefficients exist only for  $k = 10, 20, 30, \dots, 3650$ , thus yielding a total of 182 nonzero Fourier Coefficients that can be used to model the seasonal component of the residual error  $\text{GSR}_{\text{residue1}}(t)$ .

The overall decomposition leads to the following regression model:

$$\begin{aligned} \text{GSR}(t) = & \text{GSR}_{\text{regression}}(t) + \text{GSR}_{\text{trend}}(t) \\ & + \text{GSR}_{\text{seasonal}}(t) + \text{GSR}_{\text{residue2}}(t), \end{aligned} \quad (6)$$

where the multivariable linear regression, trend, and seasonal components are described in (4), (6), and (7), respectively. The normality test of the residual term  $\text{GSR}_{\text{residue2}}(t)$  in Minitab [47] yields a near zero mean (mean =  $-1.013E - 06$ ) and standard deviation of 0.393. The residual term  $\text{GSR}_{\text{residue2}}(t)$  exhibits quasinormal distribution behavior between the 1% and 95% percentiles with a computed skewness of 0.09 and kurtosis of 3.17. The normality plot is slightly skewed to the right of zero mean, and the curve is sharper than the normal distribution curve.

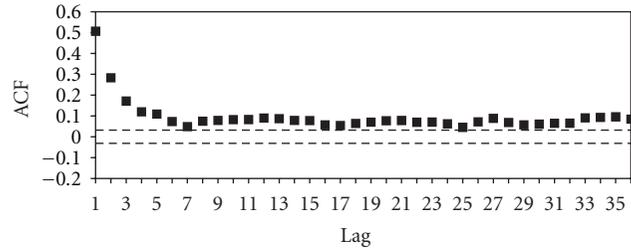
*2.1. Time-Series ARMA Modelling for the Residual Term (Stochastic Component).* The Box-Jenkins models [44] are only applicable to stationary time series. The identification of an appropriate box-Jenkins model for a particular time-series would first require a check for stationarity. If the residual term exhibits a normal distribution behavior with zero mean and constant variance, then it resembles white noise error and there is no need for further ARIMA modeling.

The behavior of the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots can help identify the ARMA model that best describes the resulting stationary time-series. Table 1 summarizes the ARMA model selection criteria [45]. In general, if the ACF of the time series value either cuts off or dies down fairly quickly, then the time series values should be considered stationary. On the other hand, if the ACF dies down extremely slowly, then the time series values may be considered nonstationary. If the model is adequate, then these plots should show all spikes within the 95% Confidence Interval (CI) bounds ( $\pm 1.96/\sqrt{N}$ ), where  $N$  is the sample size. If the series is non-stationary, one could try to apply differencing or log-transformation and then check if these make the series stationary. A stationary time-series would have a quasinormal distribution with zero mean and constant variance.

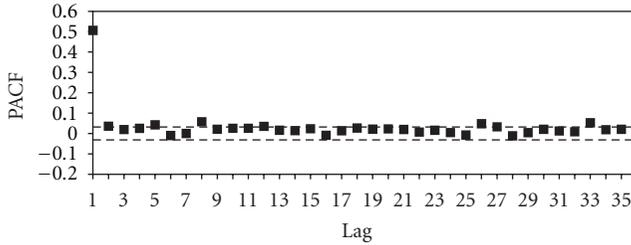
Figures 1(a) and 1(b) show, respectively, the ACF and PACF plots for the residual component as obtained from Minitab. Note that the ACF decays in an oscillating form after

TABLE 1: Behavior of ACF and PACF for each of the general non-seasonal models.

Model	ACF	PACF
AR ( $p$ )	Spikes decay towards zero. Coefficients may oscillate.	Spikes decay to zero after lag $p$
MA ( $q$ )	Spikes decay to zero after lag $q$	Spikes decay towards zero. Coefficients may oscillate.
ARMA ( $p, q$ )	Spikes decay (either direct or oscillatory) to zero beginning after lag $q$	Spikes decay (either direct or oscillatory) to zero beginning after lag $p$



(a)



(b)

FIGURE 1: (a) Autocorrelation function (ACF) and (b) partial Autocorrelation function (PACF) of GSR residual component for Al-Ain, UAE (1995–2004).

few lags within CI bounds thus implying a fairly stationary time-series. Differencing the residual component made the ACF and PACF more unstable with further deterioration of their behavior thus implying that differencing is not appropriate for this data. The PACF plot cuts off quickly after 2 lags indicating that an AR (2) or higher could be adequate.

The time-series analysis of the residual stochastic component of the mean daily GSR data is conducted using SPSS and Minitab. The nature of the ARMA model used is first described followed by explanation of the model selection process with diagnostic measures used to validate the selected model.

The nonseasonal Autoregressive-moving average model of order ( $p, q$ ), for example, ARMA ( $p, q$ ) is described by the equation [45]

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \dots - \theta_q \epsilon_{t-q}, \quad (7)$$

where  $\delta$  is a random shock and  $\phi_1, \phi_2, \dots, \phi_p$  are unknown autoregressive model coefficients that are estimated from sample data. The constant parameter is proportional to the mean value  $\mu$ , which is zero in our case.  $\theta_1, \theta_2, \dots, \theta_q$  are unknown moving average model coefficients that depend on the sample data. On the other hand,  $\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_{t-q}$  are statistically independent random shocks that are assumed to be randomly selected from a normal distribution with zero mean and constant variance.

The ARMA ( $p, q$ ) model parameters ( $\phi_i, i = 1, 2, \dots, p$  and  $\theta_j, j = 1, 2, \dots, q$ ) described in (7) are estimated in SPSS or Minitab using Least Square methods with known values of  $y_t$  representing the residual component data. Based on results seen in the ACF/PACF curves of Figure 1, an ARMA Model is executed in SPSS for  $p = 0, 1, 2, 3, 4, 5$  and  $q = 0, 1, 2, 3, 4, 5$  to select the parsimonious model that best describes the data. The estimated parameters of the selected ARMA ( $p, q$ ) models should have  $t$ -values higher than 2.0 in order to be judged significantly different from zero at the 5% level. Moreover, the coefficients should not be strongly correlated with each other in order to yield a parsimonious ARMA model, whilst passing the diagnostic checks. A parsimonious model is desirable because including irrelevant lags in the model increases the coefficient standard errors and therefore reduces the  $t$ -statistics. Models that incorporate large numbers of lags tend not to forecast well as they fit data specific features, explaining much of the noise or random features in the data. Therefore, model coefficients with  $P$  values higher than 0.05 are insignificant and should be eliminated to avoid overfitting as they have a little effect on the prediction model [45].

The differencing of the time-series led to deterioration of the ACF/PACF behavior, and thus no differencing was made ( $d = 0$ ), so ARIMA ( $p, 0, q$ ) will be considered, which we refer to as ARMA ( $p, q$ ) models. Different models can be obtained for various combinations of AR and MA individually and collectively. The best model is obtained with following diagnostics.

(a) *Low Akaike Information Criteria (AIC)/Schwarz-Bayesian Information Criteria (SBC)*. AIC and SBC are given in SPSS by

$$\begin{aligned} \text{AIC} &= -2 \log L + 2m, \\ \text{SBC} &= -2 \log L + m \ln(n), \end{aligned} \quad (8)$$

where  $m = p + q$ ,  $n = N - d$ , with  $N$  being the number of sample data points and  $d$  the order of differencing (here  $d = 0$ ), and  $L$  is the likelihood function. Since  $-2 \log L$  is approximately equal to  $\{n(1 + \log 2\pi) + n \log \sigma^2\}$  where  $\sigma^2$  is the model mean-square error (MSE), AIC can be written as  $\text{AIC} = \{n(1 + \log 2\pi) + n \log \sigma^2 + 2m\}$  and because the first term in this equation is a constant, it is usually omitted while comparing between models. Moreover, SBC can also be written as  $\text{SBC} = \log \sigma^2 + (m \log n)/n$ . SBC selects the more parsimonious model and is better than AIC for large samples. The best model should have the lowest AIC/SBC value and the least MSE.

(b) *Plot of Residual Autocorrelation Function (ACF)*. The appropriate ARMA model once fitted should have a residual error whose ACF plot varies within the 95% CI bounds ( $\pm 1.96/\sqrt{N}$ ), where  $N$  is the number of observations upon which the model is based.

(c) *Nonsignificance of Autocorrelations of Residuals via Portmanteau Tests (Q-Tests Based on Chi-Square Statistics) Such As Box-Pierce or Ljung-Box Tests (White Noise tests)*. Once the optimal ARMA ( $p, q$ ) model for the residual GSR time-series is selected, there is a need to check the white noise test if the ACF/PACF correlograms show significant spikes at one or more lags that could be just by chance. These tests indicate whether there is any correlation in the time-series or whether the abnormal spikes encountered in the ACF and PACF of the residual error are just a set of random, identically distributed variables overall.

The Ljung-Box Q-statistics can be used to check if the residuals from the ARMA ( $p, q$ ) model behave as a white-noise process [45, 48]. Ljung and Box [48] used the Q statistic:

$$Q = (n)(n+2) \sum_{k=1}^K \frac{r_k^2}{n-k}, \quad (9)$$

which yields a more accurate variance of ACF [variance becomes  $(n-K)/n^2$  instead of  $1/n$ ] compared to the statistic defined earlier by Box and Pierce [49].  $K$  is the degrees of freedom representing the maximum lags considered.  $n = N - d$  with  $N$  being the number of data points,  $d$  the degree of differencing (no differencing is assumed in this work so  $d = 0$ ), and  $r_k$  is the sample ACF at lag  $k$ .

Under the null-hypothesis that all values of autocorrelation  $r_k = 0$ , the Q Statistic is compared to critical values from chi-square distribution  $\chi^2$  distributed with  $K$ -degrees of freedom. If the model is correctly specified, the residuals should be uncorrelated and Q should be small and consequently the probability value should be large. A white noise process would ideally have  $Q = 0$ . Therefore, if  $Q > \chi_{DF,\alpha}^2$  at the specified degrees of freedom (DF) and significance level ( $\alpha$ ), then the null hypothesis can be rejected.

(d) *Check the Normality Test of the Residual Error of the Selected ARMA Model*. The residual error should resemble white noise with zero mean and constant variance.

Another method to ensure that the selected ARMA model yields a stationary residual error is by checking the *Unit-root rule for stationarity*.

Assume that the lag operator in (7) is  $B$ , then  $By = y_{t-1}$ . Then (7) can be written in the form (assume zero mean;  $\delta = 0$ ):

$$\begin{aligned} (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) y_t \\ = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) \epsilon_t \end{aligned} \quad (10)$$

or

$$\phi(B) y_t = \theta(B) \epsilon_t, \quad (11)$$

where

$$\begin{aligned} \phi(B) &= 1 - \sum_{i=1}^p \phi_i B^i \\ \theta(B) &= 1 - \sum_{i=1}^q \theta_i B^i. \end{aligned} \quad (12)$$

The stationarity of the time-series sequence  $y_t$  requires that all the roots of the polynomial  $\phi(B)$  should lie outside the unit circle [45].

Several ARMA models were analyzed based on the above criteria and two models surfaced out to be the best, namely, ARMA (2, 1) and ARMA (4, 3). Tables 2 and 3 show the estimated parameters for the ARMA (2, 1) and ARMA (4, 3) models, respectively, as obtained from SPSS.

The SPSS output shown in Tables 2 and 3 indicate that all estimated model coefficients have  $P$  value less than 0.05 ( $\alpha = 5\%$ ). This implies that all the coefficients of the two selected Box-Jenkins ARMA models are significant since the null hypothesis  $H_0 : \phi = 0$  (AR) or  $\theta = 0$  (MA) can be rejected for the preset significance level  $\alpha$  (chosen as 0.05). Other models tested resulted in coefficients with  $P$ -values greater than 0.05 and thus the coefficients will have little effect on model description (overfitting). Table 4 shows the estimated coefficients for few of the ARMA models studied to show the over-fitting behavior. Note that model ARMA (4, 4) is overfitted from ARMA (4, 3). The same behavior can be noted from models ARMA (2, 2) and ARMA (3, 1), which are overfitted from ARMA (2, 1) model.

The ACF and PACF plots of the residual error for the ARMA (2, 1) and ARMA (4, 3) models are displayed in Figures 2(a) and 2(b) and Figures 3(a) and 3(b), respectively. The plots vary within the 95% CI bounds. The spikes in the ACF and PACF at lag 7 are due to random events and thus cannot be explained. However, since it is a 95% confidence interval, one can expect this to happen once in every twenty lags and so we will not be concerned with this.

Ljung-Box test results for the selected ARMA (2, 1) and ARMA (4, 3) Models obtained using SPSS and JMulTi [50] are shown in Tables 5 and 6, respectively. The Ljung-Box Q-statistic values correspond to  $P$ -values greater than  $\alpha = 0.05$  thus indicating that the test is not significant and residuals appear to be uncorrelated. This implies that for both models, the GSR residual resembles random white noise. Both models will, therefore, be useful to forecast the GSR weather data, with ARMA (2, 1) being the more parsimonious and thus will be used hereafter. Model ARMA (2, 1) also has the lowest BIC value as witnessed from Table 4 and hence is the most suitable model for long-term prediction of GSR data.

For the selected ARMA (2, 1) model, the computed AR and MA roots of the polynomials  $\phi(B)$  and  $\theta(B)$  in (12) using JMulTi software are shown in Table 7. Note that all roots lie outside the unit circle and hence imply that ARMA (2, 1) model will yield a stationary residual error.

TABLE 2: SPSS least square estimation of the model parameters for ARMA (2, 1) model of residual GSR component.

Parameter	Lag	Estimate	Standard error	$t$ value	$P$ value
AR (1)	1	1.45044	0.01889	76.7886	0.000
AR (2)	2	-0.45681	0.01720	-26.5608	0.000
MA (1)	1	0.96992	0.00999	97.0868	0.000

TABLE 3: SPSS least square estimation of the model parameters for ARMA (4, 3) model of residual GSR component.

Parameter	Lag	Estimate	Standard error	$t$ value	$P$ value
AR (1)	1	0.5494	0.0202	27.2320	0.000
AR (2)	2	-0.1193	0.0130	-9.1758	0.000
AR (3)	3	1.0064	0.0129	78.0475	0.000
AR (4)	4	-0.4534	0.0175	-25.8911	0.000
MA (1)	1	0.0721	0.0125	5.7486	0.000
MA (2)	2	-0.1109	0.0122	-9.1137	0.000
MA (3)	3	0.9572	0.0121	78.8096	0.000

Once the parsimonious ARMA (2, 1) model is selected and checked for stationary residual error, the GSR regression model involving the ARMA model becomes

$$\begin{aligned} \text{GSR}(t) = & \text{GSR}_{\text{regression}}(t) + \text{GSR}_{\text{trend}}(t) \\ & + \text{GSR}_{\text{seasonal}}(t) + \text{GSR}_{\text{ARMA}}(t) + \text{WN}(t), \end{aligned} \quad (13)$$

where  $\text{WN}(t)$  is the final residual error resembling white noise with zero mean and constant variance.

Table 8 shows the result of running Levene's two-tailed variance test on the Regression model with and without ARMA modeling and measured data for years 1995–2004. This variance ratio test follows the hypothesis:

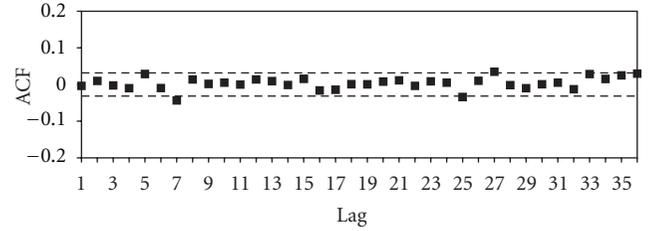
$H_0$  : the variances are identical,

$H_a$  : at least one of the variances is different from another.

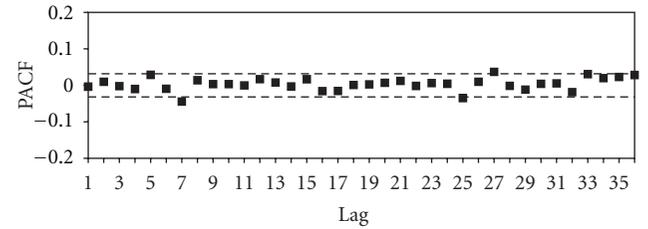
As the computed  $P$ -value in Table 8 are greater than the significance level  $\alpha = 0.05$ , one cannot reject the null hypothesis  $H_0$ . The risk to reject the null hypothesis  $H_0$  in Levene's test while it is true is 55.62% and 43.96% for regression and regression with ARMA cases, respectively. Levene's test is mostly used in samples with normal distribution.

*2.2. Comparison of the Time-Series Regression Model (with and without ARMA Model) with Measured Data for Years 1995–2004.* The “*cftool*” MATLAB GUI toolbox [51] is used to study the correlation between the regression model without (6) and with ARMA modeling (13) and the measured data set for years 1995–2004, as depicted in Figures 4(a) and 4(b). The predicted time-series regression model with and without ARMA modeling yields deterministic coefficients of 94.2% and 92.4%, respectively, as shown in Figures 3(a) and 3(b). No outliers were removed.

The statistical error parameters for both models are given in Table 9. Note the ARMA model helps decrease the RMSE,

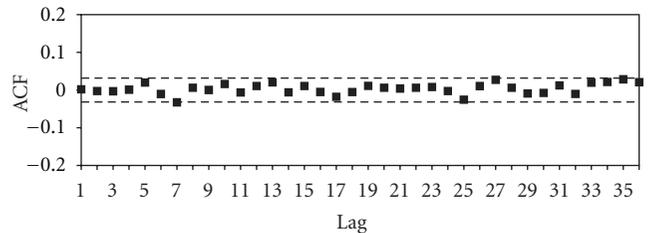


(a)

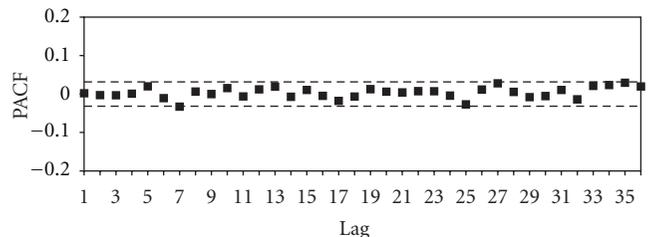


(b)

FIGURE 2: (a) Autocorrelation function (ACF) and (b) partial autocorrelation function (PACF) plots for the residual error of ARMA (2, 1) model.



(a)



(b)

FIGURE 3: (a) Autocorrelation function (ACF) and (b) partial autocorrelation function (PACF) plots for the residual error of ARMA (4, 3) model.

MABE, and MAPE error slightly and increase  $R^2$ , and hence will do a better forecasting job.

The mean daily GSR data comparison between the measured and the regression model for the years 1995–2004 in Al-Ain is shown in Figure 5.

The resulting mean monthly model data were computed and are shown in Figure 6.

Note the excellent agreement between the regression model and measured data for the period 1995–2004. Table 10 shows the mean monthly statistical error parameters for comparison of regression models with and without ARMA

TABLE 4: ARIMA model parameters.

ARMA ( $p, q$ )	AR coefficients ( $t$ value)	MA coefficients ( $t$ value)	Model remark	Normalized BIC (SPSS)	AIC, SBC, SE, SSE
(4, 3)	$\phi_1 = 0.5494$ (27.232)	$\theta_1 = 0.0721$ (5.748)		-2.159	AIC = 2408.1037
	$\phi_2 = -0.1193$ (-9.176)	$\theta_2 = -0.1109$ (-9.114)			SBC = 2451.5211
	$\phi_3 = 1.0064$ (78.048)	$\theta_3 = 0.9572$ (78.810)			SSE = 411.7070
	$\phi_4 = -0.4534$ (-25.891)				SE = 0.3362
(4, 4)	$\phi_1 = 0.5272$ (13.094)	$\theta_1 = 0.0483$ (1.116)	Over-fitting	-2.150	AIC = 2410.0453
	$\phi_2 = -0.0954$ (-4.605)	$\theta_2 = -0.0942$ (-5.385)			SBC = 2459.6652
	$\phi_3 = 0.9866$ (47.718)	$\theta_3 = 0.9420$ (49.891)			SSE = 411.7010
	$\phi_4 = -0.4369$ (-11.737)	$\theta_4 = 0.0166$ (0.452)			SE = 0.3362
(2, 1)	$\phi_1 = 1.4504$ (76.789)	$\theta_1 = 0.9699$ (97.087)		<b>-2.168</b>	AIC = 2416.2385
	$\phi_2 = -0.4568$ (-26.561)				<b>SBC = 2434.8459</b>
(3, 1)	$\phi_1 = 1.4504$ (75.183)	$\theta_1 = 0.9719$ (98.780)	Over-fitting	-2.165	AIC = 2417.9783
	$\phi_2 = -0.4461$ (-15.732)				SBC = 2442.7882
	$\phi_3 = -0.0087$ (-0.507)				SSE = 413.5336
(2, 2)	$\phi_1 = 1.4692$ (40.154)	$\theta_1 = 0.9920$ (24.575)	Over-fitting	-2.165	AIC = 2417.9862
	$\phi_2 = -0.4746$ (-13.441)	$\theta_2 = -0.0185$ (-0.510)			SBC = 2442.7962
					SSE = 413.5345
					SE = 0.3368

TABLE 5: Ljung-Box test for ARMA (2, 1) model for  $\alpha = 0.05$  and  $DF = K - (p + q)$ .

Lags ( $K$ )	12	16	18	20	24	36	48
DF	9	13	15	17	21	33	45
$\chi^2$ (DF, $\alpha$ )	16.919	22.362	24.996	27.587	32.671	47.400	61.656
Q Statistic	12.564	14.729	15.471	15.711	16.560	36.200	46.741
P-value	0.183	0.325	0.418	0.544	0.735	0.322	0.401

TABLE 6: Ljung-Box test for ARMA (4, 3) model for  $\alpha = 0.05$  and  $DF = K - (p + q)$ .

Lags ( $K$ )	12	16	18	20	24	36	48
DF	5	9	11	13	17	29	41
$\chi^2$ (DF, $\alpha$ )	11.070	16.919	19.675	22.362	27.587	42.557	56.942
Q Statistic	7.481	9.611	10.916	11.488	11.940	26.656	35.392
P-value	0.187	0.383	0.450	0.570	0.804	0.590	0.718

TABLE 7: Roots and moduli for selected ARMA models.

ARMA (2, 1)	Real	Imaginary	Modulus
AR (1)	2.1619	0.00000	2.16197
AR (2)	1.01175	0.00000	1.01175
MA (1)	1.03054	0.00000	1.03054

modeling of the residual error with the measured data for years 1995–2004.

2.3. *Validation of Time-Series Regression Model.* The resulting time-series regression model is validated with measured test data set for years 2005–2007 in Al-Ain, UAE, in order

TABLE 8: Levene’s two-tailed variance test ( $\alpha = 0.05$ ).

Parameter	Regression versus measured	Regression with ARMA versus measured
$F$ (Observed value)	0.346	0.597
$F$ (Critical value)	3.843	3.843
DF1	1	1
DF2	7298	7298
P-value (one-tailed)	0.5562	0.4396

to check the prediction accuracy of the model in forecasting GSR data. Figure 7 shows the comparison of mean daily regression with measured test data for years 2005–2007. The statistical error data computed with MATLAB are given in Table 11. Testing results yield a deterministic coefficient of 90.77% for the regression model and 92.59% for the regression-ARMA model indicating a good model prediction performance.

The resulting monthly mean GSR data comparison between test (measured) and regression model for years 2005–2007 is shown in Figure 8. Note the excellent agreement between regression model with ARMA (2, 1) and test data. Figure 8 also shows comparison results obtained using Multilayer Perceptron (MLP) and Radial Basis Function (RBF) Artificial Neural Networks (ANN) obtained by the coauthor [26] using the same model and test data sets. Figure 8 shows a better prediction performance for the Al-Ain test data. Note from Table 12 the very high  $R^2$  value (near 100%) obtained for monthly mean data validation for both time-series regression models. The low error parameters (RMSE, MAPE, MBE, and MABE) provide a clear indication

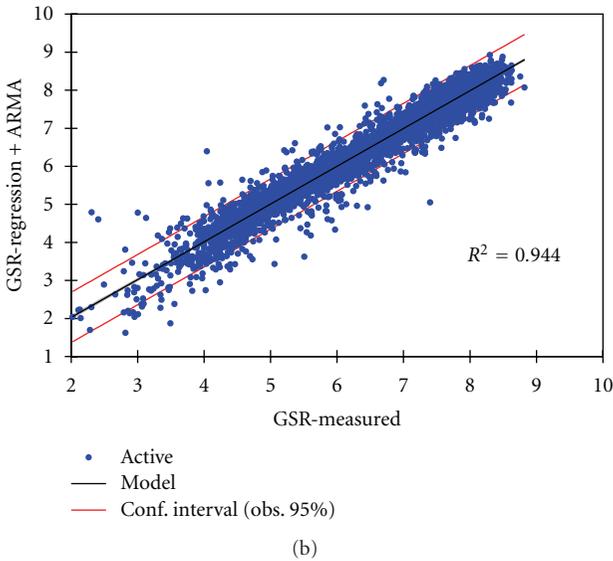
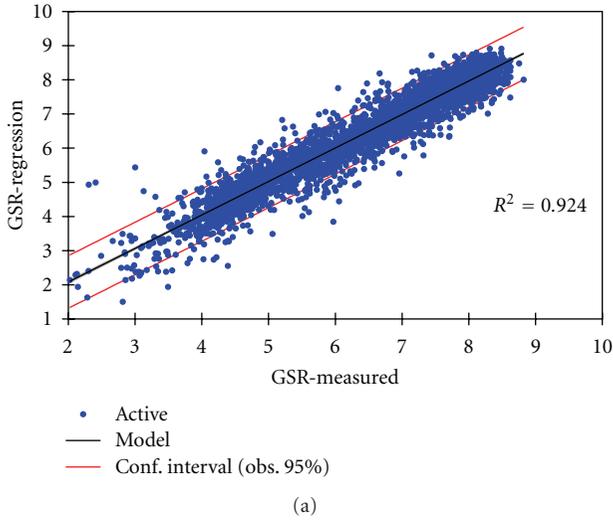


FIGURE 4: Comparison between regression and measured data for years 1995–2004, (a) with and (b) without ARMA model.

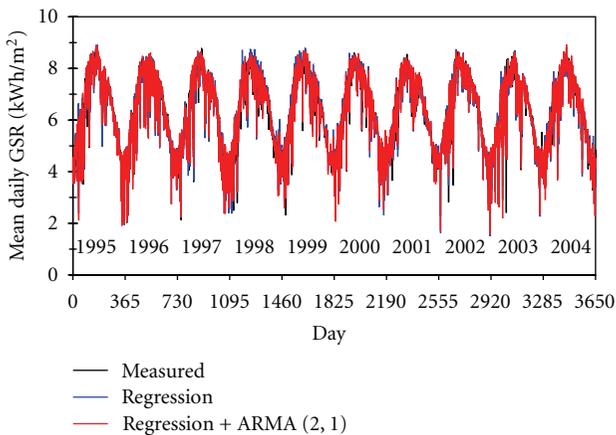


FIGURE 5: Mean daily global solar radiation data for years 1995–2004 in Al-Ain, UAE.

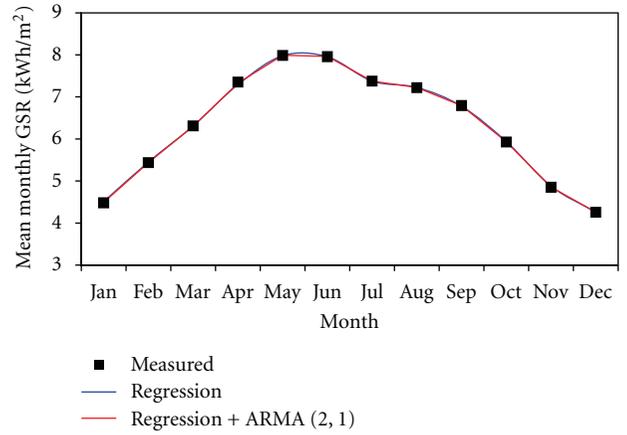


FIGURE 6: Monthly mean GSR measured and model data for years 1995–2004.

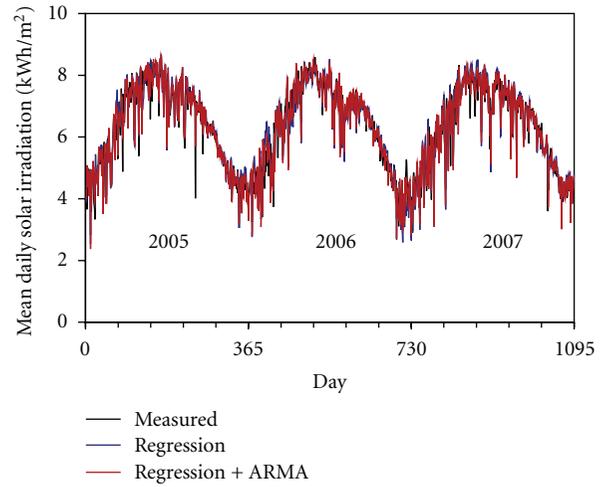


FIGURE 7: Mean daily GSR for regression model with the test data set for years 2005–2007 for Al-Ain city.

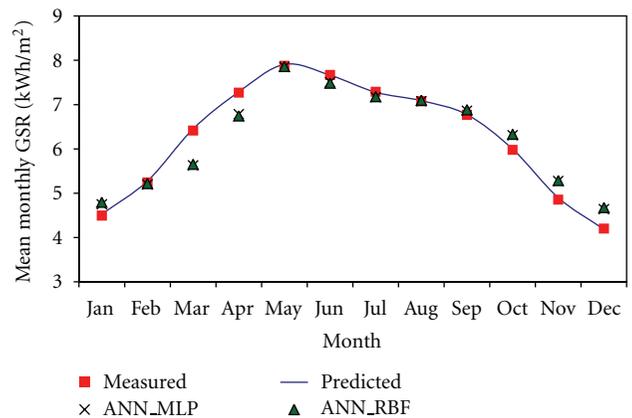


FIGURE 8: Monthly mean GSR data comparison for test data 2005–2007 in Al-Ain, UAE.

TABLE 9: Statistical error data for the mean daily GSR model for regression and regression-ARMA models versus measured model data for years 1995–2004.

Description	Regression-measured	Regression and ARMA-measured
RMSE	0.3926	0.3363
MBE	$-1.1013E - 06$	$5.8476E - 05$
MABE	0.2818	0.2380
MAPE	4.85%	4.16%
$R^2$	92.40%	94.42%

TABLE 10: Statistical error data for the mean monthly GSR model for regression and regression-ARMA models versus measured model data for years 1995–2004.

Description	Regression	Regression + ARMA
RMSE	0.0228	0.0096
MBE	$-1.47E - 04$	$6.50E - 05$
MABE	0.0152	0.0065
MAPE	0.256%	0.103%
$R^2$	99.979%	99.995%

TABLE 11: Statistical error data for the mean daily GSR error for regression and regression-ARMA models versus measured test data for years 2005–2007.

Description	Regression versus test data	Regression-ARMA versus test data
RMSE	0.4161	0.3701
MBE	-0.0464	-0.0164
MABE	0.3082	0.2704
MAPE	5.37%	4.71%
$R^2$	90.77%	92.59%

TABLE 12: Statistical error data for the mean monthly GSR error for regression models versus measured test data for years 2005–2007.

Description	Regression versus test data	Regression-ARMA versus test data
RMSE	0.0678	0.0254
MBE	-0.0475	-0.0167
MABE	0.0526	0.0217
MAPE	0.9214%	0.3825%
$R^2$	99.85%	99.98%

of the potential of these techniques for long-term GSR data prediction.

### 3. Definition of the Statistical Error Parameters

The formulas used to compute the statistical error parameters in MATLAB are given in [1]. These parameters attest to the accuracy of the models used for predicting the mean daily global solar radiation.

## 4. Conclusion

The solar radiation data taken in Al-Ain, UAE, are analyzed using time-series regression with Box-Jenkins ARMA model. Based on the behavior of ACF and PACF plots, the nonseasonal ARMA (2, 1) and ARMA (4, 3) models were identified as the two most adequate models with model ARMA (2, 1) being the most parsimonious and best suited for the long-term GSR prediction as it yielded the lowest BIC value. The model adequacy was conformed through Ljung-Box Q test and normality test. The model was validated by testing with GSR data for years 2005–2007 yielding deterministic coefficients of 92.59% and 99.98%, for mean daily and monthly GSR data comparison, respectively. The corresponding values for RMSE, MBE, and MAPE were also small confirming the adequacy of the regression-ARMA model for GSR data prediction in the city of Al-Ain, UAE. Work is underway to establish a model for the city of Abu Dhabi, UAE, using GSR measured data for the same time period in order to come up later with a more generalized GSR weather model for the UAE.

## Acknowledgments

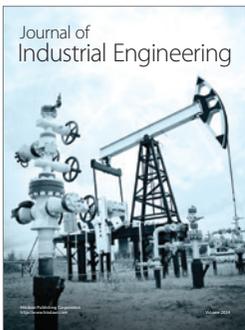
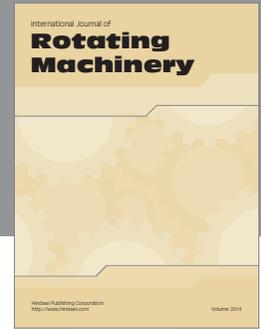
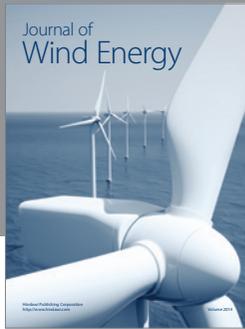
The authors would like to thank the National Center of Meteorology and Seismology (NCMS), Abu Dhabi, for providing the weather data. Thanks are also due to Maitha Al-Shamisi for preprocessing the received data. This paper is financially supported by UAE University Research Affairs under the contract no. 1542-07-01-10.

## References

- [1] A. Assi and M. Jama, "Predicting global solar radiation on horizontal using sunshine duration in al Ain, UAE," in *Proceedings of the 25th European Photovoltaic Solar Energy Conference and Exhibition /5th World Conference on Photovoltaic Energy Conversion*, pp. 508–512, Valencia, Spain, September 2010.
- [2] Y. A. G. Abdalla and G. M. Feregh, "Contribution to the study of solar radiation in Abu Dhabi," *Energy Conversion and Management*, vol. 28, no. 1, pp. 63–67, 1988.
- [3] B. G. Akinoglu and A. Ecevit, "A further comparison and discussion of sunshine based models to estimate global solar radiation," *Energy*, vol. 15, no. 10, pp. 865–872, 1990.
- [4] A. Al-Lawati, A. S. S. Dorvlo, and J. A. Jervase, "Monthly average daily solar radiation and clearness index contour maps over Oman," *Energy Conversion and Management*, vol. 44, no. 5, pp. 691–671, 2003.
- [5] N. Al Mahdi, N. S. Al Baharna, and F. F. Zaki, "Assessment of solar radiation models for the Gulf Arabian countries," *Renewable Energy*, vol. 2, no. 1, pp. 65–71, 1992.
- [6] J. Almorox, M. Benito, and C. Hontoria, "Estimation of global solar radiation in Venezuela," *Interciencia*, vol. 33, no. 4, pp. 280–283, 2008.
- [7] D. B. Ampratwum and A. S. S. Dorvlo, "Estimation of so-lar radiation from the number of sunshine hours," *Applied Energy*, vol. 63, no. 3, pp. 161–167, 1999.
- [8] A. Angström, "Solar and terrestrial radiation," *Quarterly Journal of the Royal Meteorological Society*, vol. 50, pp. 121–125, 1924.

- [9] A. Assi and M. Jama, "Estimating global solar radiation on horizontal from sunshine hours in Abu Dhabi, UAE," in *Proceedings of the 4th International Conference on Renewable Energy Sources (RES '10)*, pp. 101–108, Sousse, Tunisia, May 2010.
- [10] K. Bakirci, "Correlations for estimation of daily global solar radiation with hours of bright sunshine in Turkey," *Energy*, vol. 34, no. 4, pp. 485–501, 1999.
- [11] R. A. Ball, L. C. Purcell, and S. K. Carey, "Evaluation of solar radiation prediction models in North America," *Agronomy Journal*, vol. 96, no. 2, pp. 391–397, 2004.
- [12] N. Elagib and M. G. Mansell, "New approaches for estimating global solar radiation across Sudan," *Energy Conversion and Management*, vol. 41, no. 5, pp. 419–434, 2000.
- [13] E. O. Falayi, J. O. Adepitan, and A. B. Rabi, "Empirical models for the correlation of global solar radiation with meteorological data for Iseyin, Nigeria," *International Journal of Physical Sciences*, vol. 3, no. 9, pp. 210–216, 2008.
- [14] J. Fortin, F. Antcil, L. Parent, and M. Bolinder, "Comparison of empirical daily surface incoming solar radiation models," *Agricultural and Forest Meteorology*, vol. 148, no. 8-9, pp. 1332–1340, 2008.
- [15] Z. Jin, Y. Wu, and Y. Gang, "General formula for estimation of monthly average daily global solar radiation in China," *Energy Conversion and Management*, vol. 46, no. 2, pp. 257–268, 2005.
- [16] A. Khalil and A. Alnajjar, "Experimental and theoretical investigation of global and diffuse solar radiation in the United Arab Emirates," *Renewable Energy*, vol. 6, no. 5-6, pp. 537–543, 1995.
- [17] H. O. Menges, C. Ertekin, and M. H. Sonmete, "Evaluation of global solar radiation models for Konya, Turkey," *Energy Conversion and Management*, vol. 47, no. 18-19, pp. 3149–3173, 2006.
- [18] F. J. Newland, "A study of solar radiation models for the coastal region of South China," *Solar Energy*, vol. 31, no. 4, pp. 227–235, 1988.
- [19] G. Podestá, G. Núñez, C. Villanueva, and M. Skansi, "Estimating daily solar radiation in the Argentine Pampas," *Agricultural and Forest Meteorology*, vol. 123, no. 1-2, pp. 41–53, 2004.
- [20] A. D. Şahin, "A new formulation for solar irradiation and sunshine duration estimation," *International Journal of Energy Research*, vol. 31, no. 2, pp. 109–118, 2007.
- [21] T. Samuel, "Estimation of global radiation for Sri Lanka," *Solar Energy*, vol. 47, no. 5, pp. 333–337, 1991.
- [22] M. T. Y. Tadros, "Uses of sunshine duration to estimate the global solar radiation over eight meteorological stations in Egypt," *Renewable Energy*, vol. 21, no. 2, pp. 231–246, 2000.
- [23] I. Togrul and H. Togrul, "Global solar radiation over Turkey: comparison of predicted and measured data," *Renewable Energy*, vol. 25, no. 1, pp. 55–67, 2002.
- [24] K. Ülgen and A. Hepbasli, "Comparison of solar radiation correlations for Izmir, Turkey," *International Journal of Energy Research*, vol. 26, no. 5, pp. 413–430, 2002.
- [25] S. Al-Alawi and H. Al-Hinai, "An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation," *Renewable Energy*, vol. 14, no. 1–4, pp. 199–204, 1998.
- [26] A. Assi and M. Al-Shamisi, "Prediction of monthly average daily global solar radiation in al Ain city, UAE, using artificial neural networks," in *Proceedings of the 25th European Photovoltaic Solar Energy Conference*, pp. 508–512, Valencia, Spain, September 2010.
- [27] A. Assi, M. Al-Shamisi, and M. Jama, "Prediction of monthly average daily global solar radiation in Al Ain City, UAE, using artificial neural networks," in *Proceedings of the 4th International Conference on Renewable Energy Sources (RES '10)*, pp. 109–113, Sousse, Tunisia, May 2010.
- [28] M. A. Behrang, E. Assareh, A. Ghanbarzadeh, and A. R. Noghrehabadi, "The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data," *Solar Energy*, vol. 84, no. 8, pp. 1468–1480, 2010.
- [29] M. Benghanem, A. Mellit, and S. N. Alamri, "ANN-based modelling and estimation of daily global solar radiation data: a case study," *Energy Conversion and Management*, vol. 50, no. 7, pp. 1644–1655, 2009.
- [30] M. Bocco, E. Willington, and M. Arias, "Comparison of regression and neural networks models to estimate solar radiation," *Chilean Journal of Agricultural Research*, vol. 70, no. 3, pp. 428–435, 2010.
- [31] H. Elminir, F. Areed, and T. Elsayed, "Estimation of solar radiation components incident on Helwan site using neural networks," *Solar Energy*, vol. 79, no. 3, pp. 270–279, 2005.
- [32] T. Krishnaiah, R. S. Srinivasa, K. Madhumurthy, and K. S. Reddy, "A neural network approach for modelling global solar radiation," *Applied Science Research*, vol. 3, no. 10, pp. 1105–1111, 2007.
- [33] M. Mohandes, S. Rehman, and T. O. Halawani, "Estimation of global solar radiation using artificial neural networks," *Renewable Energy*, vol. 14, no. 1–4, pp. 179–184, 1998.
- [34] M. Mohandes, A. Balghonaim, M. Kassas, S. Rehman, and T. O. Halawani, "Use of radial basis functions for estimating monthly mean daily solar radiation," *Solar Energy*, vol. 68, no. 2, pp. 161–168, 2000.
- [35] S. Rehman and M. Mohandes, "Artificial neural network estimation of global solar radiation using air temperature and relative humidity," *Energy Policy*, vol. 36, no. 2, pp. 571–576, 2008.
- [36] I. Tasadduq, S. Rehman, and K. Bubshait, "Application of neural networks for the prediction of hourly mean surface temperatures in Saudi Arabia," *Renewable Energy*, vol. 25, no. 4, pp. 545–554, 2002.
- [37] C. Walthall, W. Dulaney, M. Anderson, J. Norman, H. Fang, and S. Liang, "A comparison of empirical and neural network approaches for estimating corn and soybean leaf area index from Landsat ETM+ imagery," *Remote Sensing of Environment*, vol. 92, no. 4, pp. 465–474, 2004.
- [38] J. Boland, "Time-series analysis of climatic variables," *Solar Energy*, vol. 55, no. 5, pp. 377–388, 1995.
- [39] A. Zeroual, M. Ankrim, and A. J. Wilkinson, "Stochastic modelling of daily global solar radiation measured in Marrakesh, Morocco," *Renewable Energy*, vol. 6, no. 7, pp. 787–793, 1995.
- [40] M. Y. Sulaiman, W. M. H. Oo, M. A. Wahab, and Z. A. Sulaiman, "Analysis of residuals in daily solar radiation time series," *Renewable Energy*, vol. 11, no. 1, pp. 97–105, 1997.
- [41] A. Zaharim, A. M. Razali, T. P. Gim, and K. Sopian, "Time series analysis of solar radiation data in the tropics," *European Journal of Scientific Research*, vol. 25, no. 4, pp. 672–678, 2009.
- [42] G. Reikard, "Predicting solar radiation at high resolutions: a comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, 2009.
- [43] T. N. Goh and K. J. Tan, "Stochastic modeling and forecasting of solar radiation data," *Solar Energy*, vol. 19, no. 6, pp. 755–757, 1977.
- [44] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Prentice Hall, Englewood Cliffs, NJ, USA, 3rd edition, 1994.

- [45] W. Enders, *Applied Econometric Time Series*, John Wiley & Sons, New York, NY, USA, 3rd edition, 2010.
- [46] IBM SPSS Statistics for Windows, Version 19.0.0, SPSS , 2010.
- [47] Minitab, version 16.1.0, Minitab, 2010.
- [48] G. Ljung and G. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.
- [49] G. Box and D. Pierce, "Distribution of autocorrelations in autoregressive moving average time series models," *Journal of the American Statistical Association*, vol. 65, pp. 1509–1526, 1970.
- [50] H. Lutkepohl and M. Kratzig, *Applied Time Series Econometrics*, Cambridge University Press, Cambridge, UK, 2004, JMulTi.
- [51] MATLAB, version 7.10.0.499 (R 2010 a), The MathWorks, 2010.



**Hindawi**

Submit your manuscripts at  
<http://www.hindawi.com>

