WAVELET-GAUSSIAN PROCESS REGRESSION MODEL FOR FORECASTING DAILY SOLAR RADIATION IN THE SAHARAN CLIMATE

Khaled Ferkous1,*, Farouk Chellali1, Abdalah Kouzou1 and Belgacem Bekkar2

1LAADI Laboratory, Department of Electrical Engineering, Zian Achour University, Djelfa, Algeria
2University of Ghardaia, Ghardaia, Algeria

*Corresponding author. E-mail: farkous.khaled@yahoo.com; k.farkouss@univ-djelfa.dz

Abstract

Forecasting solar radiation is fundamental to several domains related to renewable energy where several methods have been used to predict daily solar radiation, such as artificial intelligence and hybrid models. Recently, the Gaussian process regression (GPR) algorithm has been used successfully in remote sensing and Earth sciences. In this paper, a wavelet-coupled Gaussian process regression (W–GPR) model was proposed to predict the daily solar radiation received on a horizontal surface in Ghardaia (Algeria). For this purpose, 3 years of data (2013–15) have been used in model training while the data of 2016 were used to validate the model. In this work, different types of mother wavelets and different combinations of input data were evaluated based on the minimum air temperature, relative humidity and extraterrestrial solar radiation on a horizontal surface. The results demonstrated the effectiveness of the new hybrid W–GPR model compared with the classical GPR model in terms of root mean square error (RMSE), relative root mean square error (rRMSE), mean absolute error (MAE) and determination coefficient ($R^2$).

Graphical Abstract

Keywords: Gaussian process regression; wavelets; hybrid models; forecasting; solar radiation; solar measurements; Ghardaia
Introduction

In order to set up solar energy technology in the early stage, knowledge about the potential of solar energy is essential. Measured data sets are used to simulate the potential functioning of solar energy systems. The study of the potential of solar energy is the starting point of any investigation on solar energy projects. The precise knowledge of global solar radiation (GSR) available in the long-term data is necessary to design and implement good energy generation [1].

The insufficient number of meteorological stations in which GSR is recorded, as well as the lack of access to solar-radiation measurement stations, has encouraged researchers to develop models suitable for predicting solar radiation using available meteorological data (relative humidity, air temperature, wind speed and sunlight duration, etc.). However, empirical models based on the solar data are the most precise [2, 3]. Among the current models, the model proposed by Angstrom [4] gives a simple formula that determines the relationship between GSR and the duration of sunlight. Unfortunately, the performance of this model did not achieve the desired goal; therefore, models that are more accurate are needed.

Autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average models have been used by many researchers [5–8]. However, the limitation of time-series models in the use of high-resolution daily data encourages researchers to use other modelling techniques, including statistical learning methods such as the artificial neural network (ANN), support vector machine (SVM), neuro-fuzzy and Gaussian process regression (GPR). In [9], radial basis function (RBF) has been used to estimate the daily GSR in Medina (Saudi Arabia) and showed that the RBF was able to predict daily GSR at high resolution. The authors of [10] used multiple-layer perception to forecast GSR in 12 regions of Turkey. Two types of delay were used (weekly and annual). The results showed a better forecast with RMSE of 91 W/m². One type of delay was used (weekly and annual). The results showed a better forecast with RMSE of 91 W/m². One study [11] used daily solar radiation outside the atmosphere, cloud cover, day length and maximum and minimum air temperature to train four algorithms, deep learning (DL), kernel and Nearest-Neighbor (k-NN), ANN and SVM to predict GSR data in four Turkish provinces. The ANN algorithm confirmed its superiority ($R^2 = 93.2\%$, RMSE = 2.157 MJ/m²/day, rRMSE = 14.10%).

Some researchers have achieved considerable results using the extreme learning machine (ELM) algorithm due to its rapid implementation and ease of training. In [12], the kernel-based extreme learning machine (KELM) has been used to model the daily GSR. Many tests were performed and the results revealed that the basis of the KELM model $T_{\text{min}}$ and $T_{\text{max}}$ achieves higher precision, in particular when using $T_{\text{max}}$ and $T_{\text{max}} = T_{\text{min}}$ inputs ($R^2 = 90.57\%$, RMSE = 2.02 MJ/m²/day, mean absolute bias error (MBE) = 1.35 MJ/m², rRMSE = 11.25%).

A newly developed machine-learning model has been used to predict GSR in [13]. The model was compared with five other machine-learning models, namely the original ELM, SVM, generalized regression neural networks, M5 model tree and auto encoder, and climatic data taken during 1961–2016 from seven stations located on the Loess Plateau of China ($R^2 = 90.20\%$, MAE = 1.774 MJ/m²/day). Another study [14] developed an empirical model for accurately predicting solar radiation in China and then compared the new model with 19 locally calibrated empirical models, based on meteorological data during 1994–2016; the results showed prediction accuracy among the models, with an average MAE of 1.69 MJ/m²/day and rRMSE of 16.2%.

In this context, the GPR algorithm is a useful means [15]. The GPR algorithm has been used successfully in recent years in remote sensing and Earth sciences [16, 17]. In addition to good computational performance and stability, GPR is simpler and generally more robust than other statistical regression tools, requires a relatively small training data set, which can adopt highly flexible kernel functions, and has the ability to provide satisfactory predictions. In [18], a GPR model was used to predict daily GSR where the results showed better performance than conventional methods. In [19], the authors used data from four years (2005–08) to develop the GPR model. The results obtained showed that a GPR model based on sunshine duration, minimum air temperature and relative humidity gave the best results in terms of MBE, root mean square error (RMSE), relative mean square error (rRMSE) and correlation coefficient ($r$). The obtained values of these indicators were 0.67 and 1.15MJ/m², and 5.2% and 98.42%, respectively. The adaptive neuro-fuzzy inference system (ANFIS) model was used in [20] and trained using data of daily solar radiation recorded on a horizontal surface in the National Research Institute of Astronomy and Geophysics (NARIG), Helwan, Egypt for 10 years (1991–2000). The model gave a good accuracy of ~96% and a RMSE of <6%.

Several previous studies, for example [7], have shown that it is not wise to use deployment of exogenous meteorological data because they complicate the model and that the use of one variable achieves better results. On this basis, the hybrid model W–GPR is proposed in this study for the first time to predict GSR.

The proposed hybrid model incorporates wavelet transformation (WT) and GPR. GPR, which is based on a Bayesian framework, provides a probabilistic and non-parametric modelling approach. Compared with a persistence model, such as ANN or SVM, GPR is easy to implement, self-adaptive to estimate hyper-parameters and capable of providing uncertainty estimates. These characteristics make GPR an attractive tool in dealing with high dimensions, small samples and complex non-linear problems.

2 Methodology

2.1 Study region and meteorological data

The study area covers the province of Ghardaia (32.2°–32.82°N, and 3.7° and 4.5°E), which is located in the desert Plateau of China ($R^2 = 90.20\%$, MAE = 1.774 MJ/m²/day). Another study [14] developed an empirical model for accurately predicting solar radiation in China and then compared the new model with 19 locally calibrated empirical models, based on meteorological data during 1994–2016; the results showed prediction accuracy among the models, with an average MAE of 1.69 MJ/m²/day and rRMSE of 16.2%.

In this context, the GPR algorithm is a useful means [15]. The GPR algorithm has been used successfully in recent years in remote sensing and Earth sciences [16, 17]. In addition to good computational performance and stability, GPR is simpler and generally more robust than other statistical regression tools, requires a relatively small training data set, which can adopt highly flexible kernel functions, and has the ability to provide satisfactory predictions. In [18], a GPR model was used to predict daily GSR where the results showed better performance than conventional methods. In [19], the authors used data from four years (2005–08) to develop the GPR model. The results obtained showed that a GPR model based on sunshine duration, minimum air temperature and relative humidity gave the best results in terms of MBE, root mean square error (RMSE), relative mean square error (rRMSE) and correlation coefficient ($r$). The obtained values of these indicators were 0.67 and 1.15MJ/m², and 5.2% and 98.42%, respectively. The adaptive neuro-fuzzy inference system (ANFIS) model was used in [20] and trained using data of daily solar radiation recorded on a horizontal surface in the National Research Institute of Astronomy and Geophysics (NARIG), Helwan, Egypt for 10 years (1991–2000). The model gave a good accuracy of ~96% and a RMSE of <6%.

Several previous studies, for example [7], have shown that it is not wise to use deployment of exogenous meteorological data because they complicate the model and that the use of one variable achieves better results. On this basis, the hybrid model W–GPR is proposed in this study for the first time to predict GSR.

The proposed hybrid model incorporates wavelet transformation (WT) and GPR. GPR, which is based on a Bayesian framework, provides a probabilistic and non-parametric modelling approach. Compared with a persistence model, such as ANN or SVM, GPR is easy to implement, self-adaptive to estimate hyper-parameters and capable of providing uncertainty estimates. These characteristics make GPR an attractive tool in dealing with high dimensions, small samples and complex non-linear problems.
region of Algeria, and at an altitude of 450 m. Rainfall during the year is low in Ghardaia; it is classified as BWh Desert climate (or hot desert, according to the Köppen-Geiger classification) [21].

The average precipitation in Ghardaia is 68 mm/year while the average annual temperature is 21.0°C. The city has great solar potential throughout the year due to its location (the average daily solar radiation received is ~6000 Wh.m⁻² on a horizontal surface, as indicated in Fig. 1).

The authors have obtained the data sets from the Applied Research Unit for Renewable Energies (URAER). The data sets contain minimum air temperature (T_{min}), maximum air temperature (T_{max}), mean air temperature (T_{mean}), minimum relative humidity (RH_{min}), maximum relative humidity (RH_{max}) and mean relative humidity (RH_{mean}) including the total daily solar radiation on a horizontal surface. Data were measured and recorded at the research unit for the 4-year period from 1 January 2013 to 31 December 2016. The first 3 years of data (2013–15) are used as the training data set while the last year (2016) is used to test the different models.

### 2.2 Refinement of data

The accuracy of the models is greatly affected by the quality of the data used. It is preferable to perform the data-cleaning procedure for improving the quality of the data by filtering it from any error or doubt.

Malfunctions in the solar-radiation meters cause some discrepancies and anomalies in the values. Various techniques have been suggested in the literature to improve data quality by filtering errors. In this study, the following strategy was adopted:

(i) In order to identify inaccurate daily solar-radiation values, the daily clearness index K is calculated. The clearness index is a measure of the clearness of the atmosphere. It is the fraction of the solar radiation (H) that is transmitted through the atmosphere relative to the total solar radiation that strikes Earth (Hₐ). For each specific day, Hₐ for any geographical location is a constant value, irrespective of changes over the year. However, solar attenuation occurs as radiation passes through the atmosphere due to some atmospheric phenomena such as aerosol extinction, cloud extinction, Rayleigh scattering and so on. Therefore, in the available solar-radiation data, all values of H should be smaller than Hₐ, which means K is a dimensionless number between 0 and 1. It is defined as the surface radiation divided by the extraterrestrial radiation (Equation (1)).

\[
K = \frac{H}{H_a} \tag{1}
\]

Hₐ as the daily extraterrestrial radiation is computed as:

\[
H_a = \frac{24 \times 3600 \times 1000}{\pi} \left[ 1 + 0.033 \cos \left( \frac{284 \times \text{day}}{365} \right) \right] \left( \cos (\phi) \sin (\omega) + \frac{284}{365} \sin (\phi) \sin (\delta) \right) \tag{2}
\]

where G₂, φ, and day are the solar constant (1367 W/m²) [22], latitude of the location and day number of the year, respectively.

\[
\delta = 23.45 \frac{\pi}{180} \sin \left( \frac{360}{365} \times \text{day} + 284 \right) \tag{3}
\]

\[
\omega = \arccos \left( -\tan (\delta) \tan (\phi) \right) \tag{4}
\]

(ii) In this work, values that are outside the range of 0.015 < K < 1 have been deleted [23].

A month is deleted from the data set if the incorrect values are >5 days in this month; if the number is <5, the values are replaced by correct values based on the interpolation [24]. Of the 1500-day data set used in this study, we found 12 days that were outside the acceptable range.

In Fig. 2, the variables used as inputs are very appropriate, where the close correlation is between the individual variables T_{min}, T_{max}, T_{mean}, Hₐ and solar radiation. The lowest value of solar radiation is recorded during the month of December and the highest value is recorded in the month of July. In terms of relative humidity, this is the opposite of solar radiation, where the highest value is recorded during the month of July and the lowest value is recorded in the month of December.
The minimum temperature data and the minimum, medium and maximum humidity data are positively skewed with the average skewness factors of 0.02, 0.37, 0.24 and 0.43, respectively, while the maximum and medium temperature data and extraterrestrial solar radiation and daily incident solar radiation are negatively skewed with average skewness factors of −0.04, −0.02, −0.24 and −0.15.

Statistically, two numerical measures of shape (skewness and excess kurtosis) can be used to test for normality. For kurtosis, the general guideline is that if the number is greater than +1, the distribution is peaked. If the skewness is not close to zero, then the data set does not have a Gaussian distribution [25]. As expected, all the data can be considered as Gaussian in their distributional behaviours (Table 1).

The correlation between $H$ and its predictors is represented by the regression line that is included in Fig. 3, noting that the degree of correlation between $T_{\text{max}}$ and $H$ is high with the greatest value of the correlation $r = 0.70$.

### 2.3 GPR

A GPR is a non-parametric model based on the Gaussian probability distribution [26], which can be defined as a collection of random variables, of which any finite number GP has a joint Gaussian distribution [27, 28]. Thus, a GP is completely specified by its second-order statistics,

$$f(x) \sim \text{GP}(m(x), k(x, x'))$$

where $m(x)$ and $k(x, x')$ are the mean and covariance functions of a real process $f(x)$, respectively.

---

**Table 1**: The objective variable and input statistics for Ghardaia city

| Variable | Min. | Max. | Mean | Standard deviation | Skewness | Kurtosis |
|----------|------|------|------|--------------------|----------|----------|
| Ghardaia City URAER Station | Latitude 32.6° N, longitude 3.8° E, elevation 450 m | | | | | |
| $T_{\text{min}}$ (°) | 1.20 | 34.90 | 17.07 | 7.98 | 0.02 | 1.79 |
| $T_{\text{max}}$ (°) | 10.70 | 47.30 | 29.25 | 9.13 | −0.04 | 1.78 |
| $T_{\text{mean}}$ (°) | 7.20 | 40.40 | 22.91 | 8.62 | −0.02 | 1.74 |
| RH$_{\text{min}}$ (%) | 0.50 | 51.00 | 21.32 | 10.77 | 0.37 | 2.62 |
| RH$_{\text{mean}}$ (%) | 11.50 | 95.50 | 50.10 | 17.60 | 0.24 | 2.19 |
| RH$_{\text{max}}$ (%) | 41.00 | 98.50 | 91.47 | 13.45 | 0.43 | 2.24 |
| $H$ (MJ/m$^2$day) | 3.49 | 30.81 | 20.59 | 6.05 | −0.15 | 1.88 |
Suppose that there is a training set \( \{(x_i, y_i), i = 1, \ldots, n\} \), and the relationship between the \( p \)-dimensional predictor \( x \in \mathbb{R}^p \) and the target variable \( y \) is expressed as:

\[
y = f(x) + \varepsilon
\]

where \( \varepsilon \) is assumed to be an independent, identically distributed Gaussian noise, \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \).

The prior on the noisy observation becomes:

\[
\text{cov}(y) = K(X, X) + \sigma^2_n I
\]

where \( I \) denotes the identity matrix of size \( n \); the joint distribution of the observed target values and the function values at the test locations prior is given by:

\[
f_* \sim \mathcal{N}
\left(
0, \begin{bmatrix}
K(X, X) + \sigma^2_n I & K(X, X) \\
K(X, X) & K(X, X)
\end{bmatrix}
\right)
\]

where \( K(X, X)_{n \times n} \) denotes the covariance (or Gram) matrix between the test point and the training set, and also for different matrices \( K(X, X), K(X, X) \) and \( K(X, X) \).

The predictive equation for GPR becomes [29, 30]:

\[
f_\ast \mid X, y, X_\ast \sim \mathcal{N}(f_\ast, \text{cov}(f_\ast))
\]

where

\[
f_\ast = K(X_\ast, X) \left[ K(X, X) + \sigma^2_n I \right]^{-1} y
\]

(10)

\[
\text{cov}(f_\ast) = K(X_\ast, X) - f_\ast K(X, X_\ast)
\]

(11)

For a single test point \( X_\ast \), the predictive distribution is a Gaussian distribution with mean and covariance given by:

\[
f_\ast = k_\ast^T[K + \sigma^2_n I]^{-1} y
\]

(12)

\[
\text{cov}(f_\ast) = k_\ast^T k_\ast - f_\ast^T k_\ast
\]

(13)

where \( K = K(X, X) \) and \( k(X) = k \) denote the vector of covariance between the test point and the \( n \) training points. In Equations (12) and (13), \( (K + \sigma^2_n I)^{-1} \) can be calculated using Cholesky factorization [31].

### 2.4 SVM

This method for domain of regression and prediction problems [32] is gaining popularity due to its many attractive features and promising experimental performance. Unlike traditional learning algorithms (such as neural networks) that use the empirical risk minimization principle [33], SVM uses
the structural risk minimization principle that limits generalization errors [34], which reduces errors in training data, making it a powerful tool used in industrial applications in recent years [35]. In this research, the kernel cache and maximum iteration values are 200 and 5,000,000, respectively.

2.5 ANN

The ANN is a structure consisting of a network composed of process units (neurons). In order to create a relationship between inputs and outputs [36], it is similar to how the human brain processes information [37]. The parameters of neurons (weights and biases) must be calibrated using the data sets; this step is called the training process. However, before this step, the ANN structure (number of layers, number neurons in each layer, transfer function, etc.) must be fixed. A large network costs time in the training process and may create an overfitting situation [38], while a small network may be insufficient to reproduce the relationship of the studied system. In order to reach best performances, the user of the ANN should find the optimum network size.

2.6 Wavelet decomposition

The main motivation for using wavelet decomposition (WD) is the simple analysis of the series obtained. For many years, WD (or wavelet transform) has been mixed with time-series models as a preprocessing technique. WD uses a set of filters to decompose the original time series iteratively, so that separate forecasting models can be applied to each component.

The continuous wavelet transform (CWT) of a function \(f(t)\) compared with the mother wavelet \(\psi(t)\) can be written by the following integral [39]:

\[
F_w(a, \tau) = |a|^{-1/2} \int_{-\infty}^{+\infty} f(t) \psi^* \left( \frac{1 - \tau}{a} \right) dt
\]  

(14)

where (*) represents the operation of the complex conjugation, \(\tau\in\mathbb{R}\) is the translational value and \(a\in\mathbb{R}^+\) is the scaling coefficient. Unlike the Fourier transformation, the CWT has been discretized and is known as the discrete wavelet transform (DWT). The approach is an implementation of the wavelet transform by the scaling and translation of the wavelets in discrete time. In this case, the wavelets are given by

\[
\psi_{n,k}(t) = |a_0^n|^{-1/2} \psi \left( \frac{1 - k\tau_0a_0^n}{a_0^n} \right)
\]  

(15)

where \(n\) and \(k\) are integers \((a = a_0^n)\):

\[
\tau = k\tau_0a_0^n
\]  

(16)

More details on wavelet transform can be found in the literature [40, 41].

The main advantage of using the wavelet method is its robustness, as it does not include any potentially flawed parametric assumptions or test procedures. Another advantage of the wavelet method is that the decomposition of the signal makes it possible to study independently the behaviour of the signal at different time scales.

2.7 Structure of the hybrid model

The main goal of the wavelet–GPR model is the forecasting of daily solar radiation in a Saharan climate. To do this, first, predictor variable time series were decomposed to approximation subseries [42] (low frequency) and detailed subseries (D) (high frequency). In the next step, approximation and detailed subseries were used in the input matrix for the wavelet–GPR model. The optimal GPR parameters are represented in the flow chart of Fig. 4 based on the wavelet transform algorithm.

![Flow chart of the proposed model of the wavelet-Gaussian process regression W-GPR](https://academic.oup.com/ce/article-lookup/10.1093/comjnl/bxj051)
Fig. 5: DWTs of the inputs of the W–GPR model from 1 January 2013 to 31 December 2015 for Ghardaia Aero

The correlation coefficients between the output of the proposed model H and the inputs (DWTs of the predictor variable) are indicated in Table 2. As the table shows, the low-frequency component of the time series with wavelet-decomposed subseries of RH<sub>mean</sub> has the highest value of the correlation coefficient ($r_{cross} = -0.791$), followed by RH<sub>min</sub> ($r_{cross} = -0.782$).

### 3 Results

#### 3.1 Performance evaluation

To assess the success of this model, four different statistical measures have been adopted, which are used frequently in the literature. These metrics are RMSE, mean absolute error (MAE), coefficient of determination ($R^2$) and normalized root mean square error (rRMSE). The performance of the proposed models (W–GPR) are tested based on the following statistical measures for $N$ number of data points:

$$R^2 = \frac{\sum_{n=1}^{N} (H_n,Obs - \bar{H}_n,Obs) \cdot (H_n,Pred - \bar{H}_n,Pred)}{\sum_{n=1}^{N} (H_n,Obs - \bar{H}_n,Obs)^2 \cdot \sum_{n=1}^{N} (H_n,Pred - \bar{H}_n,Pred)^2}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N} (H_n,Obs - H_n,Pred)^2}{N}}$$

#### Table 2: The correlation coefficient $r$ between the output of the proposed model $H$ and the inputs (discrete wavelet coefficients of the predictor variable) after decomposition (discrete wavelet Transforms)

| Input variable | A3     | D1     | D2     | D3     |
|----------------|--------|--------|--------|--------|
| $T_{min}$      | 0.669  | -0.082 | -0.057 | -0.033 |
| $T_{max}$      | 0.708  | 0.038  | 0.006  | 0.024  |
| $T_{mean}$     | 0.700  | -0.005 | -0.008 | -0.006 |
| RH<sub>min</sub> | -0.781 | -0.054 | -0.037 | -0.049 |
| RH<sub>max</sub> | -0.772 | -0.008 | -0.036 | -0.042 |
| RH<sub>mean</sub> | -0.798 | -0.029 | -0.040 | -0.048 |

In the proposed approach, we consider each wavelet-decomposed signal in its original form to capture their random attributes and their physical structure; on this basis, we insert the entire substring into the W–GPR model.
Table 3: Effect of the wavelet type on model accuracy

| db4 | db8 | sym2 | Sym8 | coif1 | coif3 | coif5 | dmey3 |
|-----|-----|------|------|-------|-------|-------|-------|
| $R^2$ (%) | 0.907 | 0.912 | 0.910 | 0.916 | 0.920 | 0.921 | 0.908 | 0.920 |
| MAE (MJ/m²/day) | 1.727 | 1.709 | 2.332 | 1.657 | 1.642 | 1.625 | 3.339 | 1.615 |
| MSE (MJ/m²/day) | 6.887 | 6.496 | 8.548 | 6.241 | 5.940 | 5.852 | 18.219 | 5.912 |
| RMSE (MJ/m²/day) | 2.624 | 2.548 | 2.923 | 2.498 | 2.437 | 2.419 | 4.268 | 2.431 |
| rRMSE (%) | 12.82 | 12.44 | 14.27 | 12.19 | 11.89 | 11.81 | 2.083 | 11.80 |

Table 4: The accuracy of the W–GPR and GPR models from 1 January 2016 to 31 December 2016

| Input combinations | Wavelet-coupled (W–GPR) model | Classical (non-wavelet) GPR model |
|-------------------|-------------------------------|---------------------------------|
|                   | $R^2$ | MAE  | MSE  | RMSE | rRMSE | $R^2$ | MAE  | MSE  | RMSE | rRMSE |
|                   | (-)  | (MJ/m²) | (MJ/m²) | (MJ/m²) | (%)  | (-)  | (MJ/m²) | (MJ/m²) | (MJ/m²) | (%)  |
| G1                | 0.893 | 1.847 | 7.769 | 2.787 | 13.601 | 0.884 | 1.906 | 8.413 | 2.900 | 14.191 |
| M1                | 0.908 | 1.716 | 6.772 | 2.602 | 12.702 | 0.895 | 1.812 | 7.275 | 2.779 | 13.562 |
| M2                | 0.901 | 1.808 | 7.308 | 2.703 | 13.191 | 0.897 | 1.840 | 7.635 | 2.763 | 13.496 |
| M3                | 0.897 | 1.801 | 7.353 | 2.745 | 13.402 | 0.897 | 1.851 | 7.576 | 2.752 | 13.431 |
| M4                | 0.899 | 1.791 | 7.382 | 2.717 | 13.264 | 0.901 | 1.807 | 7.293 | 2.701 | 13.185 |
| M5                | 0.914 | 1.673 | 6.316 | 2.513 | 12.272 | 0.914 | 1.758 | 7.304 | 2.703 | 13.197 |
| M6                | 0.912 | 1.698 | 6.492 | 2.548 | 12.446 | 0.912 | 1.767 | 7.345 | 2.710 | 13.236 |
| M7                | 0.923 | 1.625 | 5.852 | 2.4191 | 11.817 | 0.906 | 1.727 | 6.887 | 2.624 | 12.812 |
| M8                | 0.922 | 1.648 | 5.983 | 2.446 | 11.945 | 0.903 | 1.749 | 7.119 | 2.668 | 13.020 |

Fig. 6: Comparison of the performance of W–GPR with GPR

$\text{MAE} = \frac{1}{N} \sum_{n=1}^{N} |(H_{n,\text{Pred}} - H_{n,\text{Obs}})|$

$r\text{RMSE} = 100 \times \sqrt{\frac{1}{N} \sum_{n=1}^{N} (H_{n,\text{Obs}} - H_{n,\text{Pred}})^2} \left(\frac{1}{\frac{1}{N} \sum_{n=1}^{N} (H_{n,\text{Obs}})^2}\right)$

where $H_{n,\text{obs}}$ represents the observed values, $R^2$ is the coefficient of determination, $H_{n,\text{pred}}$ represents the predicted values, $\overline{H}_{n,\text{obs}}$ represents the mean value of observations and $\overline{H}_{n,\text{pred}}$ represents the mean value of predictions.

According to [48], the performance of the model by considering $r\text{RMSE}$ is defined as:

- $r\text{RMSE} < 10\%$ the performance is Excellent.
- $10\% < r\text{RMSE} < 20\%$ the performance is Good.
- $20\% < r\text{RMSE} < 30\%$ the performance is Fair.
- $r\text{RMSE} > 30\%$ the performance is Poor.

3.2 Effect of wavelet type

The choice of the mother wavelet is one of the most important factors affecting the accuracy of the prediction models [49]. There are many wavelet families that can be used (Daubechies family (db), Symlets family (sym), Haar and Coiflets family (coif), Fejer-Korovkin family (fk), discrete approximation of Meyer family (dmey)). Application of a higher decomposition level can cause a slower training process and, in some cases, can reduce the accuracy of the models. In this work, after several experiments, wavelets were selected (db1, db4, db8, sym2, sym4, sym8, coif1, coif3, dmey3). According to Table 3, the coiflet wavelet of order 3 (abbreviated as coif3) is selected as the mother wavelet ($r\text{RMSE} = 2.419$ MJ/m²-day) This is indicated in bold text in Table 3.

In this paper, extraterrestrial solar radiation $H_0$ was used as the primary predictor variable, and then the input combinations were divided into three groups. The first group depends on air temperature as the second predictor variable. The second group depends on relative humidity as the second predictor variable. The third is a mixture of all the predictor variables:
G1 = W – GPR (H₀, T) 
{ M₁ = [H₀, Tₘₐₓ] 
  M₂ = [H₀, Tₘₐₓ, Tₘᵢₙ] 
  M₃ = [H₀, Tₘₐₓ, Tₘᵢₙ, Tₘₑᵃₙ] 

G₂ = W – GPR (H₀, RH) 
{ M₄ = [H₀, RHₘₑᵃₙ] 
  M₅ = [H₀, RHₘₑᵃₙ, RHₘᵢₙ] 
  M₆ = [H₀, RHₘₑᵃₙ, RHₘᵢₙ, RHₘₐₓ] 

G₃ = W – GPR (H₀, T, RH) 
{ M₇ = [H₀, Tₘₐₓ, RHₘₑᵃₙ, Tₘᵢₙ, RHₘᵢₙ] 
  M₈ = [H₀, Tₘₐₓ, RHₘₑᵃₙ, Tₘᵢₙ, RHₘᵢₙ, Tₘₑᵃₙ] 
  M₉ = [H₀, Tₘₐₓ, RHₘₑᵃₙ, Tₘᵢₙ, RHₘᵢₙ, RHₘₑᵃₙ] 
  M₁₀ = [All predictors variables] 

In Table 4, the performance of the model is evaluated. The effectiveness of each technique (wavelets or waves) depends on the input parameters chosen.

Fig. 7: Scatter plots of the forecasted value of solar radiation versus the measured, with different input combinations using the W–GPR model. Coefficient of determination is show in each panel

Fig. 8: The spread of prediction error Pe (MJ/m²day) for the W–GPR model compared with the GPR model
Careful examination of Table 4 shows that the best performance that can be obtained is to include all inputs except the $T_{\text{mean}}$ (M 2) for the first group, with $R^2 = 0.912\%$, $r\text{RMSE} = 12.36\%$ compared with $R^2 = 0.908\%$, $r\text{RMSE} = 12.70\%$ (M 3). In the second group performance is better in M 4 with $R^2 = 0.900\%$, $r\text{RMSE} = 13.19\%$ compared with $R^2 = 0.898\%$, $r\text{RMSE} = 13.26\%$ in M 5. However, for the third group, the best is M 9, verified by $R^2 = 0.921\%$, $r\text{RMSE} = 11.81\%$ in M 9 compared with $R^2 = 0.922\%$, $r\text{RMSE} = 11.94\%$ in M 10.

By comparing the forecasts of the three models, it can be seen that the third group exceeds the expectations of the other models. The coefficient of determination not only records the highest value (0.923\%) compared with the second group (0.906\%) and third group (0.913\%), but also marks the lowest value of rRMSE (11.81\%).

By comparing the forecasts of the three models, it can be seen that the third group exceeds the expectations of the other models. The coefficient of determination not only records the highest value (0.923\%) compared with the second group (0.906\%) and third group (0.913\%), but also marks the lowest value of rRMSE (11.81\%).

Table 5: Comparison of the overall performance of the models

| Model   | $R^2$ | MAE (MJ/m²/day) | RMSE (MJ/m²/day) | rRMSE (%) |
|---------|-------|-----------------|------------------|-----------|
| SVM     | 0.731 | 3.353           | 4.510            | 22.30     |
| GPR     | 0.900 | 1.796           | 2.739            | 13.371    |
| W–SVM   | 0.898 | 1.572           | 2.883            | 13.472    |
| W–GPR   | 0.923 | 1.625           | 2.4191           | 11.81     |

Careful examination of Table 4 shows that the best performance that can be obtained is to include all inputs except the $T_{\text{mean}}$ (M 2) for the first group, with $R^2 = 0.912\%$, rRMSE = 12.36\% compared with $R^2 = 0.908\%$, rRMSE = 12.70\% (M 3). In the second group performance is better in M 4 with $R^2 = 0.900\%$, rRMSE = 13.19\% compared with $R^2 = 0.898\%$, rRMSE = 13.26\% in M 5. However, for the third group, the best is M 9, verified by $R^2 = 0.921\%$, rRMSE = 11.81\% in M 9 compared with $R^2 = 0.922\%$, rRMSE = 11.94\% in M 10.

By comparing the forecasts of the three models, it can be seen that the third group exceeds the expectations of the other models. The coefficient of determination not only records the highest value (0.923\%) compared with the second group (0.906\%) and third group (0.913\%), but also marks the lowest value of rRMSE (11.81\%).

Table 5: Comparison of the overall performance of the models

| Model   | $R^2$ | MAE (MJ/m²/day) | RMSE (MJ/m²/day) | rRMSE (%) |
|---------|-------|-----------------|------------------|-----------|
| SVM     | 0.731 | 3.353           | 4.510            | 22.30     |
| GPR     | 0.900 | 1.796           | 2.739            | 13.371    |
| W–SVM   | 0.898 | 1.572           | 2.883            | 13.472    |
| W–GPR   | 0.923 | 1.625           | 2.4191           | 11.81     |

Fig. 9 traces the prediction error ($P_e = H_{n,\text{pred}} - H_{n,\text{obs}}$ (MJ/m²/day)) in the test. Input combinations of Table 4 are used. As shown in Fig. 8, a large correlation between the measured values of daily solar radiation and the predicted values obtained by the best model (M 9 with W–GPR) with the combination [$H_0$, $T_{\text{min}}$, $T_{\text{max}}$, $RH_{\text{min}}$, $RH_{\text{max}}$, $RH_{\text{mean}}$].

Visual representation of statistics has a big role in understanding the error range. To visualize the propagation of the prediction error $P_e$, we use a box plot; based on this representation, the prediction errors are large for the GPR model, although they are relatively lower for W–GPR(M 9), as shown in Fig. 8, which confirms the results obtained in Table 4.

To better see the occurrence of large prediction errors, the expected daily global solar irradiance and values during the test period (year 2016) are shown in Fig. 10. It also clearly shows that large changes in two types of successive days (sunny days to cloudy days) lead to significant errors in the forecast.

In order to compare the results of our study, some well-established models, including non-wavelet GPR, wavelet-coupled and non-wavelet ANN and SVM models are tested in the same way. From Table 5, the excellent predictability of the WT–GPR model is demonstrated by its superiority over both the WT–ANN and WT–SVM models (rRMSE = 11.81\%) is indicated in bold text in Table 5.

Table 6 shows a statistical comparison between the results of the current study and those of previous studies; the results were compared with those of studies using algorithms from the same class. Therefore, statistical metrics generally provided results that were close to each other for different algorithms. This makes it difficult to determine the most efficient prediction algorithm. For example, [11] studied daily global solar irradiance data. The authors found the highest value of $R^2$ with 0.932\% in their study. This value is the best in Table 6. On the other hand, the rRMSE value was calculated as 14.10\% in the study. This value is the second worst value in Table 6.

According to Table 6, there is not a single algorithm giving the best performance, but it varies from region to region.

---

Table 5: Comparison of the overall performance of the models

| Model   | $R^2$ | MAE (MJ/m²/day) | RMSE (MJ/m²/day) | rRMSE (%) |
|---------|-------|-----------------|------------------|-----------|
| SVM     | 0.731 | 3.353           | 4.510            | 22.30     |
| GPR     | 0.900 | 1.796           | 2.739            | 13.371    |
| W–SVM   | 0.898 | 1.572           | 2.883            | 13.472    |
| W–GPR   | 0.923 | 1.625           | 2.4191           | 11.81     |
region. In addition, the performances of the same algorithm may be very high for one region and low for another one.

4 Conclusion

The prediction of solar radiation is very important in the management of solar energy systems. This paper investigated the possibility of the prediction of daily solar radiation with high precision by using two models: wavelet–GPR and GPR. Due to the great impact of the quality of the data used on the accuracy of the models, we initially refined the data. The daily clearness index $K$ was calculated, so values outside the range $0.015 < K < 1$ were deleted. A W–GPR model incorporating the discrete WT algorithm for preprocessing the inputs has been designed. In order to evaluate the models and test their accuracy on the prediction, we used five statistical indicators. The results showed the significant effect of the wavelet type on the precision of the W–GPR models, where the wavelet type coif3 (Coiflet) had the highest precision and the combination of input variables $[H, T_{\text{min}}, T_{\text{max}}, RH_{\text{min}}, RH_{\text{max}}, RH_{\text{mean}}]$ offered great precision compared with the other proposed W–GPR models. To demonstrate the accuracy of the W–GPR model, its predictions were compared with the classical model (GPR, SVM, W–SVM) and other methods reviewed in the literature. The results showed a significant improvement in the performance of the W–GPR model appearing in the statistical indices $R^2 = 0.923$, $\text{MAE} = 1.625 \text{MJ/m}^2\text{day}$, $\text{MSE} = 5.852 \text{MJ/m}^2\text{day}$, $\text{RMSE} = 2.4191 \text{MJ/m}^2\text{day}$, $\text{rRMSE} = 11.81\%$. Finally, this model can be used to predict daily solar radiation in areas with a similar climate and can be further improved by introducing other variables, which should be the focus of our future work.

Conflict of interest statement

None declared.

References

[1] Zeng J, Qiao W. Short-term solar power prediction using a support vector machine. Renewable Energy, 2013, 52:118–127.
[2] Dorvlo A, Ampratwum DB. Technical note modelling of weather data for Oman. Renewable Energy, 1999, 17:421–428.
[3] Mohanty S, Patra PK, Sahoo SS. Prediction and application of solar radiation with soft computing over traditional and conventional approach: a comprehensive review. Renewable and Sustainable Energy Reviews, 2016, 56:778–796.
[4] Srivastava RC, Pandey H. Estimating Angstrom-Prescott coefficients for India and developing a correlation between sunshine hours and global solar radiation for India. Renewable Energy 2013, 1:11–7.

[5] Cragg C, Conway EM, Pearsall NM. Statistical investigation of the optimal averaging time for solar irradiance on horizontal and vertical surfaces in the UK. Solar Energy, 2000, 68:179–187.

[6] Irmak S, Asce M, Irmak A, et al. Predicting daily net radiation using minimum climatological data. Journal of Irrigation and Drainage Engineering, 2003, 129:256–269.

[7] Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry, 2019, 11:240.

[8] Paoli C, Voyant C, Muselli M, et al. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy, 2010, 84:2146–2160.

[9] Benghamem M, Mellite A. Radial Basis Function Network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, 2010, 35:3751–3762.

[10] Şenkal O, Kuleli T. Estimation of solar radiation over Turkey using artificial neural network and satellite data. Applied Energy, 2009, 86:1222–1228.

[11] Aşgülüt Ü, Gürel AE, Biçen Y. Prediction of global daily solar radiation using different machine learning algorithms: evaluation and comparison. Renewable and Sustainable Energy Reviews, 2021, 135:109393.

[12] Shahaboddin S, Kasra M, Hui-ling C, et al. Machine learning models to quantify and map daily global solar radiation and photovoltaic power. Renewable and Sustainable Energy Reviews, 2020, 118:109393.

[13] Peng Y, Hao W, Li H, et al. Daily global solar radiation prediction from air temperatures using kernel extreme learning machine: a case study for Iran. Journal of Atmospheric and Solar-Terrestrial Physics, 2015, 134:109–117.

[14] Gharbi R, Kurtagić D, Slišković D. Stream water temperature prediction based on Gaussian process regression. Expert Systems with Applications, 2013, 40:7407–7414.

[15] Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry, 2019, 11:240.

[16] Paoli C, Voyant C, Muselli M, et al. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy, 2010, 84:2146–2160.

[17] Benghamem M, Mellite A. Radial Basis Function Network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, 2010, 35:3751–3762.

[18] Şenkal O, Kuleli T. Estimation of solar radiation over Turkey using artificial neural network and satellite data. Applied Energy, 2009, 86:1222–1228.

[19] Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry, 2019, 11:240.

[20] Paoli C, Voyant C, Muselli M, et al. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy, 2010, 84:2146–2160.

[21] Benghamem M, Mellite A. Radial Basis Function Network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, 2010, 35:3751–3762.

[22] Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry, 2019, 11:240.

[23] Paoli C, Voyant C, Muselli M, et al. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy, 2010, 84:2146–2160.

[24] Benghamem M, Mellite A. Radial Basis Function Network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, 2010, 35:3751–3762.

[25] Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry, 2019, 11:240.

[26] Paoli C, Voyant C, Muselli M, et al. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy, 2010, 84:2146–2160.

[27] Benghamem M, Mellite A. Radial Basis Function Network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, 2010, 35:3751–3762.

[28] Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry, 2019, 11:240.

[29] Paoli C, Voyant C, Muselli M, et al. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy, 2010, 84:2146–2160.

[30] Benghamem M, Mellite A. Radial Basis Function Network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, 2010, 35:3751–3762.

[31] Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry, 2019, 11:240.
[45] Adamowski J, Sun K. Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. *Journal of Hydrology*, 2010, 390:85–91.

[46] Rajaee T, Mirbagheri SA, Nourani V, et al. Prediction of daily-suspended sediment load using wavelet and neuro-fuzzy combined model. *International Journal of Environmental Science & Technology*, 2010, 7:93–110.

[47] Kisi O, Cimen M. A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology*, 2011, 399:132–140.

[48] Mihoub R, Chabour N, Guermoui M. Modeling soil temperature based on Gaussian process regression in a semi-arid-climate, case study Ghardaia, Algeria. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, 2016, 2:397–403.

[49] Rafiee J, Tse PW, Harifi A, et al. A novel technique for selecting mother wavelet function using an intelligent fault diagnosis system. *Expert Systems with Applications*, 2009, 36:4862–4875.

[50] Deo RC, Wen X, Qi F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Applied Energy*, 2016, 168:568–593.

[51] Feng Y, Gong D, Zhang Q, et al. Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. *Energy Conversion and Management*, 2019, 198:111780.

[52] Quej VH, Almorox J, Arnaldo JA, et al. ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. *Journal of Atmospheric and Solar-Terrestrial Physics*, 2017, 155:62–70.