RESEARCH ARTICLE

Comparative optimization of global solar radiation forecasting using machine learning and time series models

Brahim Belmahdi[1] · Mohamed Louzazni[2] · Abdelmajid El Bouardi[1]

Abstract

The increasing use of solar energy as a source of renewable energy has led to increasing the interest in photovoltaic (PV) power outputs forecasting. In the meantime, forecasting global solar radiation (GSR) depends heavily on weather conditions, which fluctuate over time. In this paper, an algorithm method is proposed, to choose the optimum machine learning techniques and time series models which minimizing the forecasting errors. The forecasted GSR belongs to the Faculty of Sciences, Abdelmake Eassadi University, Tetouan, Morocco. The selected machine learning and times series are Autoregressive Integrated Moving Average (ARIMA), Feed Forward Neural Network with Back Propagation (FFNN-BP), k-Nearest Neighbour (k-NN), and Support Vector Machine (SVM) compared with persistence model as the reference model. To compare the results, several statistical metrics are calculated to evaluate the performance of each method. The obtained results indicated that the used machine learning and time series methods were more straightforward to implement. In particular, we find that the Feedforward neural network (FFNN) and ARIMA perform better and give good approximations with the corresponding GSR output.

Introduction

Nowadays, the energy demand is increasing worldwide with increasing population and economic development (Anon 2017). The need for energy has become more important for satisfying human social and economic development. Recently, renewable energies are growing field that provides clean energy without any harmful residue or contamination alternative to fossil fuel (Kumar and Kumar 2017). Among the renewable energy sources (RES), solar photovoltaic, solar thermal, solar cylindrical transform the solar radiation into electricity and thermal energy. RES has grown in popularity in recent years and is now successfully reported in a variety of applications. The generated electricity and energy from solar

radiation require precise knowledge of the power produced for injection to the grid (Shiva Kumar and Sudhakar 2015).

Morocco has undertaken several renewable energy projects, including solar, wind, hydraulics, and biomass. The country has a large capacity for solar radiation in wide regions with a daily average solar radiation close to 5.80 kW h/m^2/day (Belmahdi et al. 2020b). The key challenge in implementing these technologies is the use of solar power as a source of electricity in photovoltaic generators (PVG), thermal solar energy (TSE), and solar concentration technologies (CPV). Such a challenge has motivated scientists and researchers to find successful methods for predicting the value of solar radiation (Belmahdi et al. 2020a). It was pointed out that the precision of the solar radiation model is inextricably tied to the accuracy of power generation from installed solar systems, and that this has an impact on management. (Badescu et al. 2013). Using an effective model for predicting solar radiation, it is possible to monitor the power generated by the photovoltaic system. In particular, the measurement, analysis, and forecasting of global solar radiation are important because it is of great significance for the success of PVG, TSE, and CPV in the development of electrical energy and its incorporation into the electrical grid. In terms of enhancing and maintaining that the electricity produced from the RES source has been well

---

Responsible Editor: Philippe Garrigues

✉ Brahim Belmahdi
belmahdi.brahim@gmail.com

[1] Energetic Laboratory, ETEE, Faculty of Sciences, Abdelmalek Essaadi University, Tetouan, Morocco

[2] Science Engineer Laboratory for Energy, National School of Applied Sciences, Chouaib Doukkali University, El Jadida, Morocco

introduced into the electrical grid without disturbances. Within that aspect, the forecasting of global solar radiation would have a huge effect on the production and maintenance of future energy systems. In the research area, there are several papers proposed and evolved different strategies for forecasting global solar radiation. The forecasting approach chosen is relevant to the available information and its specific forecast visibility. The forecast information's are based on the meteorological conditions, and it is forecasted into three different categories named i) short-term forecast, ii) medium-term, and iii) long-term. The forecasting of global solar radiation would make a meaningful contribution to the modelling, design and control management of modern energy systems, such as the connection to the microgrids (Al-Dahidi et al. 2020).

Various methods for forecasting solar radiation have been recorded in different scientific papers. The most widely used forecasting methods are grouped into three primary groups, namely conventional models, machine learning, statistical regression methods, and hybrid methods (Fan et al. 2019)(Wu et al. 2019). Conventional models (CM), also known as predictive or mathematical models, may be identified as dynamic and empiric models (Khorasanizadeh et al. 2014)(Almorox and Hontoria 2004). In research papers, machine learning (ML) techniques for time series forecasting have been recommended as supplements to computational methods (Makridakis et al. 2018). The purpose of the ML methods is the same as that of the statistical regression methods. Both attempts to improve forecast accuracy by decreasing any loss function, usually the number of squared errors. Statistical forecasting methods focused on observational and quantitative data gathered, related to time series analysis, based on the notion that observations from history will continue in the future.

Researchers have developed a variety of methods, strategies, and algorithms to forecast solar radiation employing CM, ML, or a combination of both to define hybrid methods. Artificial neural networks (ANNs) have become the most popular and commonly utilized type of neural network in literature (Pazikadin et al. 2020). It has proved high effectiveness in modelling dynamic non-linear structures relative to traditional models. For example, in (Mehleri et al. 2010) and (Notton et al. 2013) have been presented that the ANNs approximate the inclined global irradiation with high precision relative to traditional isotropic and anisotropic models Belmahdi et al. 2021. IN Feedforward back propagation algorithm was used to predict daily global solar radiation in 25 cities around the kingdom of Morocco. Several meteorological astronomical and geographical coordinate were employed as inputs data to predict the out coming output. Multiple combination parameters were adopted in order to select the most suitable configuration with optimal input data for each study location. According to statistical metrics, the obtained result are respectively, 12 inputs for Er-Rachidia, Marrakech, Medilt, Taza,

Oujda, Nador, Tetouan, Tanger, Al-Auin, Dakhla, Settat, and Safi, seven inputs for Fes, Ifrane, Beni-Mellal, and Meknes, six inputs for Agadirand Rabat, five inputs for Sidi Ifni, Essaouira, Casablanca and Kenitra, four inputs for Ouarzazate, Larache, and Al-Hoceima In terms of accuracy, R^2 of the selected best inputs parameters varies between 0.9860% and 0.9920%, the range value of MBE (%) being from −0.1076% to −0.5931%, the RMSE between 0.1990 and 0.4580%, the range value of the NRMSE is between 0.0355 and 0.8938, and the lowest value of the MAPE is between 0.0019 and 0.0060%. This technique could be used to predict other parameters for locations where measurement instrumentation is unavailable or costly to obtain. Two ANNs were designed in (Mellit et al. 2013) for use on cloudy and sunny days, respectively, using more than one year of experimental data at the Marmara University, Istanbul, Turkey. The two models are used to forecast the power output of the 50 Wp Si-polycrystalline photovoltaic module on cloudy and sunny days. The proposed ANNs are composed of 3 layers, one input layer, a single hidden layer and an output layer. Solar radiation and ambient temperature are used in the input's layers, while the generated power from the photovoltaic module is used as an output layer with a single node. The results indicate that the ANNs models outperformed generic polynomial regression (PR), multiple linear regression (MLR), analytical and one-diode model. in the (Cervone et al. 2017) authors implemented a method based on the ANNs and Analog Ensemble (AnEn) method to produce 72 hours deterministic and probabilistic forecasts of a photovoltaic system. the input is the meteorological conditions (solar radiation, ambient temperature...) and computed astronomical variables. ANNs and AnEn are implemented to forecast short-term generated power from a photovoltaic system installed in Italy. The results show that the ANN-based AnEn is perfectly adapted for global level processing. In another paper, the ANNs has proposed to forecast the short-term wind speed, solar radiation, and electrical power demand (Di Piazza et al. 2021). The methodology is implanted in the open-loop structure to perform the time series forecasting of wind speed, solar radiation and load power demand. The results indicate that the best configuration for solar radiation forecasting is using two neurons in the input, five hidden layers and one input layer with a time delay of 7. The achieved simulation results compared with the experimental data indicate that the ANNs based in the exogenous inputs model is well adapted to perform energy-related time series forecasting in the short-term time horizon. In a recent paper, the authors have proposed and developed six ANNs models for indigenous and widespread regions around the world using two datasets (Kılıç et al. 2021). The proposed method has used the evolutionary algorithm and ANNs to optimize 19 input parameters, classification and forecasting. The results have given that the absolute percentage errors APE=2.45%, validation dataset=9.93% and testing

dataset=11.03% of the indigenous and widespread regions, respectively. Fortunately, ANNs have certain drawbacks, such as the identification of the optimum number of hidden neurons, which is mostly dependent on error checks, the availability of initial values for synaptic weights, local minima problems in the learning process, etc. The ANNs or NN have been widely used as a hybrid model with numerous techniques. For example, the ANNs has combined with satellite-derived(MODIS) and land surface temperature (LST) to forecast long-term global solar radiation in Queensland, Australia (Deo and Şahin 2017). LST data from 2012 to 2014 are collected and divided into seven categories, each with three locations, with the first two groups (2012-2013) used for the simulation process and the third group (2014) used for cross-validation. The proposed technique is tuned for the monthly horizon by testing with 55 neural structures, while nine neuronal architectures are tailed with time-lagged LST for seasonal forecasting. The proposed ANNs is training using 55 neural and nine neural for the monthly horizon and seasonal forecasting, respectively. The scaled conjugate gradient algorithm (SCGA) was used in the ANN with zero lagged LST, while the Levenberg-Marquardt algorithm (LMA) was used in the ANN with time-delayed LST. The analysis indicates that an ANN consistently outperforms multiple linear regression (MLR) and autoregressive integrated moving average (ARIMA) models, with an examination yielding 39% of cumulative errors in the lowest magnitude bracket, compared to 15% and 25% for MLR and ARIMA, respectively. The MODIS model was implemented to estimate the monthly global solar radiation using data from 50 locations around China, 8 statistical models were investigated and developed (Chen et al. 2014). The models are based on three parameters named cloud fraction (CF), cloud optical thickness (COT), precipitable water vapour (PWV) and aerosol optical thickness (AOT). The first parameters were used in all models for its crucial component influencing solar radiation in the atmosphere and it used in all models. Model 2 uses both CF and COT, models 3 and 4 are only modified from models 1 and 2 by adding PWV amount, whereas models 5 and 6 also are the modified models of models 1 and 2 by adding the AOT. The last model 8 takes into account all of the variables that make up the atmosphere. The results showed that all the models provide a satisfactory result with an average RMSE of 1.247 MJ m^{-2} and MAPE of 9.9%. The models generated a reduced RMSE under cool temperature and warm temperate zones. In Iran and using MODIS model to forecast the global solar radiation in the urban area applying satellite data to identify various atmospheric parameters such as CF, COT, AOT, cloud optical depth (COD), aerosol exponent (AE)(Bamehr and Sabetghadam 2021). The models were created based on seven combinations of atmospheric variables developed under standard statistical methods, namely, MLR, and a specific class of ANNs, namely, feedforward multilayer perceptron

(FFMP). The results conclude that the ANNs is more accurate than MLR based in the regression method. On the other hand, several models and techniques are used in the research papers. From this, we found the deep learning (DL) model and Variational Bayesian Inference (VBI) has been developed and used to forecast the solar radiation using the historical information by considering the past solar radiation and weather conditions from a multi-site location in China (Liu et al. 2019). The forecasted results have been validated by using different statistical metrics and compared with Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), and Gate Recurrent Unit Networks (GRU). A combination between Autoregressive Integrated Moving Average (ARIMA) and ANN as a hybrid technique to forecast the daily global solar radiation in three different cities in Morocco (Belmahdi et al. 2020a). The historical data has been transformed to no-stationary data and finding the optimum of ARIMA and ANN. The results show that by using the time series data, a significant ACF, PACF, and AIC criteria allowed a selection of the ARIMA (2. 1. 1), ARIMA (1.1.1) as adequate models of three sites. In another paper, the same author has applied the time series models to forecast one month of mean daily global solar radiation using Autoregressive Moving Average (ARMA) and ARIMA methods (Belmahdi et al. 2020b). The ARMA (2, 1) and ARIMA (0, 2, 1) are selected as optimal models due to the minimum values of the AIC and BIC criterion. In China three methods have been applied to forecast daily solar radiation using Support Vector Regression (SVR), extreme gradient boosting (XGBoost) and empirical method utilizing different input data (Fan et al. 2018). The results have shown that the XGBoost model is best suited to forecasting DSR in humid subtropical climates. In (Álvarez-Alvarado et al. 2021) a review has presented the hybrid techniques to forecast DSR using SVM and Search Optimization Algorithms (SOA). The papers explain and the implementation of different techniques such as ANN and SVM by using the SOA such as genetic algorithms (GA) and the particle swarm optimization algorithm (PSO) were used to optimize the prediction accuracy by searching the model parameters.

Forecasting SR seems to have become a popular topic. Such technology can help solar energy to be implemented into the grid, generating positive outcomes by improving the efficiency of the energy provided to the grid to minimize the price of accessories involved with the use of this product. Various researchers use optimization algorithms with forecasting techniques to improve the efficiency of the estimated results. A hybrid support vector regression (SVR) was boosted by the Krill Herd algorithm (KHA) to forecast the SR (Mohammadi and Aghashariatmadari 2020). Results showed that the test performance of SVR-KHA has higher accuracy and lowered error for all target data compared to the classic SVR. in the other paper, the GSR has predicted through multiple

meteorological parameters and defines important parameters based on an interpretation of synaptic weights in an artificial neural network model using a weight-approach relationship (Kumar and Kaur 2020). The paper considered that temperature plays a significant role followed by humidity and pressure compared to clearness index and precipitation which have the least effect on the prediction of solar radiations. Results also suggest that the ANN-based technique is more effective compared to the empirical model. Another paper has used the Improved Particle Swarm Optimization (IPSO) algorithm to optimize the SVR parameter for forecasting solar radiation (Ghazvinian et al. n.d.). An ensemble machine learning with square root regularization and intelligent optimization. The fundamental configuration process is based on ensemble learning, firefly algorithm and square root smoothly clipped absolute deviation (SRSCAD) using a random subspace (RS) method, which splits the original data into many covariate subspaces (Dong and He 2019).

This research is specifically intended to apply an intelligent paradigm for optimization models, which minimizes the forecasting error. The machine learning and time series models techniques can include the highest potential correlations and simplifications with our proposed study. The meteorological experimental data is obtained from the Mediterranean climate region at Tetouan, Morocco. Relying on the fact that forecasting of the global solar radiation (GSR) output can be enhanced with machine learning techniques and time series models based on meteorological data. Once the model is correctly trained using the optimization method, the model iteratively creates the forecasted GSR outputs. Four different techniques named ARIMA, FFNN-BP, K-NN, and SVM are applied to forecast the hourly global solar radiation (HGSR) output and compared with a persistence model. The aims are to minimize the forecasting errors depending on resolution, scale, and forecasting variables. The possible reason for using the ML and Time series model is mainly due to the high potential for modelling nonlinear dynamic systems with exogenous parameters depending on the times. In general, implementing methodological approaches to time series issues involves the development of characteristics that define the time series under consideration, especially time shifts (lags) of the series itself (Van Belle et al. 2021). In this regard, it is possible to use ML methods for generalized linear modelling of time series as well. However, the published results in (Makridakis et al. 2018) and (Makridakis et al. 2020) confirm that the statistical time series methods typically perform better ML techniques in a univariate setting. The utility of ML techniques in empirical forecasting configuration, however, benefits from the purpose is to be able to determine an arbitrary number of input elements. These ML strategies also depend on the principle of pre-processing to deal directly with this issue in order to define the most important input variables and their corresponding lag orders. On the other hand, the conventional procedures of

forecasting are focused on modelling the dynamics of the past time series and extrapolating it into the future. The application of these approaches of multivariate time series is detailed, with the most well-known exponential smoothing (ETS) and ARIMA methods (Van Van Belle et al. 2021). There are automated model selection algorithms focused on minimizing those knowledge requirements for both ARIMA and ETS.

The remainder of this paper is organized as follows: the "Material and method" section presents and illustrates the meteorological weather station data collection. The "Performance evaluation" section presents the forecasting methodology. The "Result and discussion" section indicated and resume the performance evaluation. The "Conclusion" section presents the simulation results and comparative analysis of the forecasted HGSR output. Finally, the conclusion has been drawn in Section 6.

Material and method

This section is divided into three subsections. Detailed information is presented about the study site region as first; the second stage gives a detail about the data collection and pre-processing. A brief summarization concerning the machine learning and time series models is given in the last subsection.

Study site

The kingdom of Morocco is a country situated between the continents of North of Africa and South of Europe. Its latitudes and longitudes are 3–40 and -7–5, respectively. Morocco has 63 provinces and its total area is 446 550 km^2. Morocco has a great solar energy potential estimated by the Moroccan Agency for Sustainable Energy (MASEN) to almost 2600 kWh/m^2/year. This case puts Morocco to an attractive location for solar energy investments and management.

In this paper, Tetouan city is selected for the prediction of hourly global solar radiation. The view of the province on the Morocco map is given in Fig. 1.

Data collection

The present paper focuses on the prediction of hourly global solar radiation data at one meteorological weather station (MWS) installed on the rooftop in the Faculty of Sciences, University Abdelmalek Essadi, Tetouan, Morocco (Fig. 2), and it covers the term from 1 January 2013 to 31 December 2015.

The data set contains records of hourly global solar radiation, maximum temperature (T_{max}), the difference of temperature (ΔT), temperature ratio (T_{ratio}) and average temperature ($T_{Average}$) were included in this study. In addition to these, the clearness index and the top of atmosphere (TOA) was calculated by using Eqs. (1)–(2). These parameters are considered as input
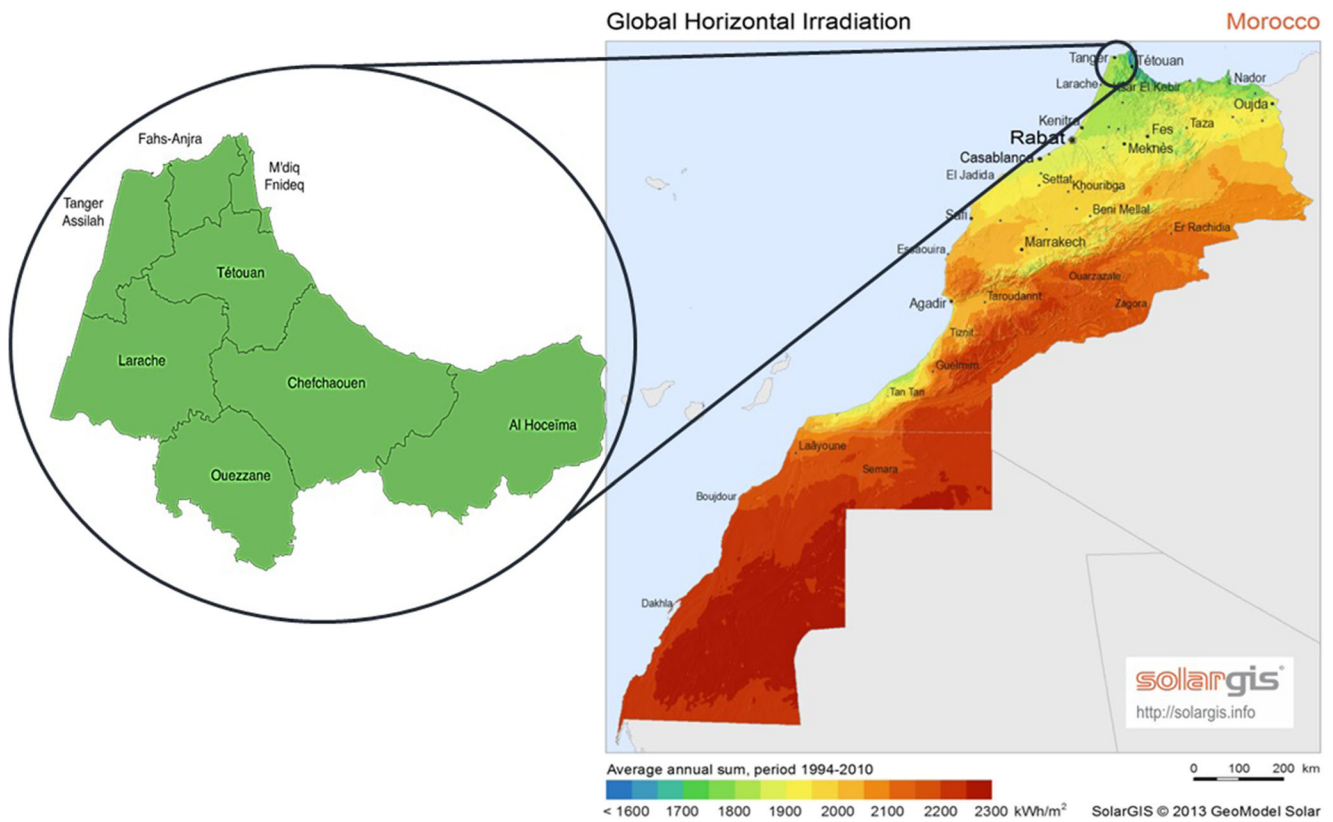
Fig. 1 Daily global solar radiation on surface horizontal distribution of the province of Tetouan-Tanger-Al Hoceima of Moroccan map

data to forecast the hourly global solar radiation output. Table 1 presents the measured data with its related notation.

The clearness index (k_t) is the ratio between the global irradiance at the surface, arriving on a horizontal plane, and the corresponding extra-terrestrial global irradiance on the horizontal planet. The following formulas of hourly clearness index are presented in equation (1):

Fig. 2 Meteorological local station in Abdelmalek Essaadi University, Faculty of Sciences

$$k_t = \frac{TOA}{HGSR} \tag{1}$$

Where the Top of Atmosphere (TOA) is computed by the following equation (2):

$$TOA = \int^{day} I_0 E_0 \sin(h) dt \tag{2}$$

With I_0 is the solar constant, h is the solar elevation and E_0 is present the correction of the Earth-Sun distance.

Selection of input parameters data by Pearson coefficient test

Since solar radiation is perturbed by many meteorological factors, it is essential to explore the relationships between solar radiation and the six pre-selected meteorological and astronomical parameters, for which an ideal forecast model can be established. The Pearson coefficient test is chosen to perform a correlation analysis by the Matlab environment. Table 1 and Appendix Fig. 11 show the values of the correlation coefficients.

The classification of input parameters shows that there is a strong correlation coefficient between global solar radiation and other parameters. The significantly correlated value must be positive and greater than 0.50 %. The classification selected is shown in Table 2.

Table 1 Input and output features and notation

Input's vector	Input Parameters	Notation	Correlation coefficient (%)
x_1	Clearness index	Kt	0.81
x_2	Delta-T	ΔT	0.67
x_3	Ratio temperature	T_{ratio}	0.68
x_4	TOA radiation	TOA	0.74
x_5	Maximum temperature	T_{max}	0.72
x_6	Average temperature.	$T_{Average}$	0.58
Output Vector	**Output Parameters**	**Notation**	
y	Hourly global solar radiation output	HGSR	

Forecasting methodology

In this paper, ML techniques and time-series models are employed to assess the proposed optimization method and select the most relevant inputs with the best configuration. The forecasting result will be compared with the persistence model as reference results. It should be mentioned that the persistence model estimates the coming output based on a single data point (i.e., 24 h from earlier HGSR outputs). The selected methods have been implemented and reported in the previous literature and have revealed the best performance characteristics. Each forecasted model is trained based on the previous inputs data namely K, the expression of the proposed method is structured as below:

$$y = f(\mathrm{x}, x_2, \cdots\cdots, x_{T-k+1}, \cdots\cdots, x_t) \qquad (3)$$

Where t is the current period (24 Hours).

The training phase method requires certain configuration parameters (i.e., lags), the lags denoted N which depend on the selected method. The Z parameters require in-depth knowledge of the system or can be found by trial and error. In this paper, we find the optimal values of K and Z for training the model by iteratively assessing the ability of the model until K and Z converge, as shown in Fig. 3A. The flowchart of the fitness function, noted by adjustment (model,X, K_i, Z_i), form the model using K and z and measures the errors resulting from the use of the forecasting model on the training data.

Once the model is correctly trained using the optimal values of z and K, the model iteratively creates the forecasted of the HGSR output of the next day, noted $\hat{y} = f(\hat{x}_1, \ldots, \hat{x}_{24})$, based on a combination of previously forecasted and realized outputs, as presented in Fig. 3B. In this regard, L denotes the lag parameter and is can be established from the configuration parameters z.

Table 2 The classification of six inputs data of Tetouan city

Study site	Parameters selected
Tetouan	Kt, TOA, T_{max} T_{ratio},ΔT, $T_{average}$

Persistence model

One of the simplest methods of forecasting the future behavior of a time series is the so-called persistence model. It implies that the future values of the time series are calculated by assuming that the conditions remain unchanged between the "current" time t and the future time $t + T_H$. For a stationary time-series the mean and variance of which do not change over time, a simple implementation of the persistence model is presented in Eq. (4):

$$y(t) = y(t + T_H) \qquad (4)$$

Where y is the earlier hourly global solar radiation output vector.

This technique simply uses the past recorded value to the model forecasting. The training phases of this technique is expressed by the following equation (5):

$$y = f(Persistence, y) \qquad (5)$$

The persistence technique counts on the past output of HGSR (L=Hours earlier), which is more appropriate when the fluctuation is smaller. In this context, L takes 24 Hours.
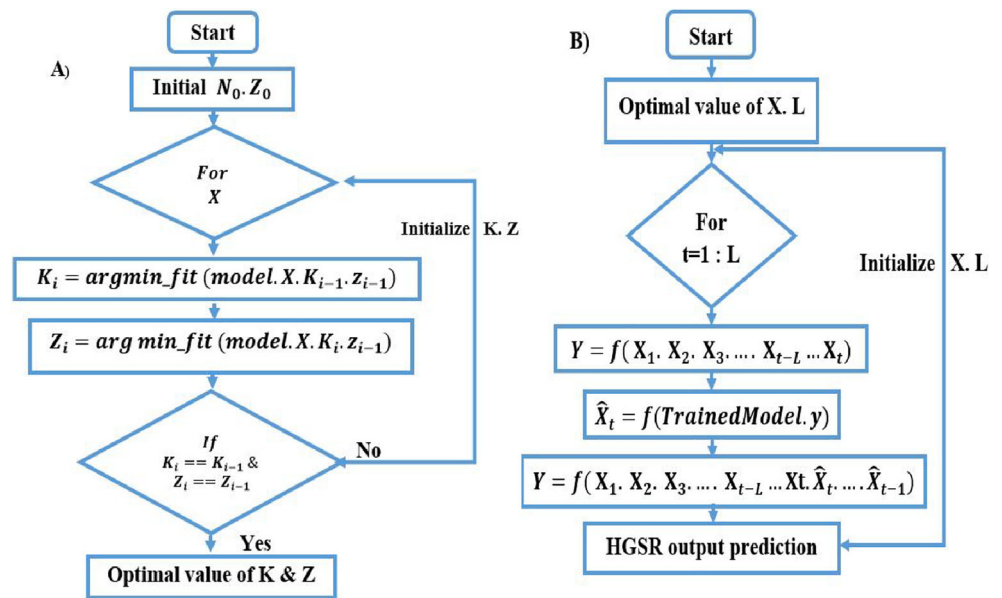
ARIMA model

In the class of forecasting model, the ARIMA model is an extension of the ARMA model. It is widely used for different modelling and forecasting applications with an acceptable level of forecast accuracy (Box et al. 2016) ARIMA for (p,d,q) can be defined as indicated in Eq. (6):

$$Y_T = \beta_1 Y_{t-1} + \beta_2 Y_{2-1} + \cdots \beta_p Y_{t-p} \varepsilon_t + \phi_1 \varepsilon_{t-1}$$
$$+ \phi_2 \varepsilon_{t-2} \phi \varepsilon_{t-2} + \phi_q \varepsilon_{t-q} + \tau \qquad (6)$$

Where Y_T present the forecasted HGSR $\beta_p Y_{t-p}$ is the linear combination Lag of Y, $\varnothing_q \varepsilon_{t-q}$ which presents the linear combination of Lagged forecasting error and τ is the constant.

We adopt in this study a Box-Jenkins method with a simple function to $z = f(p, d, q)$ obtain the optimal configuration parameters. Generally, all orders of ARIMA(p,d,q) are smaller

Fig. 3 A) Optimal forecasted model selection and B) forecasting HGSR output

than or equal to 2 (Wincek 1993). The trained ARIMA model is obtained by the given equation:

$$f[ARIMA(p,d,q),y] = \tau + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{q} \phi_i \varepsilon_{t-i} \qquad (7)$$

Where f is the prediction function, which includes the ARIMA model, τ is the constant. $\sum_{i=1}^{p} \beta_i\ Y_{i-1}$ and $\sum_{i=1}^{p} \beta_i\ Y_{i-1}$ are the amount of linear combination of Lag and linear combination of forecasting error respectively.

K-nearest-neighbours (k-NN) model

The K-NN rules are classified as a no-parametric classification algorithm in pattern recognition and are commonly used in many sciences due to their simplicity, feasibility, and intuitive existence. There are many interesting advantages of the k-NN algorithm. As a non-parametric classification system, a training procedure is not required for the k-NN algorithm. In particular, it needs no advanced knowledge of the statistical features of the training instances and can identify the query directly based on the information presented by the training collection (Li et al. 2008). This method is recognized as lazy learning because its training lagged during the execution (Korn et al. 2001). This classifier is also one of the most straightforward because the classification of the data sets is based on their class of nearest neighbours. The datasets are therefore assigned to the more similar class and k must be a positive integer. The value of k is generally small. When $k=1$, the data sets are essentially allocated to the class of its closest neighbor.

The performance of the k-NN classifier depends on the optimal distance used. There are four different types of

distance in k-NN, but in this analysis, the Euclidean distance will be applied because it is commonly used and defined by default. This distance is calculated between a test sample and the specified training samples. For example, let X_i be an input sample with r characteristics $(X_{i1}, X_{i2}, ..., X_{ir})$, N noted the total number of input samples $(i=1.2... N)$ and r the number total of $(j=1,2...,r)$. The Euclidean distance between the sample X_i and X_k $(k=1, 2...N)$ is identified as:

$$Y(X_i, X_k) = \sqrt{(X_{i1}-X_{k1})^2 + (X_{i2}-X_{k2})^2 + \cdots (X_{ir}-X_{kr})^2} \qquad (8)$$

The trained K-NN model is obtained by the given equation:

$$f(k-NN, y) = \sum_{i=1}^{r} \sqrt{(X_{ir}-X_{kr})^2} \qquad (9)$$

Feed-forward neural network with back propagation (FFNN-BP) model

One of the most widely used methods of forecasting solar energy production is artificial neural networks. The FFNN-BP is a relatively less complex neural network architecture. In FFNN-BP, information passes from the input layer to the output layer in the forward direction. The FFNN-BP network can be monolayer or multilayer, but information moves in only one direction. There is no feedback loop or cycle for processing information. In the Feed-Forward neural network (FFNN), the information reaches the output layer via the input and the hidden layer of the network. The FFNN-BP has also been used for several forecasting and pattern recognition applications (Mellit and Kalogirou 2008)(Malki et al. 2004). The relationship between the output $Y(t)$ and the inputs $(X_{t-1}; X_t$

$_{-2}\ldots X_{t-i})$ can be represented by the following expression equation:

$$Y(t) = s_1 \sum_{j=1}^{j} w_j s_2 \sum_{i=1}^{i} w_i x(t-i) \tag{10}$$

Where $Y(t)$ is an output from the network, $x(t-i)$ is the inputs to the network. Wj and Wi are the connection weights. S_1 and S_2 are the activation functions, the most commonly used function is a logistic sigmoid function given by the equation:

$$s(y) = \frac{1}{1 + e^{-x}} \tag{11}$$

The main control parameters of any FFNN are the weights. The process of estimating these parameters are known as training where optimal connection weights are determined by minimizing an objective function. The FFNN-BP forecasting model can be expressed as follows:

$$f(FFNN-BP, y) = s_1 \sum_{j=1}^{j} w_j s_2 \sum_{i=1}^{i} w_i x(t-i) \tag{12}$$

In the training step, z requires several configuration parameters. L represents the lag parameters; MN represents the maximum number of hidden neurons and FT is the Training Function. For each iteration, the model training is stopped when the error is linked below the RMSE%.

Support vector machine (SVM) model

The SVM has been introduced by Vapnik et al. based on statistical learning methodology and optimization theory as a rather strong and excellent pattern classification method (Takruri et al. 2020)(Awad et al. 2015). The main objective of SVM classification is to create a combination of simultaneous hyperplanes to maximize the minimum distance between two classes of samples. The SVM is considered an effective machine learning method and has attracted wide attention for its high performance on binary classification problems (Wu et al. 2021). The most important distinction between SVM and other methods of machine learning based on the Empirical Risk Minimization (ERM) paradigm is that not only analytical risk but also generalization flexibility is taken into account in the SVM models. In addition, the SVM describes the nonlinear response and one or multi-descriptors. The SVM output may be regular, binomial, or Poisson, as opposed to a simple linear or exponential regression, whose output has a normal distribution. In the SVM a connection objective function is introduced to the linear representation. SVM regularization is a type of shrinkage that applies a penalty function to minimize a model's difficulties, which can define significant descriptors, choose descriptors and generate fewer model formula coefficients. Based on the principle of Structured Risk

Minimization (SRM), SVMs seek to minimize an upper limit of the generalization errors instead of the empirical error as in other neural networks. Also, SVM models generate the regression function by applying a set of high dimensional linear functions. in order to solve the nonlinearity of the parameters X is introduced at first and mapped to m-dimensional feature space. A linear relationship was then established in the feature space, which obtains:

$$Y = \sum_{m=1}^{M} w_m \psi(X, X_m) + c \tag{13}$$

Where $\psi(X)$ is called the feature, which is a nonlinear mapping function from the input space the coefficients using kernel function (e.g linear function) w and c are the weight vector and the bias. The data are usually supposed to be zero and the variance term after preprocessing is negligible.

Performance evaluation

To achieve the performance accuracy of each model forecasting, analyze every parameter and case in a better way, it was necessary to use some statistical measured tools, which are very common in the kind of studies (Tian et al. 2016)(Emery et al. 2017). The calculated statistical errors are presented in Table 3.

In Table 3 υ_{max} is the number of observations at time t, $\widehat{X}_t$ and X_t are the forecasted and measured hourly global solar radiation at time t respectively.

These statistical indicators are commonly used to assess the obtained results and to compare the performance success of the proposed models used in the present study.

Result and discussion

The present study deals with the prediction of hourly global solar radiation on the horizontal surface of Tetouan city using five different forecasting models. To evaluate the adopted forecasting methodology, we applied multiple statistical performance indicators that are commonly used in the literature, are discussed. Tables 4, 5, 6, 7, 8, and 9 give multiple results of the selected models according to these metric indicators to assess the performance of these models in order to select the most appropriate one.

The measurements data cover the period from 1 January 2013 to 31 December 2015 and were carried out in a Meteorological station placed at the top of a building of the Faculty of Sciences of Abdelmalek Essaadi Tétouan University. These data are divided into the training and the testing set. The training set includes the cover the date from 1 January 2013 to 31 December 2014, which occupies 80% of

Table 3 Statistical metrics used in the study

Statistical performance indicator	Equation	Description	Performance information
Mean Bias Error (MBE)	$MBE = \frac{1}{v_{max}} \sum\limits_{t=1}^{v_{max}} (X_t - X_t)$	Denotes the arithmetical mean of the error and provides long-term status reports; as a condition, a small value is preferred.	The Positive value MBE means that the forecasted model is an overestimation, while the negative value means is an underestimation. Nevertheless, the low value of MBE indicates that the model is suitable and exhibits good performance.
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{v_{max}} \sum\limits_{t=1}^{v_{max}} (X_t - X_t)^2}$	used to assess if the existing measured and experimental values for the observational current values were adequate.	The small value of RMSE indicates the model performs better.
Normalized Root Mean Square Error (NRMSE)	$NRMSE = RMSE / \frac{1}{v_{max}} \sum\limits_{t=1}^{v_{max}} X_t$	it easier to compare datasets or models of various levels.	The low value of NRMSE indicates the model performs better.
Mean Absolute Percentage Error (MAPE)	$NRMSE = RMSE / \frac{1}{v_{max}} \sum\limits_{t=1}^{v_{max}} X_t$	It is a measure of the forecasting performance of a forecasting method.	MAPE is always positive
Test Statistic (TS)	$TS = \sqrt{(v_{max}-1) \frac{MBE^2}{SMSE^2 - MBE^2}}$	A Test Statistic is identified to evaluate behaviours, which would differentiate the zero from the dependent variable in the collected data.	The low value of TS indicates the model performs better.
Standard Deviation (σ)	$\sigma = \sqrt{\frac{v_{max}(MSE^2 - MBE)}{(v_{max}-1)}}$	Standard Deviation presents the calculation of the level difference or deformation of a range of values.	σ is always positive

the entire data. The testing set to cover the data from 1 January 2015 to 31 December 2015, which occupies the rest of the data (20%).

In this section, the results of each prediction model are studied individually by introducing several trial to select the performance of the appropriate model, therefore we have briefly presented the appropriate models in order to illustrate the prediction of 24 h of global solar radiation of two typical summer and winter days and compare it with the measurement data located at the local meteorological station of the Faculty of Sciences Abdelmalek Essadi.

The former author points out a multi-step optimization method:

- Collect the data.
- Initial and create the configuration parameters.

Table 4 Statistical tool's performance of ARIMA (p, d, q) models

Models	MBE	MBE%	RMSE	RMSE%	NRMSE	MAPE	TS	σ	σ%	AIC	BIC
ARIMA(1,0,0)	15,3876	32,252	25,951	6,815	1,1623	2,76	5,478	24,102	7,903	121.5222	122.7159
ARIMA(1,2,0)	16,0851	39,954	25,651	5,82	1,1724	2,77	8,728	20,771	8,903	114.2099	114.8016
ARIMA(1,0,2)	12,33	32,803	45,089	1,135	1,2956	2,86	6,658	19,032	2,701	132.6962	134.6857
ARIMA(2,1,0)	19,08	25,812	52,955	5,242	1,236	1,862	6,876	24,33	6,907	115.4747	116.2636
ARIMA(1,2,1)	18,992	77,443	15,711	8,379	1,0986	0,412	1,438	36,307	4,322	116.1671	116.956
ARIMA(2,1,1)	13,525	53,747	14,389	8,491	1,4848	2,414	6,876	24,5395	3,702	117.3987	118.3849
ARIMA(1,1,2)	13,418	35,875	14,387	7,811	1,4845	2,41	8,874	24,5364	2,877	111.756	113.2689
ARIMA(1,1,0)	15,183	25,652	14,453	4,638	1,4443	2,732	6,295	24,607	3,96	114.2345	115.1423
ARIMA(1,1,1)	13,0942	43,982	32,951	7,813	1,1916	1,76	9,778	27,722	2,083	115.7155	116.9259
ARIMA(2,0,0)	10,2341	34,079	34,713	7,356	0,0738	1,852	6,118	25,013	5,252	110.5625	111.7728
ARIMA(2,0,2)	11,234	43,072	54,151	6,311	0,197	0,782	4,379	46,811	6,206	109.4006	111.2161
ARIMA(2,2,0)	10,229	37,063	16,017	9,418	0,1374	1,102	9,677	41,017	7,914	116.1366	116.4544
ARIMA(2,2,2)	12,668	45,223	13,515	5,814	0,3457	0,972	9,419	9,954	6,07	114.3527	114.8293
ARIMA(2,2,1)	**9,538**	**22,923**	**22,014**	**4,774**	**0,0467**	**0,642**	**8,619**	**14,814**	**0,173**	**99.6896**	**101.2026**

Table 5 Statistical tool's performance of k-NN models

Model	K	MBE	MBE%	RMSE	RMSE%	NRMSE	MAPE	TS	σ	σ%
K-NN	1	29,766	7,102	22,69342	10,51552	9,53064	11,998	18,56077	29,884	0,87
	9	29,889	8,107	22,99866	10,65696	9,53064	11,998	19,61049	29,884	1,218
	14	30,012	8,308	23,09258	10,70048	9,53064	11,998	19,74366	30,008	1,479
	16	30,627	8,643	23,25694	10,77664	9,68436	11,998	21,4406	30,256	2,61
	10	**25,584**	**6,7**	**19,15381**	**8,87536**	**8,60832**	**11,998**	**18,33721**	**28,396**	**0,774**
	12	30,627	8,509	23,25107	10,77392	9,68436	11,998	19,75792	30,256	2,349
	40	30,873	9,313	23,79111	11,02416	9,68436	11,998	21,65979	30,38	3,393
	34	30,996	9,648	23,79698	11,02688	9,83808	11,998	21,69521	30,38	3,567
	28	31,242	9,916	23,88503	11,06768	10,29924	11,998	21,85437	31,496	4,785
	20	31,857	10,184	24,0083	11,1248	10,29924	11,998	23,6946	31,62	4,872
	24	33,087	10,586	25,28796	11,71776	12,2976	17,997	26,12524	44,64	5,307
	18	33,21	10,586	25,59907	11,86192	12,45132	17,997	26,17699	45,012	6,09
	8	33,948	10,787	25,72234	11,91904	12,75876	12,097	26,25979	45,508	8,526
	32	34,317	10,854	26,1215	12,104	12,91248	17,997	26,52222	45,632	8,874
	4	34,44	11,122	26,19781	12,13936	12,91248	11,999	26,52222	45,632	8,961
	28	38,622	11,256	29,56719	13,70064	13,8348	17,997	26,62618	47,244	10,701
	15	39,237	11,39	29,98396	13,89376	13,8348	17,997	26,67793	47,492	10,875
	18	39,36	11,926	30,16006	13,97536	13,8348	17,997	27,53215	47,492	10,962
	22	39,729	12,395	30,54748	14,15488	13,98852	17,997	27,54204	47,616	11,049

- Training the optimization method
- Validate and compute the statistical error measures.

Figure 4 presents the monthly average of daily global solar radiation on surface horizontal. It can be computed by the average of each month. The performances of the proposed models depend on the season of the year. For example, the monthly average of daily global solar radiation in winter is more fluctuate than in summer, which means that the global solar radiation is more difficult to forecast in winter. Solar radiation generates a satisfactory solar potential when the sunrise between 6:00 a.m. and 7:00 p.m., and it reaches to increase until its maximums (generally between 12:00 p.m. and 2:00

Table 6 FFNN-BP model architecture

Input combination	Activation function	Training Function (TF)	Number of neurons in the hidden layer (MN)	Final FFNN-BP Architectures
Kt, TOA, $T_{max}T_{ratio}$, ΔT, $T_{average}$	Sigmoid	**SCG**	10	6×10×1
			2	6×2×1
			15	6×15×1
			7	6×7×1
			22	6×12×1
			18	6×18×1
			23	6×23×1
			26	6×26×1
Kt, TOA, $T_{max}T_{ratio}$, ΔT, $T_{average}$		**LM**	10	6×10×1
			2	6×2×1
			15	6×15×1
			7	6×7×1
			22	6×12×1
			18	6×18×1
			23	6×23×1
			26	6×26×1

Table 7 Statistical tool's performance of FFNN-BP models

	FFNN-BP	MBE	MBE%	RMSE	RMSE%	NRMSE	MAPE	TS	σ	σ%
SCG	**FFNN-BP.10**	**26,42575**	**15,504**	**16,91**	**6,54**	**0,616132**	**1,80106**	**7,128387**	**24,05057**	**0,92137**
	FFNN-BP.2	26,86802	15,752	17,10	8,58	0,626233	1,80106	7,174755	24,25183	0,91938
	FFNN-BP.15	27,53143	16,228	17,41	6,93	0,636333	1,80106	7,273476	24,55372	0,91938
	FFNN-BP.7	27,86314	16,372	17,55	9,62	0,646434	1,80106	7,311654	24,65435	0,91838
	FFNN-BP.12	29,85336	17,848	18,55	5,83	0,676736	1,80106	7,551558	25,56002	0,91838
	FFNN-BP.18	29,85336	17,832	18,55	6,93	0,676736	1,80106	7,542486	25,66065	0,91739
	FFNN-BP.23	30,51677	17,184	22,95	4,95	0,838344	2,70159	5,351472	36,93121	0,90943
	FFNN-BP.26	33,61267	19,356	24,20	8,63	0,888847	2,70159	5,731425	37,93751	0,90644
LM	**FFNN-BP.10**	**23,88269**	**13,532**	**15,80**	**5,24**	**0,57573**	**1,80106**	**6,684111**	**23,24553**	**0,22037**
	FFNN-BP.2	29,07938	17,344	18,20	6,71	0,666635	1,80106	7,458444	25,35876	0,91838
	FFNN-BP.15	29,52166	16,428	22,55	7,15	0,828243	2,70159	5,212557	36,62932	0,90744
	FFNN-BP.7	29,85336	17,836	18,55	6,05	0,676736	1,80106	7,545636	25,66065	0,91739
	FFNN-BP.12	30,07451	16,841	22,90	5,83	0,838344	2,70159	5,278644	37,03184	0,90644
	FFNN-BP.18	30,84847	17,452	23,10	8,55	0,848445	2,70159	5,410314	37,03184	0,90843
	FFNN-BP.23	32,06472	18,256	23,65	6,38	0,868646	2,70159	5,537637	37,53499	0,90644
	FFNN-BP.26	34,49722	19,964	24,60	7,45	0,898947	2,70159	5,841801	38,2394	0,90545

p.m). In many cases, the solar radiation reaches a maximum between 10 a.m. and 2 p.m. After peak hour, it then decreases until sunset.

It is well-known that global solar radiation is related to multiple meteorological parameters.(Malvoni et al. 2016, 2017). The main meteorological data that influenced global solar radiation are presented in Table 1 and 2. The dynamic variations of each parameter make the forecasting methodology very difficult. These parameters are also significant for such issues as renewable energy. For example, the evaluation and interpretation of solar energy intensity are only possible in comparison with meteorological data acquired concurrently.

In this subsection, the training steps of our models ARIMA, FFNN-BP, K-NN, SVM and persistence model are tested and evaluated by minimizing the forecasting errors which depending on the hourly global solar radiation as output parameter and analysis of several performance metrics. In addition, the appropriate value of K and z indicate that's the

Table 8 Statistical tool's performance of SVM models

Model	Kernel Function	K	Z	MBE	MBE%	RMSE	RMSE(%	NRMSE	MAPE	TS	σ	σ%
SVM	**RBF with simplex Optimization method**	10	360	47,7034	25,437	20,52	7,19	1,411	6,4627	10,1088	49,044	7,664
		1	360	46,8485	25,687	20,83	8,16	1,428	5,3398	10,2753	49,446	8,816
		25	360	46,8485	25,659	20,43	2,92	1,394	6,5753	9,9505	49,044	7,125
		17	360	46,6775	25,632	20,38	5,21	1,394	5,4507	9,9036	49,044	7,045
		20	360	46,3356	25,381	20,25	2,91	1,394	5,5299	9,8663	48,776	6,565
		14	360	46,3356	25,298	19,82	2,43	1,365	5,5968	9,4586	48,374	5,776
		32	360	46,1646	24,047	16,92	4,55	1,156	3,6326	14,4592	34,438	8,288
		50	360	45,8226	23,631	16,92	3,45	1,156	4,7802	14,4661	34,438	8,464
		3	360	45,6516	22,851	16,74	5,46	1,139	4,8189	14,3934	34,170	6,848
		12	360	45,3097	22,824	16,69	6,46	1,139	5,0623	14,3554	34,173	6,192
		27	360	45,1387	22,629	16,69	5,46	1,139	4,6094	14,4572	34,036	6,464
		34	360	44,6258	22,574	16,42	4,80	1,122	4,4458	14,2071	33,768	5,552
		38	360	42,9161	22,575	16,29	4,91	1,122	4,7027	14,1451	33,543	5,328
		45	360	41,8901	22,101	15,48	5,44	1,054	4,4415	13,7292	32,428	2,656
		29	360	41,2062	22,017	15,33	6,55	1,054	5,0459	13,6190	32,162	2,592
		7	360	40,8642	20,182	15,21	7,55	1,037	5,0621	13,5631	32,026	2,032
		5	360	40,8642	20,155	15,21	9,55	1,037	3,8597	13,5357	32,026	1,445
		15	**360**	**34,7089**	**20,099**	**13,59**	**4,64**	**0,935**	**3,4211**	**9,30636**	**29,882**	**0,948**

Table 9 Statistical tool's performance of optimal models

Models	MBE	MBE%	RMSE	RMSE%	NRMSE	MAPE	TS	σ	σ%
ARIMA (2,2,1)	9,538	22,923	22,014	4,774	0,047	0,642	8,619	14,814	0,173
SVM (K=15, Z=360)	34,709	20,099	13,59	4,64	0,935	3,421	9,306	29,882	0,948
FFNN-BP.10	23,883	13,532	15,8	5,24	0,575	1,801	6,6841	23,245	0,221
k-NN (k=10)	25,584	6,7	19,154	8,875	8,608	11,998	18,337	28,396	0,774

forecasting models are performing well which provide a suitable platform to study the solar field

ARIMA model

Based on the proposed methodologies illustrated in Fig. 3 A) and B), several testing were found to select the appropriate and optimal values for the parameters K and $Z = \{p, d, q\}$, some values can be excluded and inferred from z. Figure 5 presents the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) plot of HGSR output which indicates a significant fluctuation of the daily cycle, while the PACF does not decay. It should be mentioned that the insolation decreases from summer to winter and increases from winter to summer. In this regard, we further assume that the differencing $d=2$ is non-negligible because the records data is taken from the middle of the summer, where the impact of difference is acceptable.

After several trials, we find that the $K=720$ (i.e., 30 days), $p=2$, $d=2$, and $q=1$ as optimal values for the ARIMA model.

Figure 6a and b illustrates the training set of the forecasting ARIMA (p,d,q) models measured by RMSE% and MAPE%. It should be mentioned that the lower values of these indicators yielded the highest forecasting accuracy with the optimal configuration of the ARIMA (p,d,q) models.
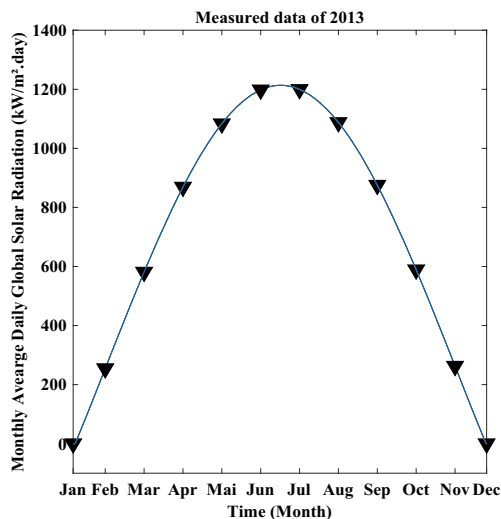
From the figures concerning changes in the order of p, d, and q, it appears clearly that the lower value of the median of NRMSE and MAPE are shown for ARIMA (2, 2, 1). The median value of RMSE and MAPE area 0, 0467 %, and 0,642 %. The presented values indicate that the ARIMA (2, 2, 1) performs better than other models and is the optimal one.

For more details, Table 4 presents the mean values of the statistical indicators for ARIMA (p, d, q). Based on this mean, it appears clearly that the ARIMA (2, 2, 1) shows the best and optimal performance than other models. As a result, the ARIMA (2, 2, 1) presents an adequate result and better than other ARIMA configurations.

k-nearest neighbour (k-NN) model

In the case of the k-NN model, the forecasting accuracy using various parameters of K to select the appropriate and optimal configuration of the proposed model by adopting the considered methodologies (Fig. 3 A and B). A different formulation for short-term global solar radiation output forecasting using the k-NN model based on meteorological data.

Figure 7 depicts the training set of the k-NN model measured by (a) NRMSE% and (b) MBE%; it is quite obvious that the NRMSE and MBE of k-NN have large values ($K=40$) and give the best performance accuracy. The mean, median and standard deviation (Sd) are 11.34 %, 10.3 %, and 1.896 % for NRMSE%, 9.68 %, 10.18 %, and 1.604 for MBE% respectively. Based on this result, it appears clearly that the optimal configuration of the k-NN model corresponding to the $K=10$.

Table 5 shows several results values of the k-NN model with multiple K. All of this K yielded the best performance, the lower value of the optimal configuration (K-NN with K=10) are 25.584 (6,7 %) for MBE%, 19.15381 (8,87536 %) for RMSE, 8.60832 for NRMSE, 11.998 for MAPE, 18.33721, 28.396, and 0.174 for T-statistic and σ % respectively.

Feed forward neural network (FFNN-BP) model

In the class of the FFNN-BP model, the best forecasting accuracy is obtained by using multiple training functions, different input combinations, several numbers of neurons in the hidden layer and the activation function is sigmoid. The
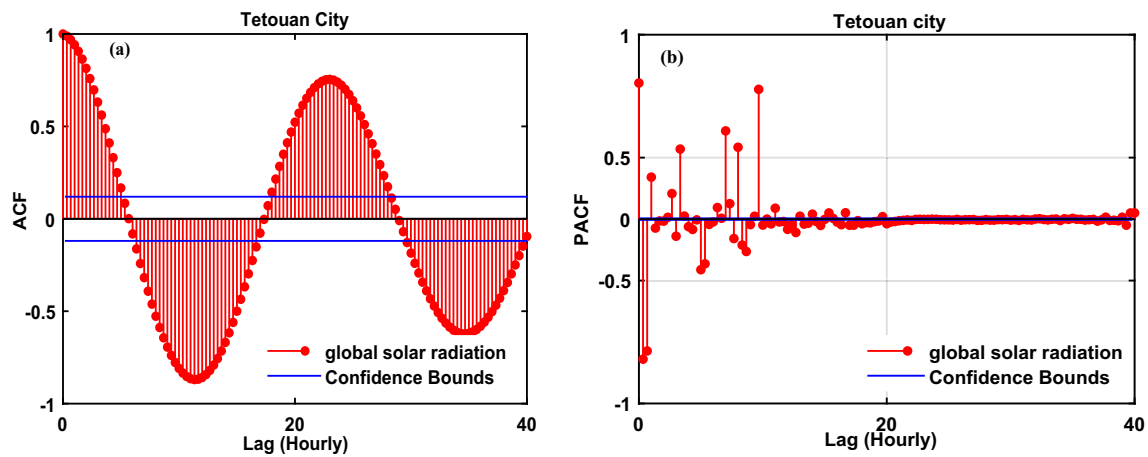

Fig. 4 Monthly average of daily global solar radiation of the year 2013

Fig. 5 (a) ACF and (b) PACF of HGSR output in Tetouan city, north of Morocco

accuracy of the model is assessed with various statistical indicators such as MBE% RMSE, NRMSE, MAPE, TS, and σ%. The forecasting accuracy of the optimized FFNN model is presented in Fig. 8, the training data set is larger than 26. The size of the training set that yielded presents the highest performance corresponding to the 10 neurons with the Gaussian activation function, 1000 epochs with the LM algorithm (the damping parameter $\mu = 0.05$), all other parameters are taken by default.

Figure 9 and Table 7 depict the training set with two algorithms named Levenberg-Marquardt(LM) and Scaled Conjugate Gradient (SCG) in terms of FFNN-BP measured by NRMSE% and MBE%. The range values of NRMSE and MBE are 3.41 %, 4.675%, 3.852 %, and 6.432 % for both LM and SCG, respectively. Based on these results, the LM algorithm performs better than the SCG training algorithm.

Table 7 presents the statistical measure of the two-training function with the optimal input combination and several hidden layers. It should be mentioned that the optimal configuration that they have a lower value of the forecasting accuracy. From the table, it appears clearly that the FFNN-BP.10 with the LM training function presents the best performance accuracy than other models. The lower values of MBE, NRMSE, RMSE, MAPE, and T-statistic are 23.88269, 0.57573, 15.8,

1.80106, and 6.684111, respectively. Based on this result the FFNN-BP.10 is the optimal configuration with the LM training algorithm and performs better than the SCG training algorithm.

Support vector machine (SVM) model

In the class of the Support Vector Machine (SVM) model, we use the radial basis function (RBF) with the kernel function and optimization method to select the optimal configuration that produces the smaller value of forecasting error measures. In Fig. 9a, the training set is larger than 50. The size of the training set that yielded the highest accuracy corresponding to 15 days (i.e., K=360), concerning the change of SVM parameters, the mean, median and Standard deviation (Sd) values of the NRMSE using the optimization method (Radial Basis Function with Kernel function method and optimization) are 1.193, 1.139 and 0.1585, respectively. While the range min and max value of the MBE% are 5.588, 20.1, and 25.69, respectively.

Table 8 describes the statistical error measure of several SVM forecasting s and multiple kernel function and optimization values. It appears clearly from the table that the lower value forecasting accuracy of SVM is K=360, the MBE%, RMSE%, NRMSE, MAPE, TS, and σ% for the optimal

Fig. 6 Influence of the learning set on the precision of the forecasting ARIMA models measured by RMSE (a) and MAPE (%) (b)
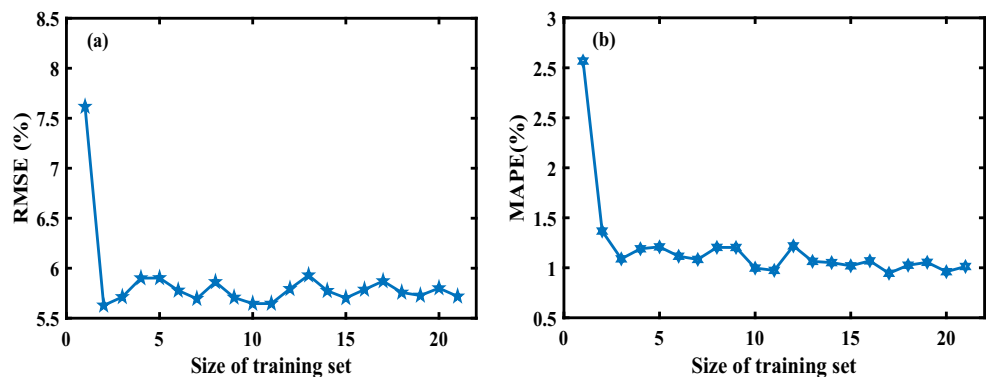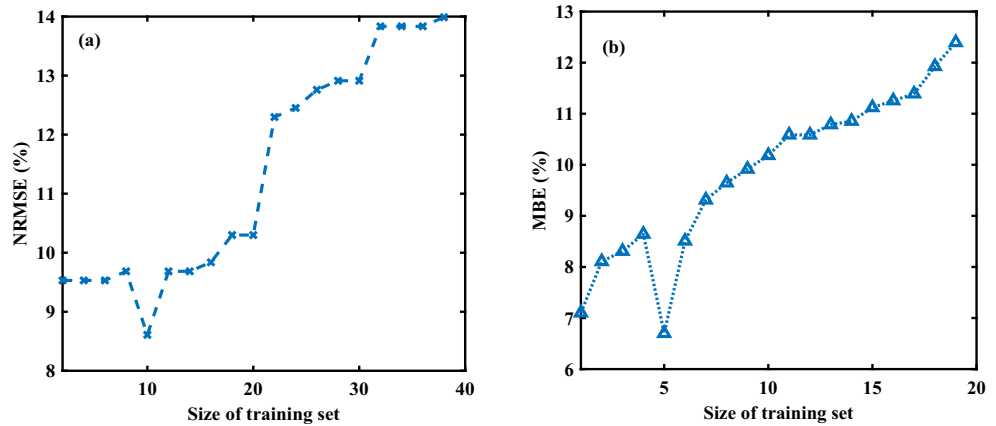
SVM configuration are 34.70 (20.09 %), 34.70894, 20.0994 (13.59 %), 4.64, 0.935, 3.42112, 9.30636, and 29.882 (0.848%), respectively.

Global solar radiation output forecasting

The objective of this subsection is to summarize the forecasting accuracy of the selected models using hourly global solar radiation output data from 15 January 2013 to 15 November 2015. Only the hourly global solar radiation forecasting is presented for the two typical summer and winter days. All these models used various inputs parameters (Kt, TOA, $T_{max}T_{ratio}$, ΔT, $T_{average}$). Table 9 shows the achieved model with the optimal configuration. The FFNN-BP.10 and ARIMA (2, 2, 1) models' forecasting s are outperformed ones than SVM (k=15, Z=360) and k-NN (K=10) models.

Figure 10A to F depicts the experimental and the forecasted HGSR output for the two typical summer and winter days (over 1000 w/m² to 1200 w/m² for summer, and 800 w/m² to 1000 W/m² for winter) from 3 years. In this case, the ARIMA (2, 2, 1) and FFNN-BP.10 forecasting are the appropriate, and forecasted curves fit the actual global solar radiation output curves well. However, some peaks and the turning point for the GSR output are not forecasted accurately. The SVM and k-NN models generate relatively less accurate

forecasting and the Persistence model is approximately close to the actual GSR output. The HGSR forecasted by several selected models is almost superimposed with those measured by the local metrological station from summer and winter days when the sun reaches the highest and lowest elevations.

In summer and winter days, it is also observed from the figures that the forecasted value of ARIMA (2, 2, 1) and FFNN-BP.10 are close to the experimental GSR. While the SVM and k-NN models show the underestimation cases for the forecasting GSR and the persistence model present slightly over forecasting and close to the corresponding GSR. Although the corresponding profile of time series data varied greatly due to varying weather conditions. The output results on the clear days provide a good platform to study the influence of cloudiness and deviation between the forecasted values by each forecasted model. In other to select the optimal and gives good approximations with the corresponding GSR.
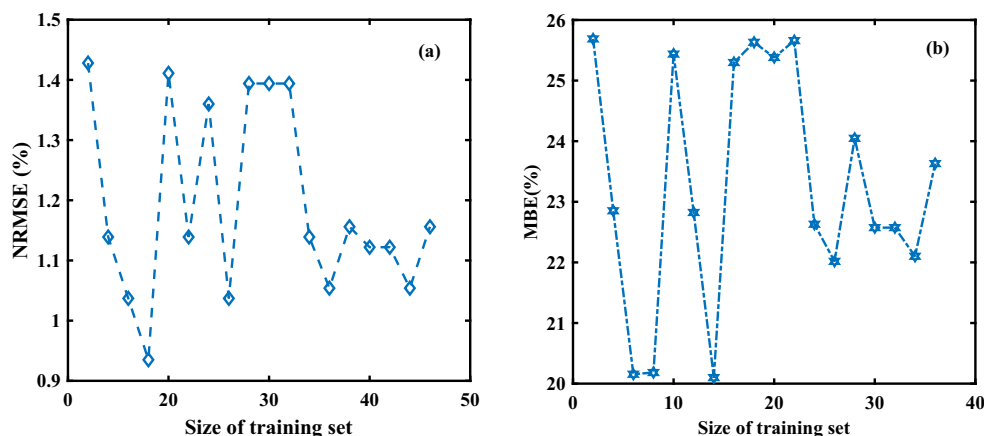
Conclusion

This paper presents the performance of five different machine learning (Persistence, k-NN, ARIMA, FFBP, and SVM) in the forecasting optimization of hourly global solar radiation. The current study considers multiple inputs data (clearness index,

Fig. 9 Training set on the forecasting error measures of (a) NRMSE and (b) MAPE of SVM model

TOA radiation, maximum, average, delta and ratio temperature are used as attributes) from local stations installed on the rooftop of the University Abdelmalek Essaadi of Tetouan. To assess the performance of the proposed models, six metrics (MBE (%), RMSE (%), NRMSE, MAPE (%), Ts and σ (%)) are discussed in this study. Then the following conclusions can be drawn based on the present investigation.
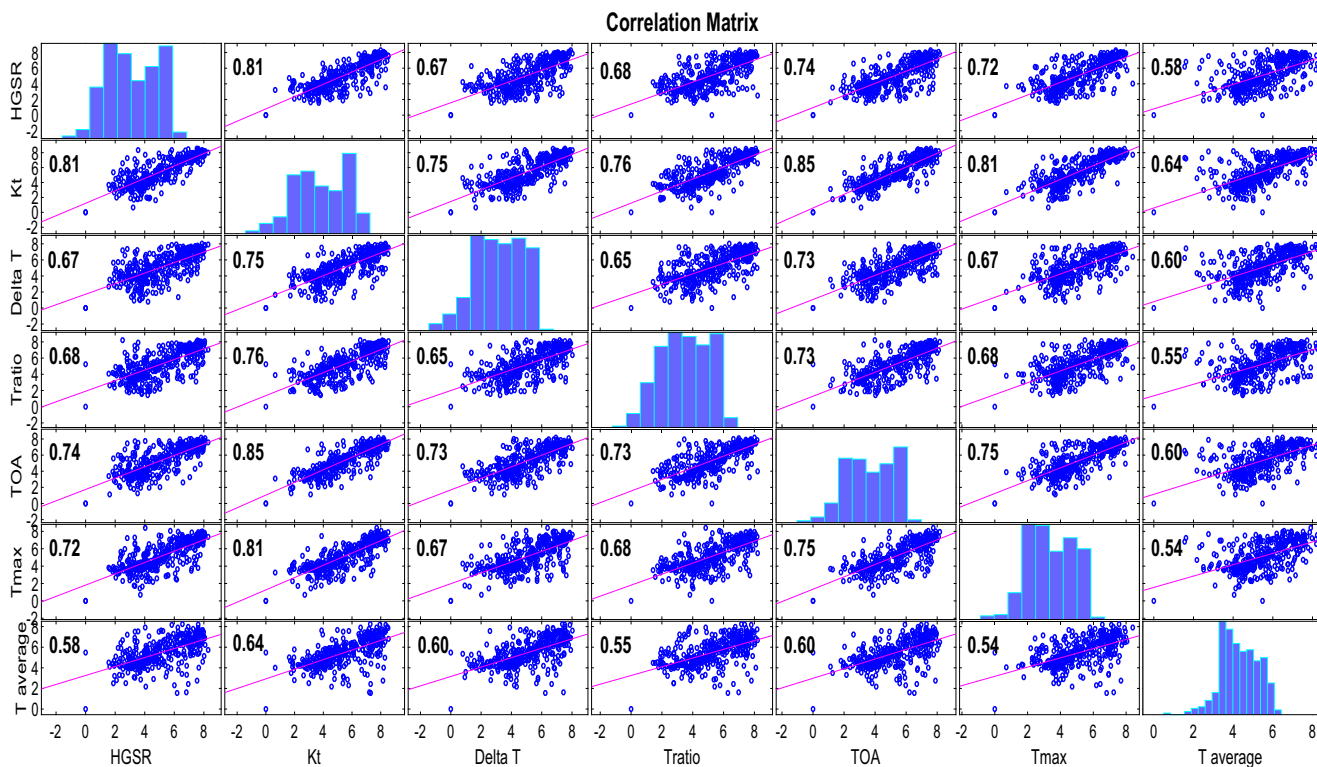
1- The forecasting methodology used in the study location has shown good results.

2- The RMSE (%) and MBE (%) values of several models employed in this study were computed to be mostly positive. The range value of the selected model measured by RMSE (%) and MBE (%) varied between 4.64 to 8.87 % and 6 to 22.93 %.

3- Based on all statistic metrics, the lower value of the selected model corresponds to the neural FFBP (6×10×1) in comparison with the others models. The appropriate one performs well and is close to the measured data.
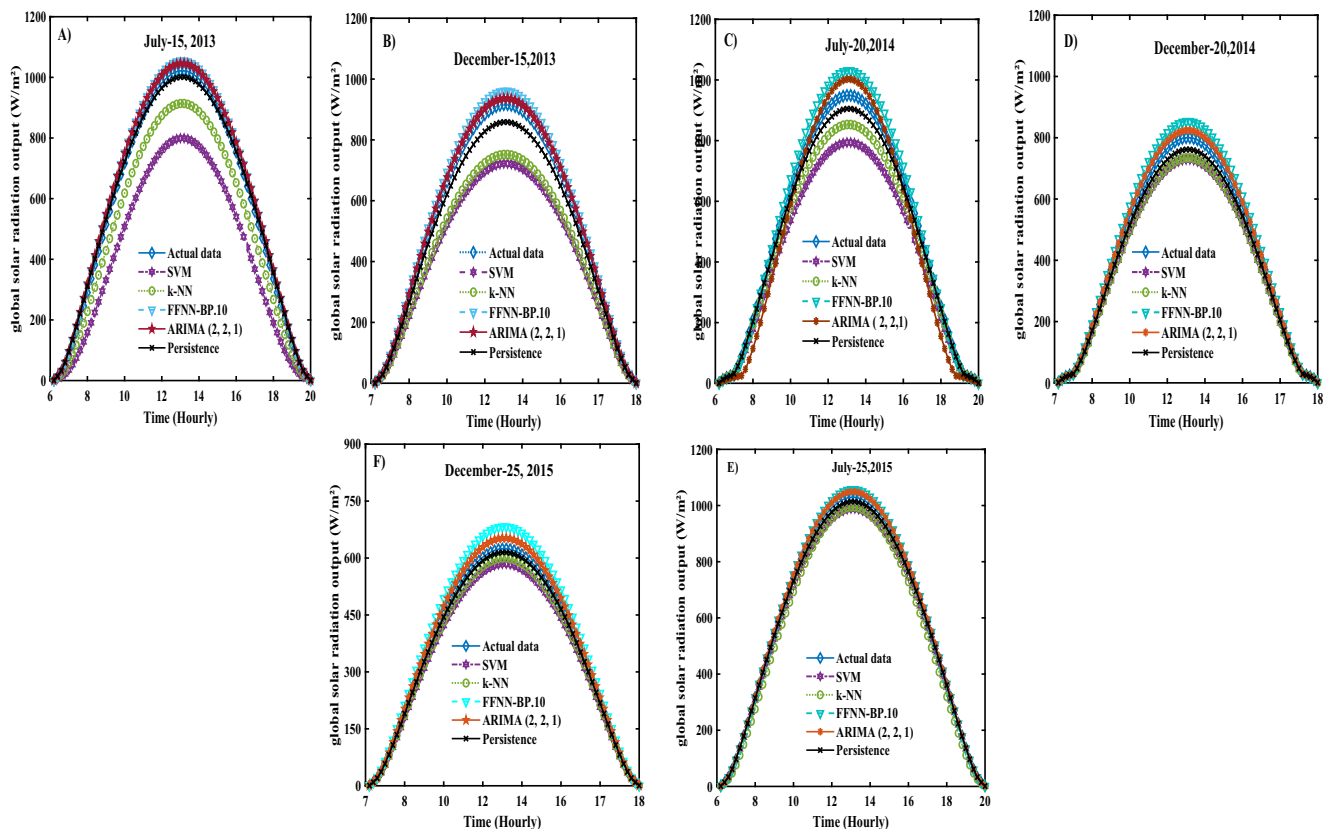
Fig. 10 Actual. vs forecasted HGSR output by using Persistence, ARIMA (2, 2, 1), FFNN-BP.10, k-NN, and SVM methods

Appendix

Fig. 11 Correlation coefficient of several inputs parameters.

Declarations

Ethics approval No ethical approval is foreseen since the research described by this paper only concerns the analysis of meteorological data that have been made public by the laboratory that collected them and that no participant has been involved not directly nor indirectly as the subject of the research itself.

Consent to participate No participant has been involved not directly nor indirectly in the research itself.

Consent for publication All the authors mentioned in the manuscript have agreed to authorship, read, and approved the manuscript, and given consent for submission and subsequent publication of the manuscript. All named authors have agreed with the order of authorship before submission, and all authors have agreed to the name of the corresponding author. The manuscript has not been published anywhere else.

Competing interests The authors declare no competing interests.

References

Al-Dahidi S, Louzazni M, Omran N (2020) A Local training strategy-based artificial neural network for predicting the power production of solar photovoltaic systems. IEEE Access 8:150262–150281

Almorox J, Hontoria C (2004) Global Solar Radiation Estimation Using Sunshine Duration in Spain. Energy Convers Manag 45(9):1529–1535

Álvarez-Alvarado JM, Ríos-Moreno JG, Obregón-Biosca SA, Ronquillo-Lomelí G, Ventura-Ramos E, Trejo-Perea M (2021) Hybrid Techniques to Predict Solar Radiation Using Support Vector Machine and Search Optimization Algorithms: A Review. Appl Sci 11(3):1044

Anon (2017) Photovoltaics for Sustainable Electricity and Buildings | Ali Sayigh | Springer

Awad M, Khanna R, Awad M, Khanna R (2015) Support vector machines for classification, pp. 39–66 in Efficient Learning Machines. Apress

Badescu V, Gueymard CA, Cheval S, Oprea C, Baciu M, Dumitrescu A, Iacobescu F, Milos I, Rada C (2013) Accuracy Analysis for Fifty-Four Clear-Sky Solar Radiation Models Using Routine Hourly Global Irradiance Measurements in Romania. Renew Energy 55:85–103

Bamehr S, Sabetghadam S (2021) Estimation of Global Solar Radiation Data Based on Satellite-Derived Atmospheric Parameters over the Urban Area of Mashhad, Iran. Environ Sci Pollut Res 28(6):7167–7179

Belmahdi B, Louzazni M, El Bouardi A (2020a) A Hybrid ARIMA–ANN Method to Forecast Daily Global Solar Radiation in Three Different Cities in Morocco. Eur Phys J Plus 135(11):925

Belmahdi B, Louzazni M, El Bouardi A (2020b) One Month-Ahead Forecasting of Mean Daily Global Solar Radiation Using Time Series Models. Optik 219:165207

Belmahdi, Brahim, Mohamed Louzazni, Mohamed Akour, Daniel Tudor Cotfas, Petru Adrian Cotfas, Abdelmajid El Bouardi (2021) Long-term global solar radiation prediction in 25 cities in Morocco using the FFNN-BP method. Front Energ Res 9:550

Box G, Jenkins G, Reinsel G, Ljung G (2016) 5th edition time series analysis forecasting and control. edited by Balding D, Cressie N, Fitzmaurice G, Givens G, Goldstein H, Molenberghs G, Scott D, Smith A, Tsay R, and Weisberg S. John Wiley & Sons

Cervone G, Clemente-Harding L, Alessandrini S, Monache LD (2017) Short-Term Photovoltaic Power Forecasting Using Artificial Neural Networks and an Analog Ensemble. Renew Energy 108:274–286

Chen JL, Xiao BB, Di Chen C, Wen ZF, Jiang Y, Lv MQ, Wu SJ, Li GS (2014) Estimation of Monthly-Mean Global Solar Radiation Using MODIS Atmospheric Product over China. J Atmos Sol Terr Phys 110–111:63–80

Deo RC, Şahin M (2017) Forecasting Long-Term Global Solar Radiation with an ANN Algorithm Coupled with Satellite-Derived (MODIS) Land Surface Temperature (LST) for Regional Locations in Queensland. Renew Sust Energ Rev 72:828–848

Di Piazza A, Di Piazza MC, La Tona G, Luna M (2021) An Artificial Neural Network-Based Forecasting Model of Energy-Related Time Series for Electrical Grid Management. Math Comput Simul 184:294–305

Dong Y, He J (2019) Global solar radiation forecasting using square root regularization-based ensemble. Math Problems Eng 2019

Emery C, Liu Z, Russell AG, Talat Odman M, Yarwood G, Kumar N (2017) Recommendations on Statistics and Benchmarks to Assess Photochemical Model Performance. J Air Waste Manag Assoc 67(5):582–598

Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, Lu X, Xiang Y (2018) Comparison of Support Vector Machine and Extreme Gradient Boosting for Predicting Daily Global Solar Radiation Using Temperature and Precipitation in Humid Subtropical Climates: A Case Study in China. Energy Convers Manag 164:102–111

Fan J, Wu L, Zhang F, Cai H, Ma X, Bai H (2019) Evaluation and Development of Empirical Models for Estimating Daily and Monthly Mean Daily Diffuse Horizontal Solar Radiation for Different Climatic Regions of China. Renew Sust Energ Rev 105:168–186

Ghazvinian H, Mousavi S-F, Karami H, Farzin S, Ehteram M, Hossain S, Fai CM, Hashim HB, Singh VP, Ros FC, Ahmed AN Haitham Abdulmohsin Afan, Sai Hin Lai, Ahmed El-Shafie (2019) Integrated support vector regression and an improved particle swarm optimization-based model for solar radiation prediction edited by Y. Li. PLoS One 14(5):e0217634

Khorasanizadeh H, Mohammadi K, Jalilvand M (2014) A Statistical Comparative Study to Demonstrate the Merit of Day of the Year-Based Models for Estimation of Horizontal Global Solar Radiation. Energy Convers Manag 87:37–47

Kılıç F, Yılmaz İH, Kaya Ö (2021) Adaptive Co-Optimization of Artificial Neural Networks Using Evolutionary Algorithm for Global Radiation Forecasting. Renew Energy 171:176–190

Korn F, Pagel BU, Faloutsos C (2001) On the 'Dimensionality curse' and the 'self-similarity blessing', pp. 96–111. In: IEEE Transactions on Knowledge and Data Engineering. Vol. 13

Kumar S, Kaur T (2020) Efficient Solar Radiation Estimation Using Cohesive Artificial Neural Network Technique with Optimal Synaptic Weights. Proc Inst Mech Eng Part A: Journal of Power and Energy 234(6):862–873

Kumar M, Kumar A (2017) Performance Assessment and Degradation Analysis of Solar Photovoltaic Technologies: A Review. Renew Sust Energ Rev 78(Supplement C):554–587

Li B, Chen YW, Chen YQ (2008) The Nearest Neighbor Algorithm of Local Probability Centers. IEEE Trans Syst Man Cybernet Part B: Cybernetics 38(1):141–154

Liu Y, Qin H, Zhang Z, Pei S, Wang C, Yu X, Jiang Z, Zhou J (2019) Ensemble Spatiotemporal Forecasting of Solar Irradiation Using Variational Bayesian Convolutional Gate Recurrent Unit Network. Appl Energy 253:113596

Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward edited by A. R. Hernandez Montoya. PLoS One 13(3):e0194889

Makridakis S, Spiliotis E, Assimakopoulos V (2020) The M4 Competition: 100,000 Time Series and 61 Forecasting Methods. Int J Forecast 36(1):54–74

Malki HA, Karayiannis NB, Balasubramanian M (2004) Short-Term Electric Power Load Forecasting Using Feedforward Neural Networks. Expert Syst 21(3):157–167

Malvoni M, De Giorgi MG, Congedo PM (2016) Data on Photovoltaic Power Forecasting Models for Mediterranean Climate. Data Brief 7:1639–1642

Malvoni M, De Giorgi MG, Congedo PM (2017) Forecasting of PV power generation using weather input data-preprocessing techniques, pp. 651–58 in Energy Procedia. Vol. 126. Elsevier Ltd

Mehleri ED, Zervas PL, Sarimveis H, Palyvos JA, Markatos NC (2010) A New Neural Network Model for Evaluating the Performance of Various Hourly Slope Irradiation Models: Implementation for the Region of Athens. Renew Energy 35(7):1357–1362

Mellit A, Kalogirou SA (2008) Artificial Intelligence Techniques for Photovoltaic Applications: A Review. Prog Energy Combust Sci 34(5):574–632

Mellit A, Sağlam S, Kalogirou SA (2013) Artificial Neural Network-Based Model for Estimating the Produced Power Ofaphotovoltaic Module. Renew Energy 60:71–78

Mohammadi B, Aghashariatmadari Z (2020) Estimation of Solar Radiation Using Neighboring Stations through Hybrid Support Vector Regression Boosted by Krill Herd Algorithm. Arab J Geosci 13(10):1–16

Notton G, Paoli C, Ivanova L, Vasileva S, Nivet ML (2013) Neural Network Approach to Estimate 10-Min Solar Global Irradiation Values on Tilted Planes. Renew Energy 50:576–584

Pazikadin AR, Rifai D, Ali K, Malik MZ, Abdalla AN, Faraj MA (2020) Solar Irradiance Measurement Instrumentation and Power Solar Generation Forecasting Based on Artificial Neural Networks

(ANN): A Review of Five Years Research Trend. Sci Total Environ 715:136848

Shiva Kumar B, Sudhakar K (2015) Performance Evaluation of 10 MW Grid Connected Solar Photovoltaic Power Plant in India. Energy Rep 1(Supplement C):184–192

Takruri M, Farhat M, Sunil S, Ramos-Hernanz JA, Barambones O (2020) Support Vector Machine for Photovoltaic System Efficiency Improvement. J Sustain Dev Energy Water Environ Syst N/A(N/A):0–0 8(3):441–451

Tian Y, Nearing GS, Peters-Lidard CD, Harrison KW, Tang L (2016) Performance Metrics, Error Modeling, and Uncertainty Quantification. Mon Weather Rev 144(2):607–613

Van Belle J, Guns T, Verbeke W (2021) Using Shared Sell-through Data to Forecast Wholesaler Demand in Multi-Echelon Supply Chains. Eur J Oper Res 288(2):466–479

Wincek MA (1993) Forecasting With Dynamic Regression Models. Technometrics 35(1):87–88

Wu L, Huang G, Fan J, Zhang F, Wang X, Zeng W (2019) Potential of Kernel-Based Nonlinear Extension of Arps Decline Model and Gradient Boosting with Categorical Features Support for Predicting Daily Global Solar Radiation in Humid Regions. Energy Convers Manag 183:280–295

Wu W, Xu Y, Pang X (2021) A Hybrid Acceleration Strategy for Nonparallel Support Vector Machine. Inf Sci 546:543–558