Eur. Phys. J. Plus (2018) **133**: 211
DOI 10.1140/epjp/i2018-12029-7

THE EUROPEAN PHYSICAL JOURNAL PLUS

Regular Article

# Estimation of the daily global solar radiation based on the Gaussian process regression methodology in the Saharan climate

Mawloud Guermoui[a], Kacem Gairaa, Abdelaziz Rabehi[b], Djelloul Djafer, and Said Benkaciali

Unité de Recherche Appliquée en Energies Renouvelables, URAER, Centre de Développement des Energies Renouvelables, CDER, 47133, Ghardaïa, Algeria

Received: 12 January 2018 / Revised: 26 March 2018
Published online: 4 June 2018


**Abstract.** Accurate estimation of solar radiation is the major concern in renewable energy applications. Over the past few years, a lot of machine learning paradigms have been proposed in order to improve the estimation performances, mostly based on artificial neural networks, fuzzy logic, support vector machine and adaptive neuro-fuzzy inference system. The aim of this work is the prediction of the daily global solar radiation, received on a horizontal surface through the Gaussian process regression (GPR) methodology. A case study of Ghardaïa region (Algeria) has been used in order to validate the above methodology. In fact, several combinations have been tested; it was found that, GPR-model based on sunshine duration, minimum air temperature and relative humidity gives the best results in term of mean absolute bias error (MBE), root mean square error (RMSE), relative mean square error (rRMSE), and correlation coefficient ($r$). The obtained values of these indicators are 0.67 MJ/m$^2$, 1.15 MJ/m$^2$, 5.2%, and 98.42%, respectively.

## 1 Introduction

Solar energy is the source of most energy available on the Earth. It is expected to play a very important role in the nearest future, particularly in the developing countries. In this context, several nations have adopted strategies and visions for the development of renewable sources, in order to meet their energy needs and supplies. We can cite the 2030 plan of Algeria and Tunisia for the production of 40% of electricity needs from renewables, the 2020 plan of Morocco and Egypt for a production of 42% and 20%, respectively, from renewable sources [1]. Indeed, knowing the available solar energy amount in a specific location is of primary importance for sizing solar systems. Although, solar radiation is relatively constant at the top of atmosphere of Earth's atmosphere, local climate influences can cause huge variations in available insolation on the Earth's surface. Controlling the random nature of energy sources such as solar radiation on the ground could allow power system operators to better integrate them. This calls for the estimation of available solar energy, which involves the prediction of solar energy. The best way to quantify the amount of solar energy at a given region is the installation of many measuring stations in several places on the region under consideration, and keeps to their maintenances and their records. Unfortunately, for many developing countries, there are no long time series of solar radiation data due to the limited number of measuring stations, cost and maintenance of devices. Nevertheless, the use of approaches and methods for forecasting and estimating the solar radiation amount and its characteristics will be paramount.

In the literature, some empirical formulas have been proposed by many researchers in the purpose to estimate the surface solar radiation, based on the available meteorological parameters such as relative humidity, air temperature, wind speed, sunshine duration, etc. The best accurate estimate is expected when employing sunshine duration based models [2–12]. The reference formula in this category is that proposed by Angström-Prescott [13], which establishes a simple relationship between global solar radiation and sunshine duration. Using air temperature and humidity as predictors is also widely investigated. Although, when high precisions are needed, these models become inappropriate for the reason that the majority of them are site dependent. Here, to overcome this shortcoming, the requirement for accurate approaches such as artificial intelligence (AI) or machine learning techniques becomes even more important.

[a] e-mail: gue.mouloud@gmail.com
[b] e-mail: rab_ehi@hotmail.fr

Among these approaches, we can cite artificial neural networks (ANN), support vector machine (SVM) [14], generalized fuzzy models and extreme learning machine (ELM) [15].

The artificial neural network model (ANN) is one of the earliest machine learning methods and it is considered in the first rank. It is processed by many authors for solar components (Global, Direct and diffuse) assessment. Dorvio *et al.* [16] have estimated solar radiation through radial basis functions (RBF) and multi-layer perceptron (MLP) networks, using data of eight stations from Oman. They concluded that both RBF and MLP models agree with the observed data, where an RMSE of ($0.83\,\mathrm{MJ/m^2/day}$) was achieved by using the RBF-model and ($1.01\,\mathrm{MJ/m^2/day}$) was found for the MLP-model. Banghanem *et al.* [17] used the RBF neural network to estimate daily global solar radiation in Al-Madina site (Saudi-Arabia). In their studies, they developed four RBF models using a set of meteorological parameters as inputs. It was found that the RBF model based on sunshine duration and air temperature gives high performances ($R^2 = 98.8\%$). The MLP model has been developed by Senkal *et al.* [18] for global solar radiation prediction on horizontal surfaces, in 12 areas in Turkey. Different inputs have been used (latitude, altitude, longitude, month, mean diffuse radiation and mean beam radiation). Accordingly, they showed that, the developed MLP model gives better predictions with an RMSE of $91\,\mathrm{W/m^2}$, compared to the physical methods (RMSE = $125\,\mathrm{W/m^2}$). Sozen *et al.* [19] proposed the ANN model for predicting global solar radiation based on geographical coordinates (latitude, longitude and altitude), meteorological data (sunshine duration and mean temperature) and the corresponding month of the year, from 17 stations in Turkey. The mean absolute percentage error (MAPE) was found to be less than 6.7% and the correlation coefficient reaches 99.89% for testing stations. Mellit *et al.* [20] proposed a hybrid model that combines Markov chain with neural network called ANN-MTM (Markov Transition Matrix). They used geographic coordinates as inputs to the proposed model while the output is the daily global solar radiation. The obtained results indicate an RMSE lower than 8% and the correlation coefficient is between 90 and 92%. Gani *et al.* [21] proposed the use of nonlinear autoregressive (NAR) model as a new methodology for daily global solar radiation prediction on a horizontal surface, over seven cities with different climate conditions in Iran. In their work, they used only the day of the year as the input to the proposed prediction model. They found that the adaptive neuro-fuzzy inference system (ANFIS) achieves high performance and outperforms the proposed NAR in term of statistical indicators. Zeng *et al.* [22] applied an RBF-model for the short-term prediction of hourly global solar radiation in three different regions. They used data recorded from 1991 to 2005 which include transitivity, relative humidity, wind speed, sky cover and hourly global solar radiation. In their work, they used a new concept based on 2-D representation of hourly solar radiation and instead of direct prediction of hourly global solar radiation they predicted firstly the transmissivity, which was then used to calculate solar radiation through extraterrestrial radiation. The achieved results were compared to those of linear regression models, local linear regression (LLR) and autoregressive (AR) model as a benchmark model. They found that the RBF model outperforms the AR and LLR methods in terms of the prediction accuracies. Mohamed *et al.* [23] proposed a simple model based on neural networks for forecasting the average daily global solar radiation, over five cities in Kuwait. The suggested MLP-model was tested based on five years (2007–2011) of measurements. The obtained results in terms of MAPE and correlation coefficients ($R^2$) are 85.6 and 94.75%, respectively.

Some researchers use kernel based machine learning methods, especially the support vector machine (SVM), for estimating solar radiation components. Bektas [24] investigated a least squares support vector machines (LS-SVM) method to forecast the next day global solar radiation, taking the daily mean temperature, the daily maximum temperature, sunshine duration and the solar insolation of the previous day as inputs. The attained results show that the LS-SVM is a good method for modeling solar insolation with a correlation coefficient of about 99.294%. Support vector machine regression was applied by Chen *et al.* [14] in order to estimate the global solar radiation components in China. Therefore, seven SVM models have been developed, using different combinations. All SVM models give high precisions and outperform empirical methods. The best model is observed when the sunshine duration is used as input. Lanre *et al.* [25] proposed a hybrid model that combined the fire-fly algorithm (FFA) with the SVM model for global solar radiation prediction, using meteorological parameters. The proposed FFA-SVM model was compared to the ANN and genetic algorithm (GA) approaches. They showed that the proposed FFA-SVM model outperforms the ANN and GA model in terms of statistical indicators (RMSE = $1.87\,\mathrm{MJ/m^2/day}$, $R^2 = 72.80\%$, $r = 85.32\%$, and MAPE = $1.52\,\mathrm{MJ/m^2/day}$).

Recently, another machine learning algorithm called extreme learning machine (ELM), has obtained considerable attention in the scientific area, due to its performance, fast training and easy implementation. Shamshirband *et al.* [26] used kernel extreme learning machine (KELM) for modeling daily global solar radiation. They developed three KELM models based on maximum and minimum air temperature. Several tests have been carried out, and the achieved results reveal that the KELM model achieves higher accuracy, especially when $T_{\max}$ and $T_{\max}$-$T_{\min}$ are used as inputs (MABE = $1.35\,\mathrm{MJ/m^2}$, RMSE = $2.02\,\mathrm{MJ/m^2}$, RRME = $11.25\,\mathrm{MJ/m^2}$, $R^2 = 90.57\%$).

New models are continuously proposed, such as combined approaches between different techniques in the goal to improve the precision. Cao *et al.* [27] proposed a new method for forecasting diffuse horizontal solar irradiation (DHI) using the combination of recurrent back propagation network and wavelet transform (WT), for Shanghai region (China). The obtained RMSE ($\mathrm{MJ/m^2}$) with and without wavelet analysis are 0.7193 and 2.8168, respectively. The ANFIS model is used by Rahoma *et al.* [28] aiming the prediction of the daily global radiation in Egypt. Data of ten years (1991–2000) have been employed for developing the model. The mentioned method is the combination of fuzzy logic and neural network techniques; the obtained results clarified that the fuzzy model gives better accuracy ($R^2 = 96\%, \mathrm{RMSE} < 6\%$).

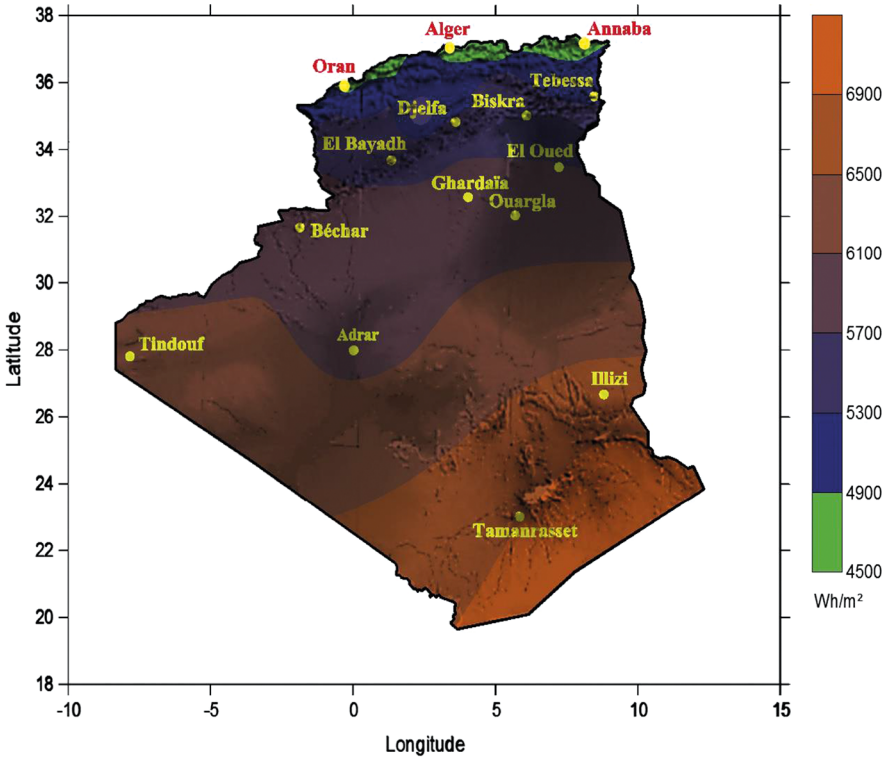


**Fig. 1.** Mean yearly global solar radiation received on horizontal surface [31].

Mellit *et al.* [29] have applied an adaptive ANFIS model for estimating the monthly mean clearness index ($K_t$) sequences and the daily global solar radiation in Algeria, based on some geographical and meteorological parameters. Also, a comparison to the ANN technique has been done using statistical indicators (RMSE and MAPE). The outcomes indicate that the ANFIS model gives better performance against ANN architectures (RBFN, MLP and RNN). Mohammadi *et al.* [30] have also used the ANFIS approach to identify the most relevant input parameters over three cities (Isfahan, Kerman and Tabass) in Iran. They considered in their study, three cases with 1, 2 and 3 inputs. In this respect, 9 ANFIS-models were developed to evaluate each case in terms of RMSE, MABE and correlation coefficient ($R^2$). For the first case, the sunshine hour and maximum possible sunshine hour have been identified as the relevant input parameters, while in the second case, they have recognized sunshine duration, extraterrestrial solar radiation and maximum possible sunshine hour as the appropriate inputs. In the last case and in addition to the inputs mentioned in the previous cases, they observed that maximum temperature ($T_{\max}$) can be added as an input, especially for Isfahan and Kerman sites.

In the present work, the purpose of our attempt is the estimation of the daily global solar radiation by means of the Gaussian process regression (GPR) model. We have developed this approach using data of four years, recorded from January, 1 (2005) to August, 31 (2008). Data were collected in Ghardaïa region, Algeria. The data used in this work in order to validate the proposed GPR model include daily global solar radiation (DHR), minimum air temperature ($T_{\min}$), maximum air temperature ($T_{\max}$), mean air temperature ($T_{\mathrm{mean}}$), minimum relative humidity ($RH_{\min}$), maximum relative humidity ($RH_{\max}$), mean relative humidity ($RH_{\mathrm{mean}}$), sunshine duration ($S$).

The paper is organized as follows: sect. 2 describes devices, data collection and site location. Section 3, deals with model performance. Results and discussion of the proposed model are covered in sect. 4. Finally, the last section is dedicated to the conclusion of the work.

## 2 Material and methods

### 2.1 Experimental setup

Algeria is located in the center of North Africa along the Mediterranean coast, between the latitudes of 19° and 38° North and longitudes of 8° West and 12° East. Its southern region comprises a large part of the Sahara (nearly 86% of the total area of the country, which represents 2048297 km$^2$). Its geographical position in the solar belt and favorable climatic conditions (abundant sunshine duration throughout the year), make Algeria a strategic actor in the field of solar technology; it is even considered one of the countries with the highest levels of solar radiation in the world with an estimated solar potential of more than 5 billion GWh/year [31].

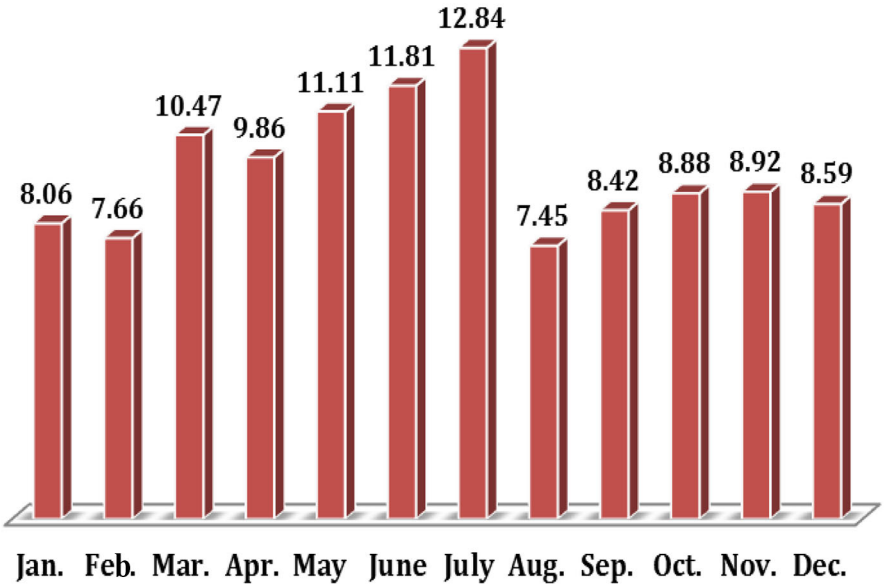


**Fig. 2.** Monthly mean sunshine duration in Ghardaïa site.


**Fig. 3.** Radiometric station: (1) Pyranometer for global solar radiation measurement on horizontal surface; (2) pyranometer for global solar radiation measurement on inclined surface; (3) pyranometer for diffuse solar radiation measurement on horizontal surface; (4) peryheliometer for DNI measurement.

The case of study considered in this paper is the Ghardaïa site. It is an arid area located in the south of Algeria about 600 (km) far from the capital city (latitude of 32.6° N, longitude of 3.8° E and altitude of 450 m). Ghardaïa is labeled by an exceptional sunshine duration (more than 3000 (hours/year)) and significant insolation where the mean daily global solar radiation received on a horizontal surface is about 6000 $Wh/m^2$ (figs. 1, 2).

The experimental data used in this work have been performed every 10 minutes by an EKO radiometric station with high precision, installed on the roof of the Applied Research Unit for Renewable Energies URAER (fig. 3). The station has the following parts:

- A fixed one consists of two EKO MS-64 pyranometers for the measurement of global solar radiation received on a horizontal plane and on an inclined surface at the latitude of the site.
- A moving part, able to track the sun from sunrise to sunset and it is equipped with an EKO MS-101D pyrheliometer, for measuring the direct normal irradiance DNI component.

**Table 1.** Technical specifications of used instruments.

| | Pyranometer MS-64 | Pyrheliometer MS-101D |
|---|---|---|
| Directional response | $< \pm 10\,\mathrm{W/m^2}$ | $< \pm 10\,\mathrm{W/m^2}$ |
| Temperature response | $< \pm 1\%$ | $< \pm 1\%$ |
| Non-linearity | $< \pm 0.2\%$ | $< \pm 0.2\%$ |
| Tilt response | $< \pm 0.2\%$ | – |
| Operating temperature range (°C) | −40–+80 | −20–+60 |
| Wavelength range (nm) | 305–2800 | 200–4000 |

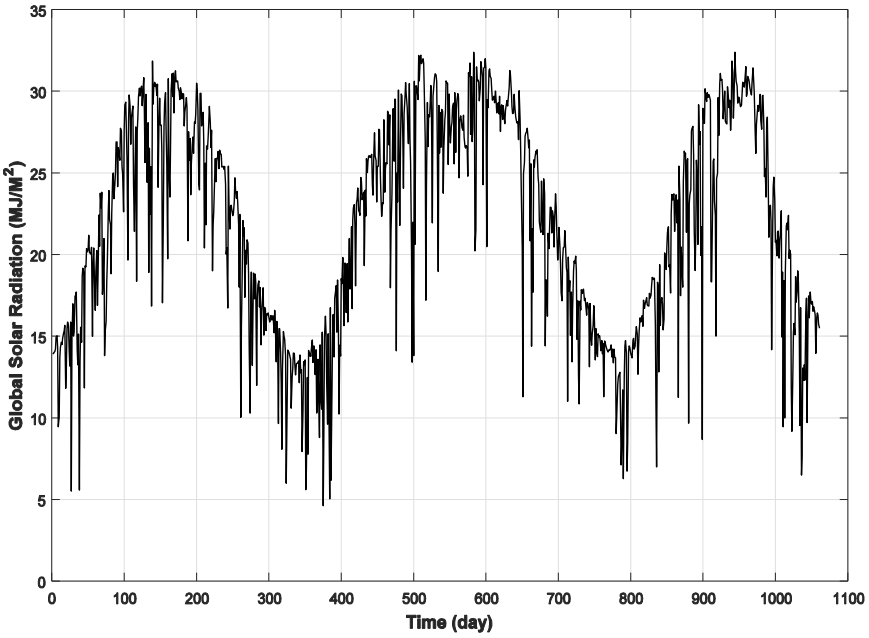


**Fig. 4.** Daily evolution of the global horizontal radiation.

Another EKO MS64 pyranometer is considered for the measurement of the diffuse component on a horizontal surface, equipped with a shadow band for hiding the radiant flux coming directly from the sun. Meteorological sensors are also included with this station for the measurement of the following parameters: air temperature, atmospheric humidity, wind speed and direction, atmospheric pressure (table 1).

More technical specifications of used radiometer are reported in table 1. We have used measurements of the daily global solar radiation recorded during four years (2005 to 2008), and the behavior of global solar radiation, sunshine duration and sunshine fraction are illustrated in figs. 4, 5 and 6, respectively.

## 3 Performance measurements

The performance of the GPR model is evaluated using different statistical indicators such as mean absolute bias error (MABE), root mean square error (RMSE), relative mean square error (rRMSE) and correlation coefficient ($r$) [23,30, 31].

The MABE gives the mean absolute value of the bias error. Its expression is given by:

$$\mathrm{MABE} = \frac{1}{n}\sum_{i=1}^{n}|\grave{H} - H|, \tag{1}$$

where $\grave{H}$ is the estimated value and $H$ is the measured one.

The RMSE represents the difference between predicted and measured values. In fact, RMSE identifies the model's accuracy. It is calculated by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\grave{H} - H)^2}. \tag{2}$$

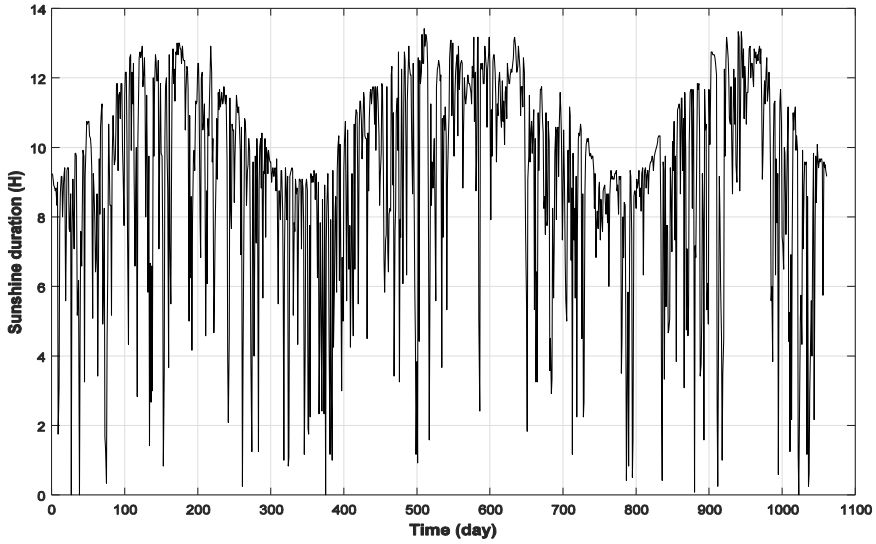


**Fig. 5.** Daily evolution of the sunshine duration.

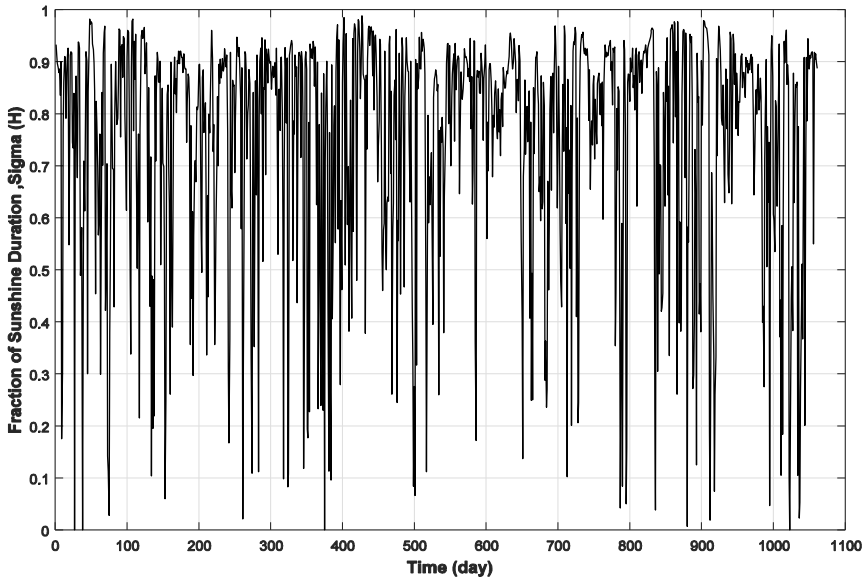


**Fig. 6.** Sunshine fraction.

The rRMSE is calculated, dividing the RMSE by the average of the measured data

$$\text{rRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\grave{H} - H)^2}}{\frac{1}{N}\sum_{i=1}^{n} H} \times 100. \tag{3}$$

The performance of the model is defined by the rRMSE range as follows [31,32]:

- Excellent, if rRMSE $< 10\%$.
- Good, if $10\% <$ rRMSE $< 20\%$.
- Fair, if $20\% <$ rRMSE $< 30\%$.
- Poor, if rRMSE $> 30\%$.

The $R^2$ and $r$ indicate the strength of a linear relationship between the measured and the predicted values. They are calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(\grave{H} - \acute{\overline{H}}) \cdot (H - \overline{H})}{\sqrt{\sum_{i=1}^{n}(H_P - \acute{\overline{H}}) \cdot \sum_{i=1}^{n}(H - \overline{H})}}. \tag{4}$$

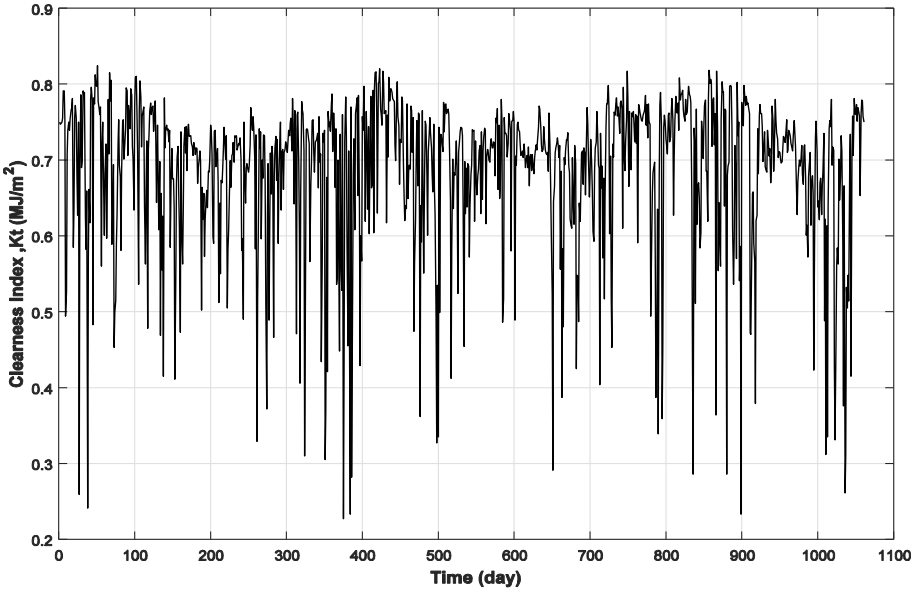


**Fig. 7.** Clearness index ($K_t$) evolution.

## 4 Results and discussion

As mentioned above, the main objective of this paper is the assessment of the GPR method for daily global solar radiation prediction, using the available meteorological parameters (*e.g.* air temperature, sunshine ratio, relative humidity). In this context, several GPR-models are developed based on different inputs. The target output is the clearness index ($K_t = \frac{\mathrm{GHR}}{\mathrm{GH_0R}}$) which is defined in the appendix. The behavior of $K_t$ is shown in fig. 7.

For the first model, air temperature, relative humidity and the mixture between them have been considered ($T_{\min}, T_{\max}, (T_{\max}\text{-}T_{\min}), T_{\mathrm{mean}}, RH_{\min}; RH_{\max}, (RHT_{\max}\text{-}RH_{\min}), RH_{\mathrm{mean}}$):

$$K_t = \mathrm{GPR}_1\{T, RH\}.$$

The second model uses only the sunshine ratio as input parameters

$$K_t = \mathrm{GPR}_2\{SS\}.$$

The third GPR model takes into account the sunshine ratio and air temperature as input parameters

$$K_t = \mathrm{GPR}_3\{T, SS\}.$$

The fourth model takes both relative humidity and sunshine ratio as inputs

$$K_t = \mathrm{GPR}_4\{RH, SS\}.$$

The last one considers all available data (sunshine duration, air temperature and relative humidity) as input parameters:

$$K_t = \mathrm{GPR}_5\{T, RH, SS\}.$$

The collected data are divided into two subsets: the first one is dedicated for training (540 samples) and the second subset is used for the test phase. The achieved statistical results are presented in table 2. In this stage, we can deduce the following observations:

- When several inputs have been used, the model accuracy has been influenced.
- The models which use air temperature as an input has reached high performance, compared to those based on relative humidity.
- The model performance doesn't improve significantly when combining air temperature and relative humidity.

**Table 2.** The studied GPR models with different input attributes.

| Models | Input parameters | MABE ($MJ/m^2$) | RMSE ($MJ/m^2$) | rRMSE | $r$ (%) |
|---|---|---|---|---|---|
| GPR models-1 | $(RH_{max}, RH_{min})$ | 2.01 | 3.028 | 13.69 | 88.70 |
| | $(RH_{max}, RH_{max}\text{-}RH_{min})$ | 2.00 | 3.036 | 13.72 | 88.64 |
| | $(RH_{min}, RH_{max}\text{-}RH_{min})$ | 1.99 | 2.93 | 13.26 | 89.38 |
| | $(RH_{max}, RH_{max}\text{-}RH_{min}, RH_{min})$ | 2.01 | 2.99 | 13.55 | 88.92 |
| | $(RH_{max}, RH_{mean})$ | 2.00 | 2.88 | 13.02 | 89.73 |
| | $(RH_{min}, RH_{mean})$ | 1.97 | 2.84 | 12.86 | 90.00 |
| | $(RH_{min}, RH_{mean}, RH_{max})$ | 2.01 | 2.95 | 13.36 | 89.23 |
| | $(RH_{max}, RH_{mean}, RH_{max}, RH_{max\text{-}min})$ | 2.00 | 2.94 | 13.28 | 89.34 |
| | $(T_{max}, T_{min})$ | 1.82 | 2.62 | 11.86 | 91.60 |
| | $(T_{min}, T_{max}\text{-}T_{min})$ | 1.80 | 2.60 | 11.75 | 91.75 |
| | $(T_{max}, T_{max}\text{-}T_{min})$ | 1.82 | 2.62 | 11.85 | 91.6 |
| | $(T_{max}, T_{min}, T_{max}\text{-}T_{min})$ | 1.82 | 2.62 | 11.84 | 91.62 |
| | $(T_{max}, T_{mean})$ | 2.04 | 2.84 | 12.85 | 90.14 |
| | $(T_{min}, T_{mean})$ | 1.78 | 2.58 | 11.67 | 91.86 |
| | $(T_{max}, T_{mean}, T_{max})$ | 1.78 | 2.57 | 11.63 | 91.92 |
| | $(T_{max}, T_{mean}, T_{max}, T_{max}\text{-}T_{min})$ | 1.77 | 2.56 | 11.58 | 91.99 |
| | $(T_{max}, RH_{max})$ | 1.99 | 2.744 | 12.40 | 90.94 |
| | $(T_{min}, RH_{min})$ | 1.97 | 2.756 | 12.46 | 90.75 |
| | $(T_{mean}, RH_{mean})$ | 2.03 | 2.81 | 12.73 | 90.31 |
| | $(T_{max}, T_{min}, RH_{max}, RH_{min})$ | 1.82 | 2.58 | 11.65 | 92.65 |
| | $(T_{max}, T_{min}, T_{max}\text{-}T_{min}, TRH_{max}, RH_{min}, RH_{max}\text{-}RH_{min})$ | 1.82 | 2.60 | 11.77 | 91.81 |
| | $(T_{max}, T_{min}, T_{mean}, T_{max}\text{-}T_{min}, TRH_{max}, RH_{min}, RH_{mean}, RH_{max}\text{-}RH_{min})$ | 1.81 | 2.60 | 11.73 | 91.85 |
| | $(T_{max}, T_{min}, T_{mean}, TRH_{max}, RH_{min}, RH_{mean})$ | 1.83 | 2.58 | 11.66 | 91.97 |
| | $(T_{max}, T_{mean}, TRH_{max}, RH_{mean})$ | 1.98 | 2.73 | 12.35 | 91 |
| | $(T_{min}, T_{mean}, RH_{min}, RH_{mean})$ | 1.85 | 2.60 | 11.76 | 91.74 |
| GPR models-2 | $(SS)$ | 0.80 | 1.24 | 5.59 | 98.23 |
| GPR models-3 | $(SS, T_{max}, T_{min})$ | 0.73 | 1.196 | 5.40 | 98.32 |
| | $(SS, T_{max}, T_{min}, T_{mean})$ | 0.726 | 1.179 | 5.33 | 98.36 |
| | $(SS, T_{max}, T_{min}, T_{max}\text{-}T_{min})$ | 0.73 | 1.195 | 5.40 | 98.33 |
| | $(SS, T_{max}, T_{min}, T_{max}\text{-}T_{min}, T_{mean})$ | 0.72 | 1.17 | 5.32 | 98.37 |
| | $(SS, T_{max}\text{-}T_{min}, T_{mean})$ | 0.74 | 1.92 | 5.39 | 98.33 |
| | $(SS, T_{max}\text{-}T_{min})$ | 0.77 | 1.20 | 5.43 | 98.28 |
| | $(SS, T_{mean})$ | 0.73 | 1.20 | 5.43 | 98.31 |
| GPR models-4 | $(SS, RH_{max}, RH_{min})$ | 0.80 | 1.23 | 5.58 | 98.29 |
| | $(SS, RH_{max}, RH_{min}, RH_{mean})$ | 0.80 | 1.23 | 5.58 | 98.27 |
| | $(SS, RH_{max}, RH_{min}, RH_{max}\text{-}RH_{min})$ | 0.81 | 1.24 | 5.60 | 98.28 |
| | $(SS, RH_{max}, RH_{min}, RH_{max}\text{-}RH_{min}, RH_{mean})$ | 0.80 | 1.24 | 5.59 | 98.27 |
| | $(SS, RH_{max}\text{-}RH_{min}, RH_{mean})$ | 0.81 | 1.25 | 5.64 | 98.24 |
| | $(SS, RH_{max}\text{-}RH_{min})$ | 0.82 | 1.24 | 5.62 | 98.21 |
| | $(SS, RH_{mean})$ | 0.80 | 1.23 | 5.56 | 98.25 |
| GPR models-5 | $(SS, T_{max}, RH_{max})$ | 0.71 | 1.18 | 5.34 | 98.35 |
| | $(SS, T_{min}, RH_{min})$ | 0.67 | 1.15 | 5.2 | 98.42 |
| | $(SS, T_{mean}, RH_{mean})$ | 0.68 | 1.16 | 5.26 | 98.39 |
| | $(SS, T_{max}\text{-}T_{min}, RH_{max}\text{-}RH_{min})$ | 0.76 | 1.19 | 5.39 | 98.31 |
| | $(SS, T_{mean}, T_{max}\text{-}T_{min}, RH_{max}\text{-}RH_{min}, RH_{mean})$ | 0.68 | 1.16 | 5.25 | 98.40 |
| | $(SS, T_{max}, T_{min}, T_{mean}, T_{max}\text{-}T_{min}, RH_{max}, RH_{min}, RH_{max}\text{-}RH_{min}, RH_{mean})$ | 0.68 | 1.15 | 5.22 | 98.42 |
| | $(SS, T_{max}, T_{min}, T_{mean}, RH_{max}, RH_{min}, RH_{mean})$ | 0.68 | 1.15 | 5.2 | 98.43 |

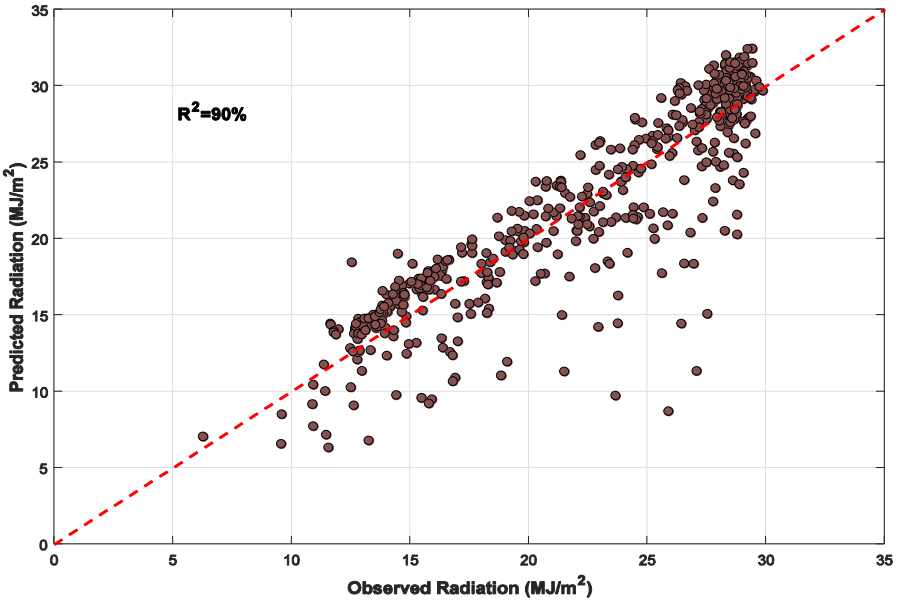


**Fig. 8.** Measured global radiation *versus* estimated GPR model ($RH_{\min}, RH_{\mathrm{mean}}$).

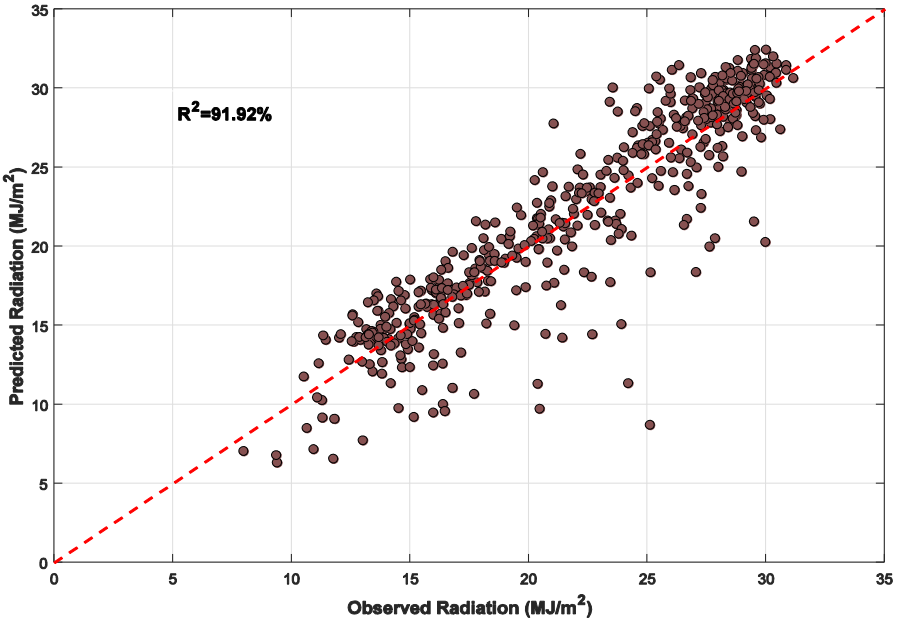


**Fig. 9.** Measured global radiation *versus* estimated GPR model ($T_{\min}, T_{\mathrm{mean}}, T_{\max}$).

Another important remark is when using only sunshine ratio as an input, high performance is obtained due to the strong correlation between the sunshine ratio and the clearness index. Therefore, combining sunshine duration with air temperature and relative humidity boosts the model performances. In order to analyze the performance of three GPR models: GPR$\{RH_{\min}, RH_{\max}\}$, GPR$\{T_{\min}, T_{\max}, T_{\max}\text{-}T_{\min}\}$ and GPR$\{SS\}$, the dispersion between the estimated and measured data values is shown in figs. 8, 9 and 10, respectively. As can be seen from these plots, the dispersions in the first and the second cases are strong where the correlation coefficients reach 90% and 91.92%. In the third case, GPR$_3\{SS\}$, the dispersion of the scatter plot is very weak ($r = 98.21\%$).

The second experiment aims at making a comparison of the best GPR model thus found with other AI methods. That is way the GPR model is compared to two ANN architectures, namely the radial basis function (RBF) and the multi-layer perceptron (MLP). The best combination $\{S, T_{\min}, RH_{\min}\}$ has been chosen as the input for the mentioned networks. Moreover, before building the RBF and MLP, the database is subdivided into learning (training and validation) and test sets. As we know, the GPR is a nonparametric method in which the validation data are not

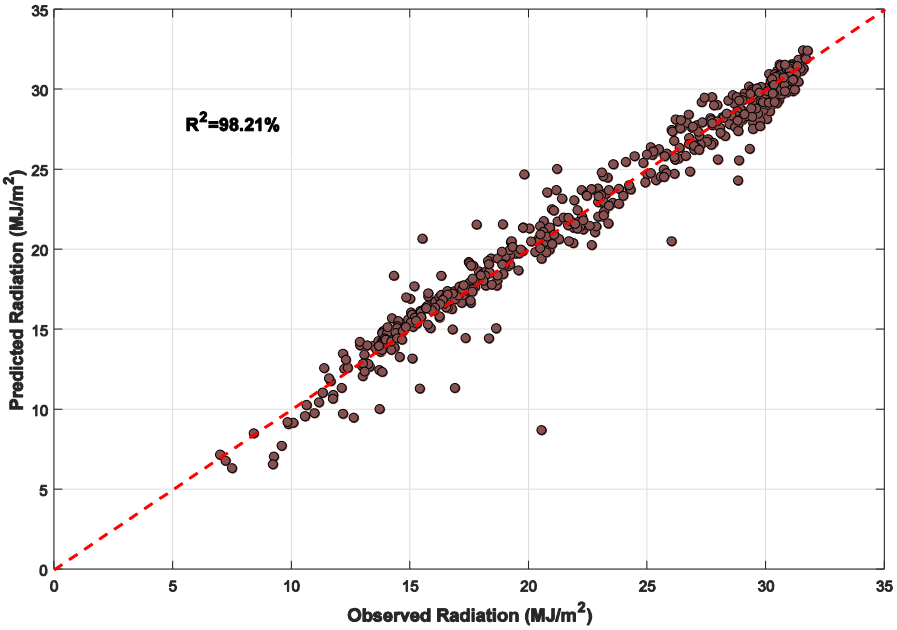


**Fig. 10.** Measured global radiation *versus* estimated GPR model ($SS$).

**Table 3.** Performance of the GPR model *versus* neural networks models.

| Models | Input parameters | MABE (MJ/m$^2$) | RMSE (MJ/m$^2$) | rRMSE | $r$ (%) |
|---|---|---|---|---|---|
| GPR | $\{S, T_{\min}, RH_{\min}\}$ | 0.67 | 1.15 | 5.20 | 98.42 |
| RBF | $\{S, T_{\min}, RH_{\min}\}$ | 0.70 | 1.17 | 5.23 | 98.38 |
| MLP | $\{S, T_{\min}, RH_{\min}\}$ | 0.73 | 1.242 | 5.61 | 98.17 |

needed. Indeed, 540 samples are used for training and the rest for testing. In neural networks models (RBF and MLP), 540 samples are used for the learning phase and the rest for the test. The learning subset is divided into training data (380 samples) and validation phase (180 samples).

It is worth highlighting that the optimal parameters of the RBF and MLP are obtained using the cross-validation scheme. After several trials, an RBF model with 6 neurons in the hidden layer and optimal width of all kernel functions equal to 0.7 has been found (we assume that all kernel functions have the same width $\sigma$). For the MLP neural network, the optimal architecture is one hidden layer containing 9 neurons and the learning rate equal to 1, using the Levenberg-Marquardt algorithm for training. From table 3, we can observe that the three models revealed significant results. However, the proposed model exhibits slight improvements against the compared methods.

As can be seen from figs. 11 and 12, it is clearly observed that even due to the random nature of the daily global solar radiation, the estimated values obtained by GPR$\{SS, T_{\min}, RH_{\min}\}$ follow the measured ones with great correlation.

In the third experiment, the performance of the three models has been considered against the number of training samples and the processing time. From tables 4–6, the following remarks can be drawn:

- In the case where the number of training equals 10 samples, the RBF model loses its generalization capacity, unlike MLP and GPR models (figs. 13, 14). In addition, the GPR is very fast in the training phase (0.2 s) but both models require a calculation time equal to 92 and 95.63 s.
- In the case of 50 samples, the MLP gives good results in terms of RMSE and $r$ with a processing time equal to 230 seconds, followed by GPR with fast training time (0.286 s).
- Starting from 100 samples, the GPR model is more efficient in terms of both statistical indicators and processing time.
- From 250 days, the performance of the three models is closer, with a slight dominance for the GPR.
- If we take the case of 550 days, the GPR needs only 35 seconds for training, while the MLP and RBF need 410 and 689 seconds, respectively, to achieve their calculations.

As a consequence, we can conclude that the GPR model has proved its efficiency daily global solar radiation prediction.

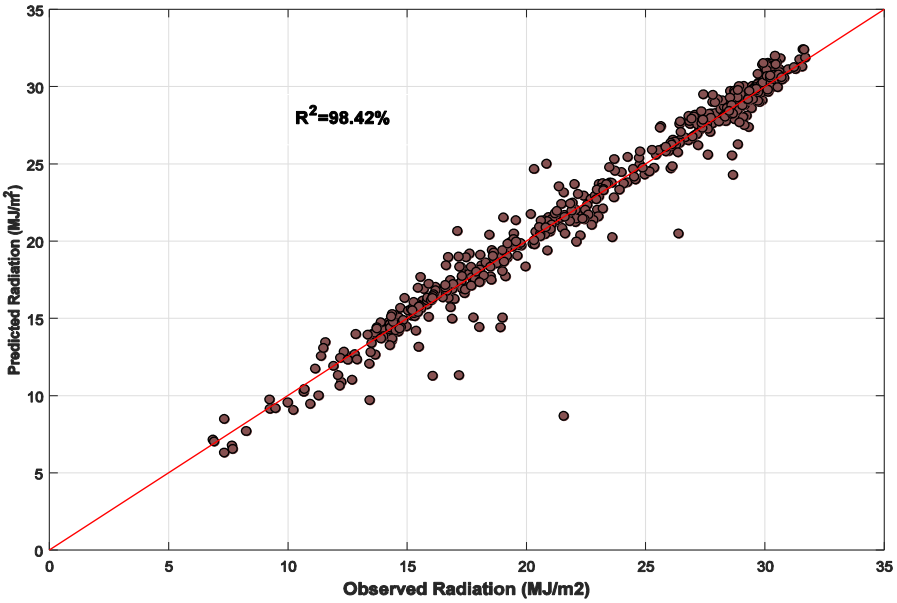


**Fig. 11.** Measured global radiation *versus* estimated GPR model ($SS, T_{\min}, RH_{\min}$).

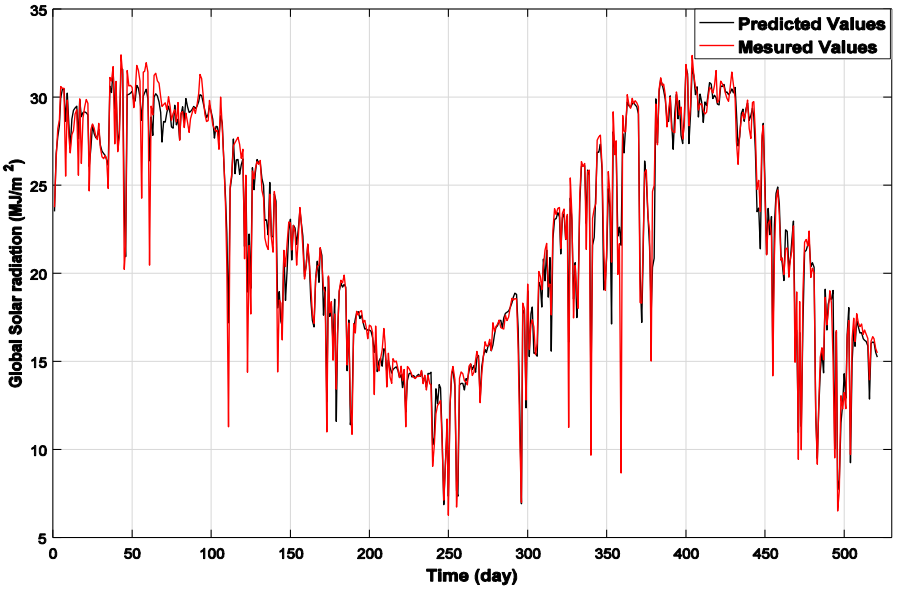


**Fig. 12.** Measured global radiation *versus* estimated best GPR model.

## 5 Conclusion

In this study, a new GPR model was proposed for estimating the daily global solar radiation on a horizontal surface. The main advantage of this approach is related mainly to its simplicity of implementation, fast training speed and accuracy. Five GPR models were developed using different input attributes including: $T_{\min}$, $T_{\max}$, $T_{\max}$-$T_{\min}$, $RH_{\min}$, $RH_{\max}$, $RH_{\max}$-$RH_{\min}$ and $SS$. The estimated values have been assessed against the measured data using five statistical indicators. The conducted examination showed the appreciable effect of input parameters on the precision of GPR models. It has been demonstrated that $H = \mathrm{GPR}\{SS, T_{\min}, RH_{\min}\}$ provide better accurate precision than the other proposed GPR models. It should be noted that the fraction of sunshine duration $SS$ is the most relevant parameter which assures high prediction accuracies.

In order to emphasize the GPR precision, its predictions are compared with two neural networks models, namely, MLP and RBF. The experimental results show an improvement in the predictive accuracy and in the capability of generalization. However, the statistical indices (MABE = 0.67 MJ/M$^2$, RMSE = 1.15 MJ/m$^2$, rRMSE = 5.20, $r$ = 98.42%) proved that the GPR model gives best accuracies when compared to the MLP and RBF architectures.

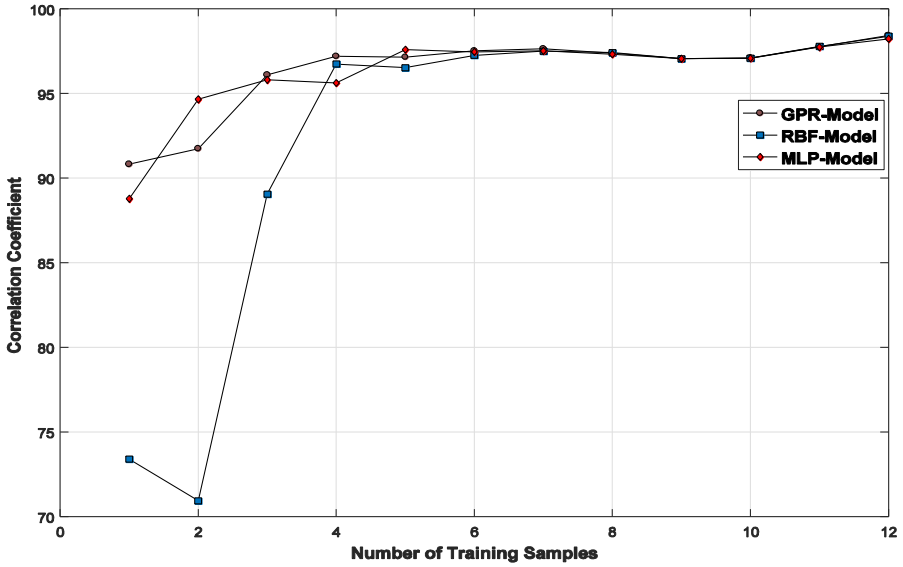


**Fig. 13.** Performance of GPR models against number of training samples in terms of correlation coefficient.

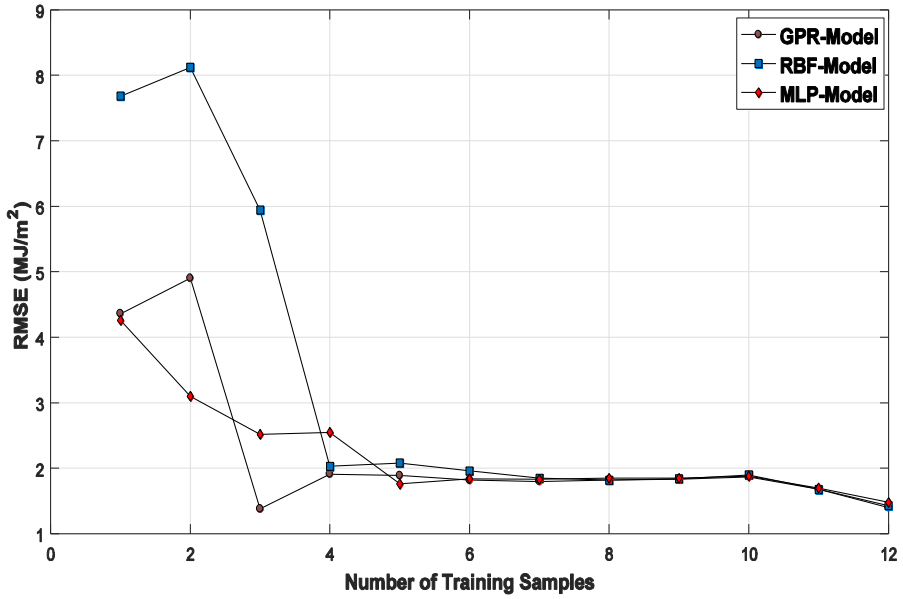


**Fig. 14.** Performance of GPR models against number of training samples in terms of RMSE.

The comparison between observed and estimated values clearly show that the proposed GPR proves remarkable improvement in the accuracy of the prediction for the day-by-day global solar radiation. Accordingly, it can be used also in regions that have a similar climate. Furthermore, other cases with different climate conditions will be considered in the future investigations.

## Appendix A.

The Gaussian process regression has become increasingly a powerful tool for data-driven modeling and forecasting. GPR models are a Bayesian non parametric approach that can be applied to solve a serious competitor for real supervised learning applications. It has been used to response surface modeling [32], system identification [33], calibration of spectroscopic analyzers [34,35] and ensemble learning [36].

The main idea behind the GPR theoretical is to place a prior directly on the space of functions. It can be defined by its mean function and covariance function, written as

$$f(x) \sim \mathcal{GP}(m(x), k(x, \hat{x})). \tag{A.1}$$

**Table 4.** Best GPR model performance against number of training samples.

| GPR models | Number of training samples | MABE ($MJ/m^2$) | RMSE ($MJ/m^2$) | rRMSE | $r$ (%) | Time (s) |
|---|---|---|---|---|---|---|
| Input parameters $\{S, T_{\min}, RH_{\min}\}$ | 10 samples | 3.25 | 4.36 | 16.41 | 90.82 | 0.206 |
| | 50 samples | 3.86 | 4.90 | 18.20 | 91.74 | 0.286 |
| | 100 samples | 1.38 | 1.38 | 2.27 | 96.10 | 0.304 |
| | 150 samples | 1.08 | 1.91 | 7.20 | 97.20 | 0.375 |
| | 200 samples | 1.04 | 1.89 | 7.25 | 97.15 | 0.477 |
| | 250 samples | 1.00 | 1.82 | 7.03 | 97.52 | 0.710 |
| | 300 samples | 0.95 | 1.8 | 6.80 | 97.64 | 0.793 |
| | 350 samples | 1.00 | 1.82 | 6.80 | 97.39 | 1.10 |
| | 400 samples | 1.00 | 1.83 | 6.62 | 97.06 | 1.60 |
| | 450 samples | 0.99 | 1.87 | 6.76 | 97.10 | 1.65 |
| | 500 samples | 0.90 | 1.68 | 6.11 | 97.76 | 2.39 |
| | 550 samples | 0.82 | 1.40 | 5.26 | 98.42 | 3.01 |

**Table 5.** RBF model performance against number of training samples.

| RBF models | Number of training samples | MABE ($MJ/m^2$) | RMSE ($MJ/m^2$) | rRMSE | $r$ (%) | Time (s) |
|---|---|---|---|---|---|---|
| Input parameters $\{S, T_{\min}, RH_{\min}\}$ | 10 samples | 5.99 | 7.68 | 28.92 | 73.40 | 92 |
| | 50 samples | 5.24 | 8.12 | 30.21 | 70.94 | 185 |
| | 100 samples | 4.06 | 5.94 | 22.08 | 89.06 | 199 |
| | 150 samples | 1.15 | 2.03 | 7.64 | 96.74 | 213 |
| | 200 samples | 1.15 | 2.08 | 7.97 | 96.52 | 212 |
| | 250 samples | 1.07 | 1.96 | 7.57 | 97.25 | 221 |
| | 300 samples | 1.03 | 1.85 | 7.10 | 97.50 | 220 |
| | 350 samples | 0.98 | 1.82 | 6.81 | 97.41 | 223 |
| | 400 samples | 0.99 | 1.84 | 6.65 | 97.05 | 240 |
| | 450 samples | 0.998 | 1.90 | 6.83 | 97.09 | 297 |
| | 500 samples | 0.90 | 1.68 | 6.13 | 97.78 | 350 |
| | 550 samples | 0.83 | 1.43 | 5.34 | 98.38 | 410 |

Let us consider a regression of input data $x$ containing $d$ variables. In the machine-learning approach, the main objective is to learn the functional relationship between the inputs of $d$-dimensional $x \in \mathbb{R}^d$ and the desired output variable $y$,

$$y = f(x), \tag{A.2}$$

where $\mathbb{R}$ denotes the real space and $f$ the unknown function.

The unknown function $f$ can be estimated by means of the following linear combination of basic functions:

$$\widehat{f}(x, w) = \sum_{j=1}^{M} W_J \phi_J(X), \tag{A.3}$$

$\{\phi_j(X)\}_{j=1}^{M}$ represents a set of basis functions which can be linear or nonlinear and $w = [w_1, \ldots, w_M]^T$ is the unknown vector for $M$ basis function of $f$:

$$y = \sum_{j=1}^{M} w_j \phi_j(x) + \varepsilon, \tag{A.4}$$

where $\varepsilon$ represents the error term.

**Table 6.** MLP model performance against number of training samples.

| MLP models | Number of training samples | MABE ($MJ/m^2$) | RMSE ($MJ/m^2$) | rRMSE | $r$ (%) | Time (s) |
|---|---|---|---|---|---|---|
| Input parameters $\{S, T_{\min}, RH_{\min}\}$ | 10 samples | 2.72 | 4.26 | 16.05 | 88.78 | 95.63 |
| | 50 samples | 2.50 | 3.10 | 11.5 | 94.66 | 230 |
| | 100 samples | 1.67 | 2.52 | 9.37 | 95.80 | 260 |
| | 150 samples | 1.47 | 2.55 | 9.62 | 95.62 | 287 |
| | 200 samples | 1.01 | 1.76 | 6.75 | 97.59 | 350 |
| | 250 samples | 1.045 | 1.84 | 7.10 | 97.46 | 398 |
| | 300 samples | 1.05 | 1.83 | 7.02 | 97.52 | 420 |
| | 350 samples | 1.03 | 1.85 | 6.97 | 97.32 | 486 |
| | 400 samples | 1.04 | 1.85 | 6.70 | 97.05 | 510 |
| | 450 samples | 0.99 | 1.88 | 6.79 | 97.08 | 564 |
| | 500 samples | 0.92 | 1.70 | 6.24 | 97.74 | 643 |
| | 550 samples | 0.89 | 1.48 | 5.57 | 98.23 | 689 |

In most linear and nonlinear regression models, a set of training data $D = \{X, Y\}_{i=1}^{N}$ of $N$ observation is needed to estimate the unknown weights $w$, and the basis function $\phi_j(X)$ can be defined as the projection of the data from the original space into high-dimensional space which is not the case in the proposed GPR models.

In the work of Rasmussen *et al.* [37], they mentioned that the basic block of the GPR is a GP that assumes Gaussian priors for function values specified which is identified by its second-order statistics:

$$f(x) \sim GP(m(x), k(x, \acute{x})), \tag{A.5}$$

where $m(x)$, $K(x, \acute{x})$ represent the mean and the covariance function of $f$. By definition GP is a collection of finite sets of random variables which follow Gaussian distribution [33]. Under GP, the prior distribution of $f$ is a Gaussian

$$p(f|X, \theta) \sim \mathcal{N}(0, K), \tag{A.6}$$

where $\theta$ represents hyper parameters.

Equation (A.6) states that a Gaussian distribution with zeros mean and the $N * N$ $K$ is a covariance matrix of $f$.

If the error term $\varepsilon$ in eq. (A.4) is assumed to be independent and follows Gaussian distribution, the likelihood function of the training target is also Gaussian:

$$p\{y|f, \sigma^2\} \sim \mathcal{N}(f, \sigma^2 I), \tag{A.7}$$

where $\sigma^2$ and $I$ represent the variance and identity matrix, respectively, of the error model. Then the posterior distribution of $f$ can be formulated using Bayes' rule:

$$p(f|y, X, \theta, \sigma^2) = \frac{p(y|f, \sigma^2)p(f|X, \theta)}{p(y|X, \theta, \sigma^2)}. \tag{A.8}$$

Under Gaussian process modeling, the posterior distribution of $f$ is also Gaussian, since both prior and likelihood function follow Gaussian distributions. The mean and covariance of the posterior function are given by [38]

$$\mu = K^T(K + \sigma^2 I)^{-1}, \tag{A.9}$$

$$\Sigma = K - K^T(K + \sigma^2 I)^{-1}K. \tag{A.10}$$

A central role for the GPR methodology is played by the covariance function $k(\cdot, \cdot)$ which embeds the geometrical distribution of the training data. In GPR literature some commonly used kernel functions include squared exponential or Gaussian kernel [38]:

$$k(x, \acute{x}|\theta) = \sigma_f^2 \exp\left(-\frac{r^2}{2l^2}\right), \quad \theta = \{\alpha, l, \sigma_f^2\} \tag{A.11}$$

and the family of covariance function is

$$k(x,\grave{x}|\theta)=\sigma_f^2\frac{2^{1-\upsilon}}{\Gamma(\upsilon)}\left(\frac{\sqrt{2\upsilon r}}{l}\right)^{\upsilon}k_{\upsilon}\left(\frac{\sqrt{2\upsilon r}}{l}\right),\quad \theta=\{\upsilon,l,\sigma_f^2\}. \tag{A.12}$$

In eqs. (A.11) and (A.12), $r=|x-\grave{x}|$ denote the Euclidean distance between two points and $\theta$ represents the hyper parameters associated with each covariance function. The marginal probability distribution can be estimated by integration over the latent function $f$ [38]:

$$p(y|X)=\int p(y|f,\sigma^2)p(f|X,\theta)\mathrm{d}f. \tag{A.13}$$

The log marginal likelihood is obtained

$$\log p(y|X)\propto-\frac{1}{2}y^T(K+\sigma^2I)^{-1}y-\frac{1}{2}\log|K+\sigma^2I|-\frac{N}{2}\log(2\pi). \tag{A.14}$$

Then the gradient-based algorithm is applied to estimate the unknown parameters $(\theta,\sigma^2)$ can be estimated from eq. (A.13) using the rule for conditioning Gaussian [38], predictive distribution of any test data $x_*$ can be evaluated:

$$p(f_*|x_*,y,X,\theta,\sigma^2), \tag{A.15}$$

where the mean $m$ and $\sigma^2$ variance of the prior distribution $p(y_*|y)$ can be derived by the following equation:

$$m(x_*)=\phi(x_*)^T\mu=K_*^T(K+\sigma^2I)^{-1}y, \tag{A.16}$$

$$\vartheta^2(x_*)=\phi(x_*)^T\Sigma\Phi(x_*)=K_{**}-K_*^T(K+\sigma^2I)^{-1}K_*, \tag{A.17}$$

$K_*=[K(x_*,x_1),\ldots,K(x_*,x_N)]^T$, $K_{**}=K(x_*,x_*)$, $\mu$ and $\Sigma$ are the posterior mean and variance of $f$.

The prior mean which represents the best output estimate for the considered data and the variance which represents the confident interval associated with the model.

## Appendix B.

Clearness index $K_t$, which is the ratio of global solar radiation on horizontal surface (GHR) to the extraterrestrial solar radiation on horizontal surface ($GH_0R$).

The extra-terrestrial solar radiation ($H_0$) is given as

$$GH_0R=\frac{24*3600}{\pi}I_0f\left(\cos\lambda\cos\delta\sin w_s+\frac{\pi}{180}w_s\sin\lambda\sin\delta\right), \tag{B.1}$$

where $I_0$ is the solar constant equal to $1360\,\mathrm{W/m^2}$, $f$ is the eccentricity correction factor, $\lambda$ is the latitude of the region, $\delta$ is the solar declination and $W_s$ is the mean sunrise hour angle:

$$f=1+0.33\left(\cos\frac{360\lambda}{365}\right), \tag{B.2}$$

$$\delta=23.45\sin\left[\frac{360(284+n)}{365}\right], \tag{B.3}$$

$$ws=\arccos(-\tan\lambda\tan\delta),$$

where $n$ represents the number of days.

## Nomenclature

| | | | |
|---|---|---|---|
| $a$-$b$: | Angstrom-Prescott regression coefficients | AI: | Artificial intelligence |
| $k_t$: | Clearness index | RBF: | Radial basis function neural network |
| $SS$: | Sunshine ratio | MLP: | Multi-layer perceptron |
| GHR: | Global horizontal radiation ($\mathrm{M \cdot J/M^2}$) | LS-SVM: | Least squares support vector machine |
| GHR0: | Extra-terrestrial solar radiation ($\mathrm{M \cdot J/M^2}$) | ELM: | Extreme learning machine |
| DHI: | Diffuse horizontal solar irradiation ($\mathrm{M \cdot J/M^2}$) | FFA: | Fire-fly algorithm |
| DNI: | Direct normal irradiance ($\mathrm{M \cdot J/M^2}$) | ANN-MTM: | Artificial neural networks Markov transition matrix |
| $I0$: | Solar constant ($1360\,\mathrm{W/m^2}$) | | |
| $f$: | Eccentricity correction factor | ANFIS: | Adaptive neuro-fuzzy inference system |
| **Greek letters** | | LLR: | Local linear regression |
| $\lambda$: | Latitude of the region (0) | KELM: | Kernel extreme learning machine |
| $\delta$: | Solar declination (0) | WT: | Wavelet transform |
| $\omega_s$: | Mean sunrise hour angle (0) | ARMA: | Mixed auto-regressive moving average |
| **Acronyms** | | MBE: | Mean absolute bias error |
| GPR: | Gaussian process regression | RMSE: | Root mean square error |
| ANN: | Artificial neural networks | rRMSE: | Relative mean square error |
| SVM: | Support vector machine | $r$: | Correlation coefficient |

## References


1. H. Othenio, J. Awange, *Energy Resources in Africa* (Springer, 2016).
2. A.S.S. Dorolvo, D.B. Ampratwum, Renew. Energy **17**, 421 (1999).
3. J.Y. Almorox, C. Hontoria, Energy Convers. Manag. **45**, 1529 (2004).
4. M. Benghanem, A.A. Joraid, Saudi Arabia. Renew. Energy **32**, 2424 (2007).
5. D.B. Ampratwum, A.S.S. Dorolvo, Appl. Energy **63**, 161 (1999).
6. Y.A.G. Abdallah, Int. J. Sol. Energy **16**, 111 (1994).
7. H. Duzen, H. Aydin, Energy Convers. Manag. **58**, 35 (2012).
8. S. Benkaciali *et al.*, Rev. Energies Renouv. **19**, 617 (2016).
9. Mawloud Guermoui *et al.*, Leonardo Electron. J. Pract. Technol. **15**, 35 (2016).
10. J. Almorox, C. Hontoria, M. Benito, Appl. Energy **88**, 1703 (2011).
11. M.S. Mecibah, T.E. Boukelia, R. Tahtah, K. Gairaa, Renew. Sustain. Energy Rev. **36**, 194 (2014).
12. S. Mohanty, P.K. Patra, S.S. Sahoo, Renew. Sustain. Energy Rev. **56**, 778 (2016).
13. J.A. Prescott, Trans. R. Soc. South Austr. **64**, 114 (1940).
14. J.L. Chen, G.S. Li, S.J. Wu, Energy Convers. Manag. **75**, 311 (2013).
15. M. Şahin, Y. Kaya, M. Uyar, S. Yildirim, Int. J. Energy Res. **38**, 205 (2014).
16. A.S.S. Dorolvo, J.A. Jervase, A. Al-lawati, Appl. Energy **71**, 307 (2002).
17. M. Benghanem, A. Mellit, Saudi Arabia. Energy **35**, 3751 (2010).
18. O. Şenkal, T. Kuleli, Appl. Energy **86**, 1222 (2009).
19. A. Sözen, E. Arcaklioğlu, M. Özalp, Energy Convers. Manag. **45**, 3033 (2004).
20. A. Mellit, M. Benghanem, A. Hadj-Arab, A.A. Guessoum, Sol. Energy **79**, 469 (2005).
21. A. Gani, K. Mohammadi, S. Shamshirband, J. Piri, Theor. Appl. Climatol. **125**, 679 (2016).
22. J. Zeng, W. Qiao, *Short-term solar power prediction using an RBF neural network*, in *Power and Energy Society General Meeting 2011* (IEEE, 2011).
23. M. Bou-Rabee, S.A. Sulaiman, M. Saad Saleh, S. Marafi, Renew. Sustain. Energy Rev. **72**, 434 (2017).
24. B.B. Ekici, Measurement **50**, 255 (2014).
25. S. Shamshirband, K. Mohammadi, H. Chen, C. Ma, J. Atmos. Sol.-Terr. Phys. **134**, 109 (2015).
26. J.C. Cao, S. Cao, Energy **31**, 3435 (2006).
27. W.A. Rahoma, U.A. Rahoma, A.H. Hassan, J. Comput. Sci. **7**, 1605 (2011).
28. A. Mellit, S.A. Kalogirou, S. Shaari, A. Hadj Arab, Renew. Energy **33**, 1570 (2008).
29. K. Mohammadi, S. Shamshirband, A. Kamsin, P.C. Lai, Zulkefli Mansor, Renew. Sustain. Energy Rev. **63**, 423 (2016).
30. M.R. Yaich, A. Bouhanik, *Atlas solaire Algérien* (Centre de développement desénergies renouvelables, 2012) `www.cder.dz`.
31. Guermoui Mawloud *et al.*, Eur. Phys. J. Plus **133**, 22 (2018).
32. A. Rabehi, G. Mawloud, L. Djemoui, Int. J. Ambient Energy (2018) `https://doi.org/10.1080/01430750.2018.1443498`.
33. L.L.T. Chan, Y. Liu, J. Chen, Ind. Eng. Chem. Res. **52**, 18276 (2013).

34. W. Ni, L. Nørgaard, M. Mørup, Anal. Chim. Acta. **813**, 1 (2014).
35. K. Wang, T. Chen, R. Lau, Chemometr. Intell. Lab. **105**, 1 (2011).
36. Y. Liu, Z. Gao, Appl. Polym. **132**, 1 (2015).
37. A.Y. Sun, D. Wang, X. Xu, J. Hydrol. **511**, 72 (2014).
38. C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes For Machine Learning* (MIT Press, 2006).