Eur. Phys. J. Plus (2022) 137:289
https://doi.org/10.1140/epjp/s13360-022-02398-z

THE EUROPEAN PHYSICAL JOURNAL PLUS

Regular Article

Estimating daily global solar radiation in hot semi-arid climate using an efficient hybrid intelligent system

Mehdi Jamei[1,a], **Iman Ahmadianfar**[2], **Mozhdeh Jamei**[3,4], **Masoud Karbasi**[5], **Ali Asghar Heidari**[6], **Huling Chen**[7]

[1] Faculty of Engineering, Shohadaye Hoveizeh Campus of Technology, Shahid Chamran University of Ahvaz, Dashte Azadegan, Ahvaz, Iran
[2] Department of Civil Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran
[3] Water Science and Engineering Dept, Ferdowsi University of Mashhad, Mashhad, Iran
[4] Khuzestan Water and Power Authority, Ahvaz, Iran
[5] Faculty of Agriculture, University of Zanjan, Zanjan, Iran
[6] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
[7] College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, Zhejiang, China

Received: 13 April 2021 / Accepted: 17 January 2022

Abstract Solar energy is one of the most important renewable energy sources. Assessing the area's solar potential needs analyzed information about the dataset of the measured global solar radiation (GSR). Researchers recently detected the high potential of state-of-the-art artificial intelligence (AI) methods in successfully estimating the GSR. In this study, a novel hybrid AI-based tool consisting of a least square support vector machine (LSSVM) integrated with improved simulated annealing (ISA) is proposed to predict the GSR over the Ahvaz synoptic station located in the South-West of Iran. The potential of the proposed hybrid paradigm so-called LSSVM-ISA was evaluated by using multivariate adaptive regression spline (MARS), generalization regression neural network (GRNN), and multivariate linear regression with interactions (MLRI). For precise assessment of efficiency of the AI models, various statistical metrics and validation methods were used to assess the precision of the developed models. A comparison of the obtained results indicated that the LSSVM-ISA method performed better than the MARS, GRNN, and MLRI models. The achieved RMSE values of the MARS, GRNN, and MLRI models were decreased by 9%, 16%, and 30% using the LSSVM-ISA model. Finally, the results demonstrated that the LSSVM-ISA model could be successfully employed for accurately estimating GSR.

1 Introduction

Solar energy is the electromagnetic energy emitted from the sun. Solar radiation is referred to as global solar radiation (GSR) that is the total of direct shortwave radiation received from the sun and diffuse sky radiation that has been scattered across the atmosphere [1]. The GSR data are essential for several theoretical and practical applications such as solar energy systems, architecture, agriculture, meteorological, and climatological models [2]. Pyranometers use thermoelectric, photoelectric, pyroelectric, or bimetallic elements as sensors to measure the global solar irradiance (radiant flux density/m^2) at the meteorological stations [1]. However, the GSR measurements are limited and only available at a few meteorological stations due to the high cost of solar measuring instruments, the accurate equipment calibration, installation, and maintenance requirements [1, 3, 4]. Besides, the missing GSR data, incorrect data, measurements in a short and discontinuous period are other problems at many stations [5, 6].

Therefore, these problems have resulted in the recommendations for alternative techniques to accurately determine the GSR, including utilizing empirical, deterministic, and artificial intelligence (AI) methods [7, 8]. Recently, AI methods have been broadly applied to predicting the GSR.

AI techniques that have been employed for estimating the GSR consist of artificial neural network (ANN) [9], radial basis function (RBF) [10], genetic programming (GP) [11], adaptive neural fuzzy inference system (ANFIS), multiple layer perceptron (MLP), multiple linear regression (MLR) [12], extreme learning machine (ELM) [13], support vector machine (SVM) [14], random forest (RF) [15], extreme gradient boosting (XGB) [16], adaptive regression spline (MARS) [17], M5 model tree [18], logistic regression (LR) [19], and least square support vector regression (LSSVM) [20]. These AI models have been developed for predicting GSR by in the previous studies, which are explained below.

To estimate daily GSR at the Dezful station in Iran's southern west, Behrang et al. (2010) used ANN and radial basis function (RBF) approaches [21]. The input dataset was comprised of a variety of meteorological variables. As a consequence of their findings, it was concluded that the input dataset should include five variables: average air temperature, daylight hours, relative humidity, wind speed, and the day of the year. The genetic programming (GP) and simulated annealing (SA) models were used in a hybrid model by Mostafavi et al. (2013) to estimate the GSR in Iran [22]. Using maximum and lowest air temperatures as input factors, they

[a] e-mail: m.jamei@shhut.ac.ir (corresponding author)

Published online: 01 March 2022

were able to accurately forecast the GSR. A fuzzy genetic (FG) model was developed by Kisi (2014) to forecast the SR in Turkey [23]. The suggested model performed better than the MLP and ANFIS models in the testing phase and showed promising outcomes compared to them. Multiple linear regression (MLR), the adaptive neuro-fuzzy inference (ANFIS), and four empirical equations were all used by Hatice Citakoglu (2015) to make predictions about Turkish solar radiation (SR) [24]. Authors have shown that the ANN technique performs better than the other empirical equations in predicting the SR.

Using the extreme learning machine (ELM), support vector machine (SVM), genetic programming (GP), and artificial neural network (ANN) approach, Shamshirband et al. (2016) suggested estimating monthly GSR in Iran [25]. According to the authors, the ELM technique outperforms the ANN, GP, and SVR models in terms of accuracy. A SVM model was used by Belaid and Mellit (2016) to estimate the daily and monthly GSR in Ghardaïa (Algeria) [26]. According to their findings, the suggested SVM model exhibited high accuracy when tested against the MLP. Using the random forest (RF) and the firefly algorithm (FA), Ibrahim and Khatib (2017) developed a hybrid model for forecasting hourly GSR [27].

In another research, the models of extreme gradient boosting (XGB) and support vector regression (SVR) were compared with empirical models for predicting daily GSR in another study conducted by Fan et al. (2018) [28]. Following the acquired findings, the XGB and SVR were more exact than the empirical models. Using the MARS model created by Li et al. (2019), they compared it to the analytic neural network (ANN) and logistic regression (LR) models [29]. The findings showed that the MARS approach was superior to the ANN and LR methods in terms of accuracy. Introducing a dynamic ANFIS model, Kisi et al. (2019) evaluated the SR model against MARS, M5 model trees, and least square support vector regression (LSSVM) [7]. The findings demonstrated that the suggested strategy might outperform the MARS, M5 model tree, and LSSVR models when predicting the SR. Gürel et al. (2020) compared the ability of the ANN method with empirical models, time series (holt-winters (HW)), and response surface methodology (RSM) for modeling the GSR in Turkey [30]. Regarding the results, the ANN presented the best results compared to the other peers. Fan et al. (2020) proposed three hybrid SVMs with bat algorithm (SVM-BAT), particle swarm optimization (SVM-PSO), and whale optimization algorithm (SVM-WOA) for the prediction of daily diffuse solar radiation [31]. These methods were compared with the SVM, MARS, and XGB methods. Their results indicated that the SVM-BAT could further enhance the prediction precision in diffused solar radiation compared to the SVM, SVM-PSO, SVM-WOA, and EGB models. Alizamir et al. (2020) evaluated the potential of six machine learning methods, MLP, gradient boosting tree (GBT), MARS, classification and regression tree (CART), and ANFIS based on fuzzy c-means clustering (ANFIS-FCM) and subtractive clustering (ANFIS-SC) to estimate the SR from two stations in Turkey and the USA [32]. The results indicated that the GBT method provided better accuracy in modeling the SR than the MLP, MARS, ANFIS, and CART models.

Regarding the above literature on predicting the GSR using the AI methods, the motivation for exploring more accurate and reliable AI methods is still a challenging task. Among different AI methods, the LSSVM methods are one of the robust and reliable AI methods that have a promising predictive ability. Generally, AI models, especially the LSSVM model, are susceptible to their setting parameters, so incorrect selection yields an optimal local solution. To address this, combining AI models with meta-heuristic algorithms can be a suitable approach to succeed. Hence, this paper introduces a new hybrid of the LSSVM with an improved version of the SA (ISA) method to optimize the control parameters of the LSSVM. For evaluation purposes, the LSSVM-ISA model was compared with several AI models on the GSR modeling, including generalization regression neural network (GRNN), multivariate linear regression with interactions (MLRI), MARS, and empirical equations.

The rest of this study is described as follows. Section 2 expresses the materials and methods, including empirical models, the dataset used in this paper, the procedure for data quality control, the introduction of the hybrid SA and the LSSVM model. It also provides a summary of other AI models employed in this study (i.e., MARS, GRNN, and MLRI). The results are discussed in Sect. 3. Validation of the proposed model with traditional approaches is presented in Sect. 4. Uncertainty analysis is implemented in Sect. 5. Finally, the study conclusions are expressed in Sect. 6.

2 Materials and methods

2.1 Empirical models to estimate global solar radiation

Numerous empirical models have been developed to estimate the GSR based on meteorological variables such as minimum daily temperature, maximum daily temperature, monthly temperature, sunshine hour, extraterrestrial radiation, relative humidity, albedo, precipitation, cloudiness, and evaporation. [33, 34]. The important empirical models and their equations forms are illustrated in Table 1. Some of these models were modified to solve the problem of the availability of meteorological data by researchers.

2.2 Study area and data processing

The current study focuses on global solar radiation assessment at Ahvaz city. The meteorological data collected from Ahvaz synoptic station belong to the IR of Iran Meteorological Organization (IRIMO), over historical data of 10 years (1 July 2009–1 July 2019). Ahwaz station is located in the Khuzestan province in the southwest of Iran with latitude 31 20 N, longitude 48 40 E, and elevation 22.5 m (Fig. 1). Ahwaz has a hot semi-arid climate with hot and long summers and moderate and short winters. This region's annual

Table 1 The most cited empirical models for predicting global solar radiation

Model	Equation	Empirical coefficient (Parameter)	Source
Angstrom and Prescott	$R_s = \left(a + b\left(\frac{n}{N}\right)\right)R_a$	*a, b*	[36]
Swartman and Ogunlade	$R_s = a + b\left(\frac{n}{N}\right) + c \cdot \mathrm{RH}$	*a, b, c*	[37]
Hargreaves	$R_s = a\sqrt{(T_{\max} - T_{\min})} \cdot R_a$	*a*	[38]
Bristow and Campbell	$R_s = a\left(1 - \exp\left(-b(T_{\max} - T_{\min})^c\right)\right)R_a$	*a, b, c*	[39]
De Jong and Stewart	$R_s = a(T_{\max} - T_{\min})^b \cdot \left(1 + cP + dP^2\right)R_a$		[40]
Allen	$R_s = a\left(\sqrt{T_{\max} - T_{\min}}\right)R_a$		[41]
Donatelli and Campbell	$R_s = a\left(1 - \exp\left(-b \cdot \frac{(T_{\max} - T_{\min})^c}{(T_{\mathrm{month}})}\right)\right)R_a$	*a, b, c*	[42]
Hunt	$R_s = a\sqrt{(T_{\max} - T_{\min})} \cdot R_a + b$	*a, b*	[43]
Hunt	$R_s = a\sqrt{(T_{\max} - T_{\min})} \cdot R_a + bT_{\max} + cP + dP^2 + e$	*a, b, c, d, e*	[43]
Goodin, Hutchinson, Vanderlip, and Knapp	$R_s = a\left(1 - \exp\left(-b.\frac{(T_{\max} - T_{\min})^c}{(R_a)}\right)\right)R_a$	*a, b, c*	[44]
Elagib and Mansell	$R_s = a\left(\exp\left(b \cdot \left(\frac{n}{N}\right)\right)\right)R_a$	*a, b, c*	[45]
Chen, Ersi, Yang, Lu, and Zhao	$R_s = \left(\mathrm{a} \cdot ln(T_{\max} - T_{\min}) + b \cdot \left(\frac{n}{N}\right)^c + d\right)R_a$	*a, b, d*	[46]

where R_s is the global solar radiation (MJ /m^2/day); R_a extraterrestrial radiation (MJ /m^2/day); n the actual duration of sunshine (hr); N maximum possible duration of sunshine or daylight hours (hr); n/N relative sunshine duration; T_{min} minimum daily air temperature (°C); T_{max} maximum daily air temperature (°C); T_{Month} mean air temperature of the month (°C); RH the relative humidity (%); and P daily precipitation (mm)

Table 2 Descriptive statistics of meteorological variables and global solar radiation data for the Ahvaz stations located in Iran

Variables	Day	S_h (h)	T_{ave} (°C)	W_s(m/s)	R_h (%)	GSR(kWh/m^2/day)
Minimum	1	0	3.64	0.85	4.96	0.21
Maximum	366	13	44.55	11.82	88.97	8.49
Range	365	13	40.91	10.97	84.01	8.28
Mean	183.1	8.602	26.79	3.594	30.03	5.397
Std. Deviation	105.4	3.442	10.25	1.351	19.41	1.918
Skewness	2.109E-05	− 1.194	− 0.1415	0.890	0.743	− 0.404
Kurtosis	− 1.2	0.434	− 1.337	1.124	− 0.5069	− 0.889
Correlation with GSR	− 0.064	0.772	0.786	0.236	− 0.784	1.000

mean, maximum, and minimum air temperatures are 26.3, 33.4, and 19.2 °C, respectively. The annual mean precipitation is 202 mm, and the annual mean relative humidity is 43.3% [35]. The collected meteorological data included the average daily temperature (T_{ave}, °C), sunshine hours (S_h, hr), relative humidity ($R_{h,\%}$), average wind velocity (W_s, m/s) at 10 m and day of the year in the range of [1 366]. These data were used as input values to estimate the AI-based models' predictive daily global solar radiation (GSR, kWh/m2/day). The predictor time series and observed values of GSR were shown in 10 years (3653 days) in Fig. 2. Moreover, Table 2 summarizes the descriptive statistics of the input time series for ten years.

The histogram of the frequency related to all the time series variables and the probability density function of the normal distribution are illustrated in Fig. 3. According to Fig. 3 and Table 1, the mean daily sunshine hours and average daily temperature dataset have the highest skewness and kurtosis, respectively. However, corresponding to [47], the skewness and kurtosis values for the range of [− 1.96 + 1.96] are acceptable to prove normal univariate distribution. Also, a visual correlation matrix between the predictor variables (Day, S_h, T_{ave}, W_s, and R_h) and the target variable (GSR) for Ahwaz synoptic station was indicated to survey the degree of influence of predictive parameters on GSR in Fig. 4.

According to the correlation analysis based on the Pearson correlation coefficient [48], it seems the average daily temperature (T_{ave}) and relative humidity (R_h). Daily sunshine hours (S_h) due to having the highest Pearson correlation ($r_p = 0.79$, −0.78, and 0.77, respectively) with the target are a more significant element in GSR estimating, whereas the day and mean wind velocity (W_s) by yielding the lowest magnitude of the correlation ($r_p = -0.06$ and 0.24, respectively) play a weaker role in daily GSR estimating process. However, a certain degree of the above analysis depends on the predictive variables' nonlinearity dependence with the objective parameter.

There is one 'thumb of rule' for data partitioning, and the researcher utilized different divisions between training and testing datasets, which has influenced the results of problems [49]. In this research, the daily time series from 01 July 2009 to 1 July 2019

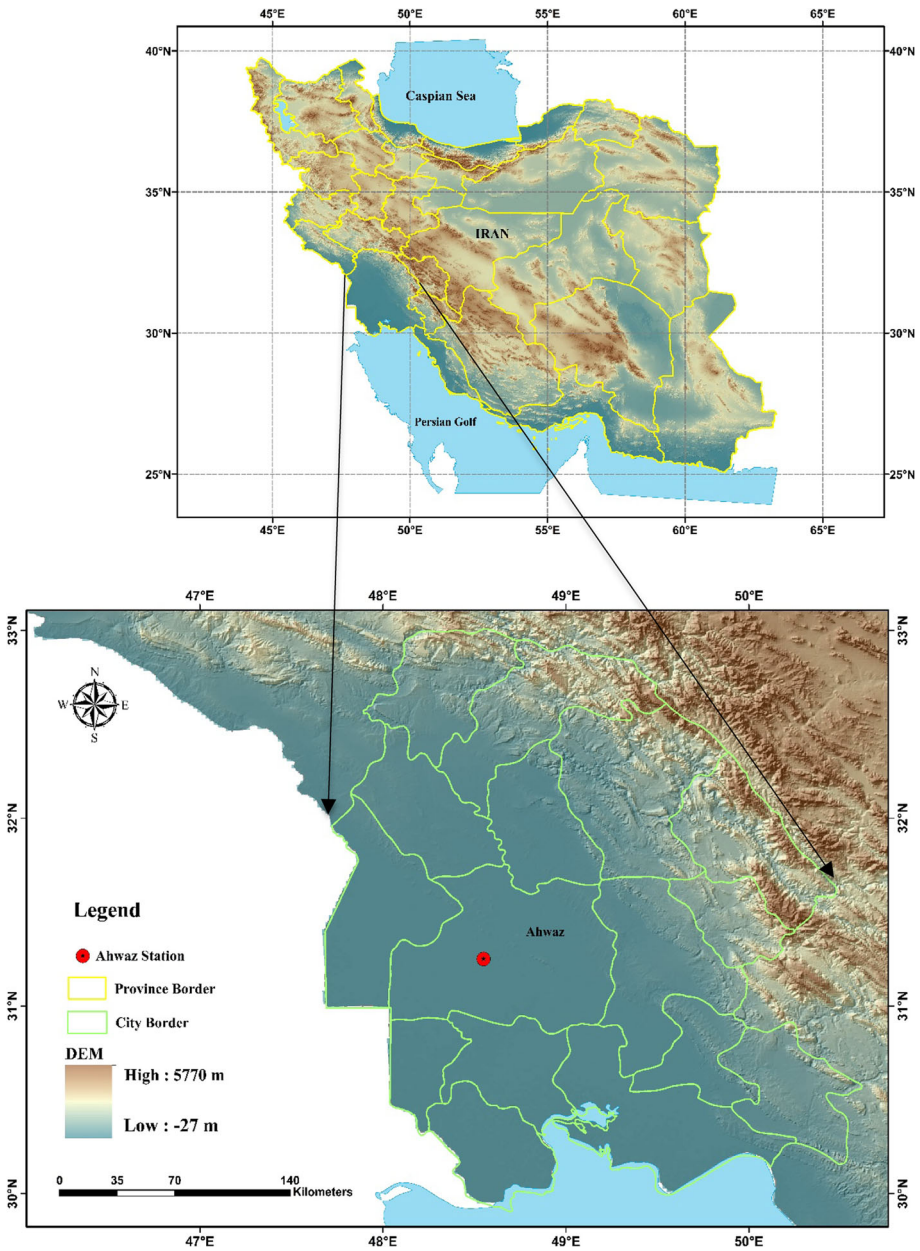

Fig. 1 Study area: location of Ahvaz synoptic station

were split into 80% (01 July 2009 to 1 July 2017) for the training dataset and 20% (02 July 2017 to 1 July 2019) as the test set. Furthermore, all the datasets were normalized between zero and one range based on Eq. 1:

$$x_{\text{nor}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{1}$$

where x is the original variable value, x_{nor} is the normalized value, and x_{max} and x_{min} are the maximum and minimum of the variable (x), respectively.

2.3 The methodology of predictive models

2.3.1 Improved simulated annealing

Kirkpatrick et al. [50] introduced the simulated annealing (SA) as a single-based optimizer. The annealing process inspired this algorithm in metallurgy. SA has two main phases: heating and cooling. The proposed method employs a temperature factor (t) to transition from heating to cooling. Although the SA is a powerful optimizer, it may be stuck in the local solutions because it is

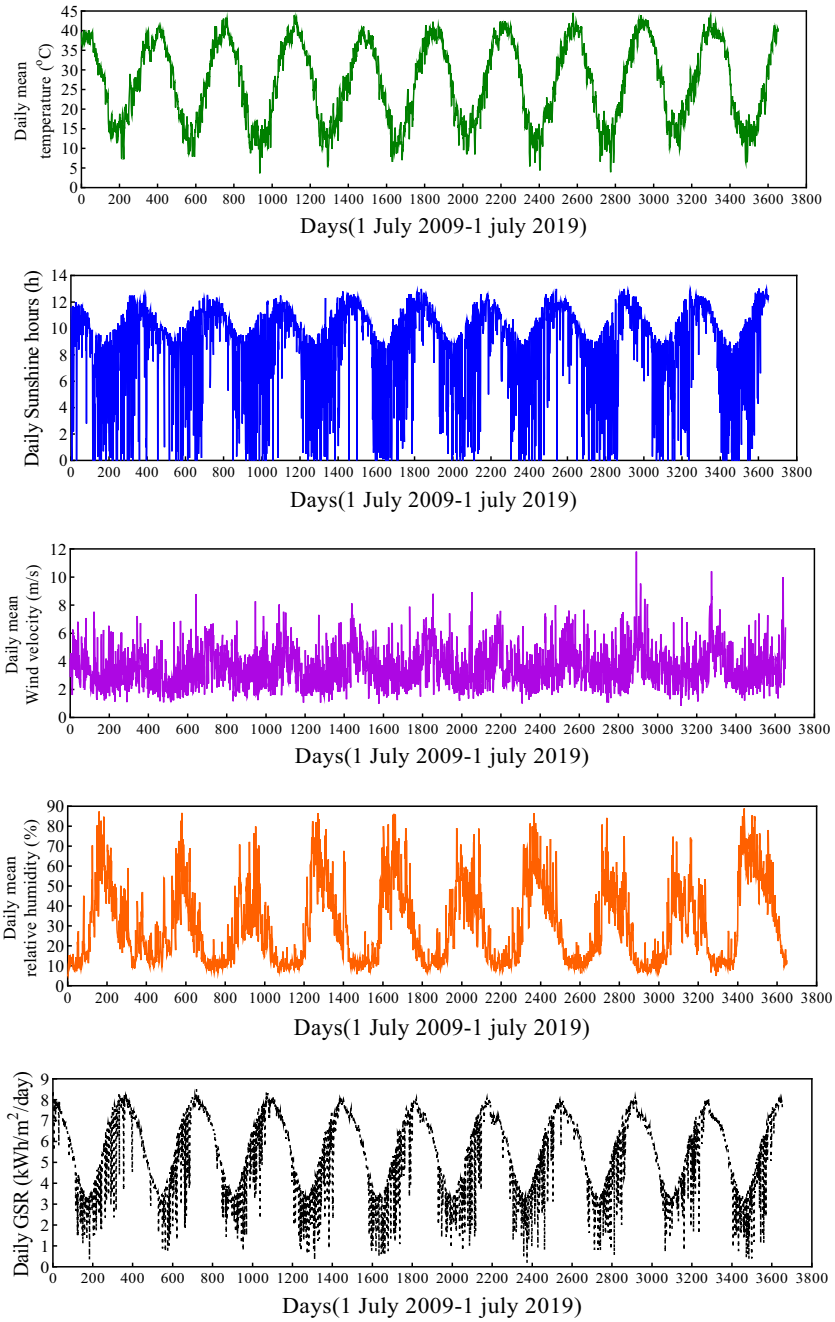

Fig. 2 The measured predictive and output variable over 10 years (1 July 2009–1 July 2019) at Ahvaz synoptic station

susceptible to the initial solutions [51]. An improved version of the SA algorithm is used to alleviate this shortcoming, which is called the ISA in this study. The proposed algorithm can be defined in the following stages:

Stage 1: Initialization: Produce an initial random solution (S_0). Specify the minimum and maximum temperature ($t_{\min}$ and $t_{\max}$) and the total number of iterations (MaxIt). Set $t = t_{\max}$ and $S_{best} = S_0$ (where x_{best} is the best-so-far solutions). The objective function of x_{best} is Z^* ($Z^* = Z(S_{\text{best}})$).

Stage 2: Main loop: In this step, the local search of the SA is defined. Also, two adaptive parameters ρ and T are formulated. The parameter ρ is a scale factor and sets how the position of solutions changes. The t parameter is updated in each iteration and is a termination condition for determining the end of a local search operation.

While $t > t_{\min}$

for $it = 1$ to *MaxIt.*

1. Produce the new solution $u_{\text{it},i}$:

$$u_{\text{it},i} = S_{\text{it},i} + \mu \times \left(S_{\max,i} - S_{\min,i}\right) \times \text{rand } n \tag{2}$$

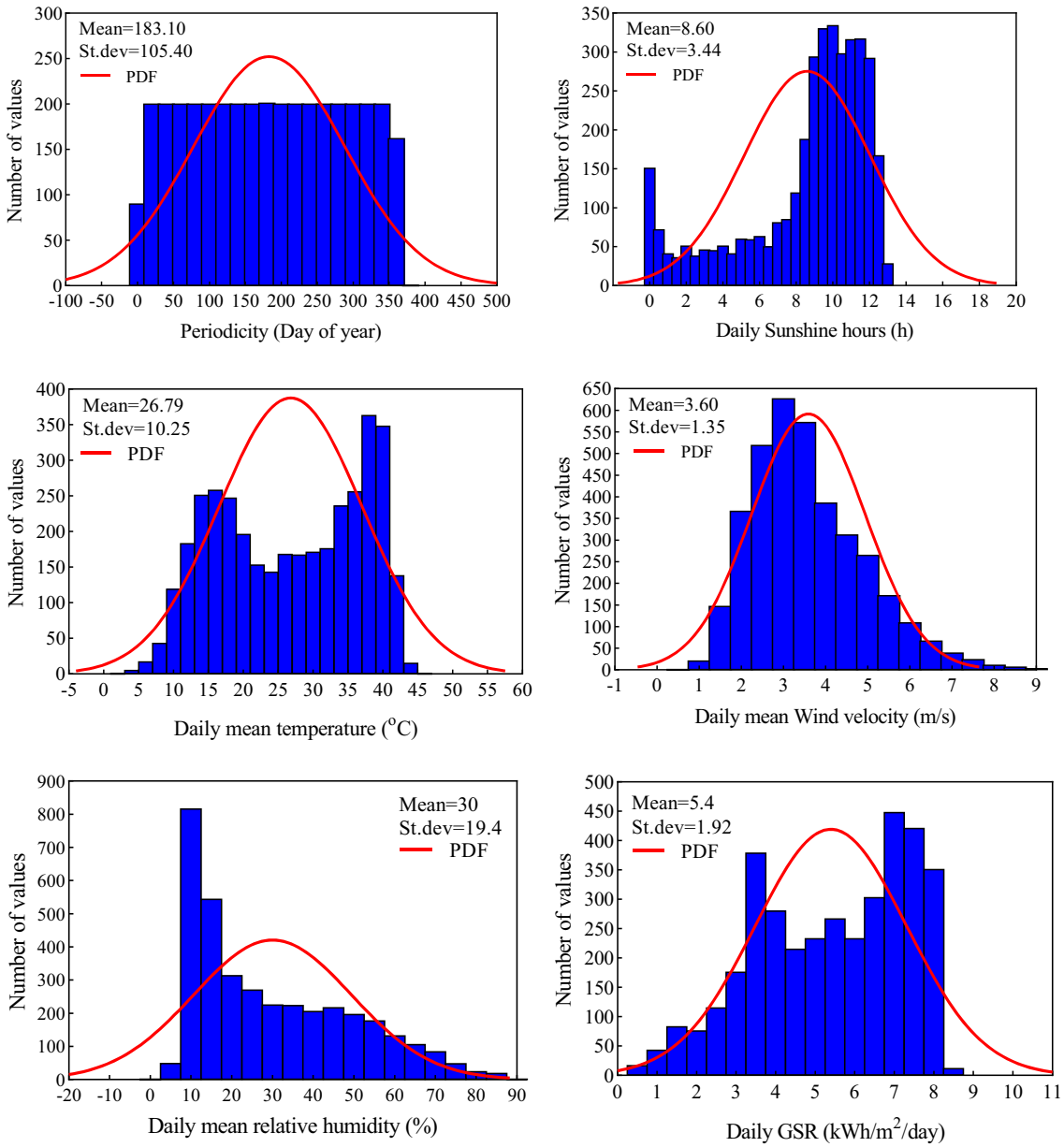

Fig. 3 The histogram of the frequency and probability density function of the predictive and objective variable

where i is a randomly chosen integer value from $[1, D]$. $S_{\min}$ and $S_{\max}$ are the lower and upper bounds of the problem. $randn$ is a random normal distribution number and the (average $= 0$) and (standard deviation $= 1$). μ is an adaptive factor which decreased by the equation $\mu = \mu_0 \times \exp(-a)$. Based on [51], $a = 1.01$ and the initial value of μ_0 is 1.

(b) Check constraints on the solution $u_{it,i}$:

$$u_{it,i} = \begin{cases} S_{\min} + \left(u_{\mathrm{it},i} - S_{\max}\right) & \text{if } u_{\mathrm{it},i} > S_{\max} \\ S_{\max} + \left(S_{\min} - u_{\mathrm{it},i}\right) & \text{if } u_{\mathrm{it},i} < S_{\min} \\ u_{\mathrm{it},i} & \text{if } S_{\min} \le u_{\mathrm{it},i} \le S_{\max} \end{cases} \tag{3}$$

2. Determine the $\Delta Z^* = Z\left(u_{\mathrm{it},i}\right) - Z^*$ and $\Delta Z = Z\left(u_{\mathrm{it},i}\right) - Z\left(S_{\mathrm{it},i}\right)$.
3. If $\Delta Z^* \le 0$, set $S_{\mathrm{best}} = u_{\mathrm{it},i}$, $Z^* = Z(S_{\mathrm{best}})$.
4. If $\Delta Z \le 0$, set $S_{\mathrm{it+1}} = u_{\mathrm{it},i}$.
5. If $\Delta Z > 0$, update the S_i with the solution u with probability $\exp(-\frac{\Delta Z}{T})$.

end for

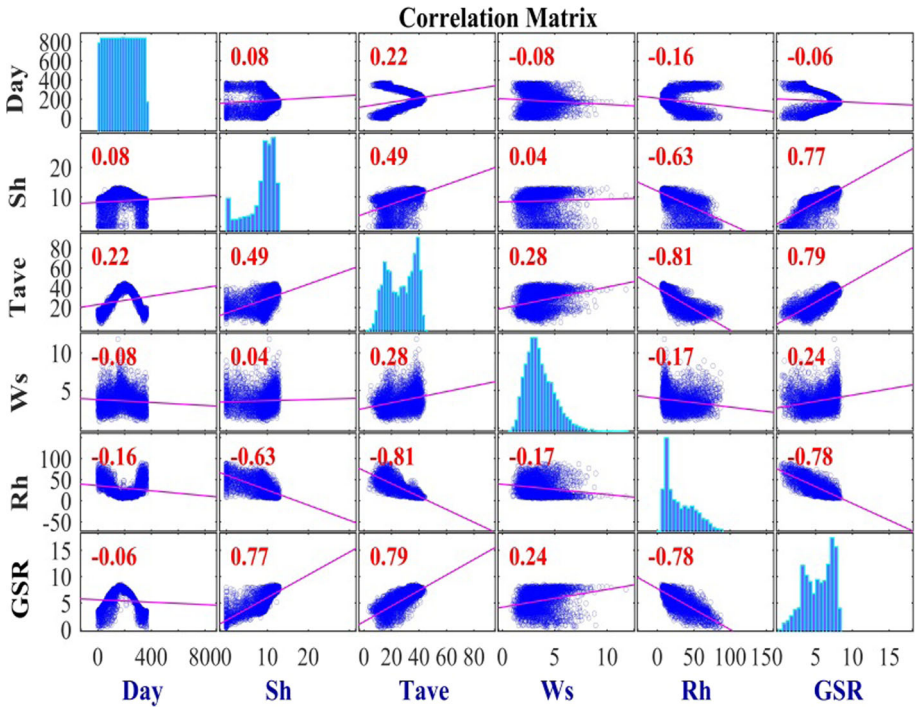

Fig. 4 The correlation matrix of the implemented variables

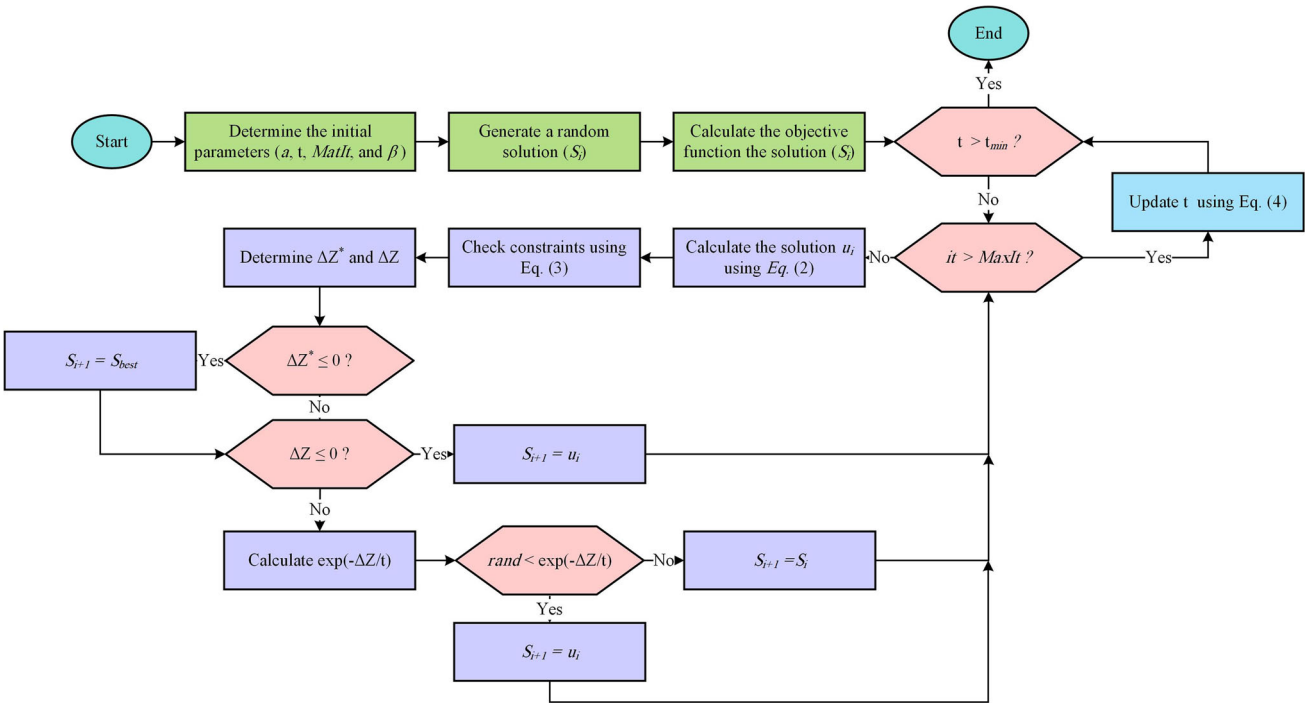

Fig. 5 Flowchart of the ISA algorithm

Decrease the T utilizing the following equation,

$$t = \beta \times t_{\max} \tag{4}$$

end

Stage 3: Presentation of the best solution S_{best}.

It is noteworthy that in Eq. (4), β is a constant number in the range of (0.5, 1). The flowchart of the ISA algorithm is displayed in Fig. 5.

2.3.2 Least square support vector machine (LSSVM)

The LSSVM is a different version of the support vector machine (SVM), which was presented by Suykens and Vandewalle [20]. The LSSVM employs a set of linear equations to increase its convergence speed, while the SVM uses a quadratic programming technique for training [52]. The simple structure and high speed of LSSVM convergence make it widely used in regression and classification fields [52, 53]. In this model, the training dataset is described by (x_m, y_m), $m = 1, 2, \ldots, M$, where x_n and y_m are the input and output dataset. The main formulation of the LSSVM is expressed as [54]:

$$\text{Minimize}_{\omega,c,\eta}\ Z(\omega,\eta) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{n=1}^{M}\eta_n^2 \tag{5}$$

$$\text{Subjectto :}$$

$$y_m = \omega^T\phi(x_m) + c + \eta_m, m = 1,2,\ldots,M \tag{6}$$

where ω^T is the vector of the transposed output layer, γ is a penalty parameter, η is a regression error, $\phi(x_m)$ is a nonlinear function, and c is a bias parameter to be estimated. Using the Lagrange method, Eqs. (5) and (6) can be defined as,

$$L(\omega, c, \eta, \alpha) = Z\left(\omega,\eta) - \sum_{m=1}^{M}\alpha_m(\omega^T\phi(x_m) + c + \eta_m - y_m\right) \tag{7}$$

where α_m is Lagrange multiplier. Based on the Karush–Kuhn–Tucker (KKT) conditions, the following solutions are achieved:

$$\frac{\partial L}{\partial \omega} = 0 \rightarrow \omega = \sum_{m=1}^{M}\alpha_m\phi(x_m) \tag{8}$$

$$\frac{\partial L}{\partial c} = 0 \rightarrow \sum_{m=1}^{M}\alpha_m = 0 \tag{9}$$

$$\frac{\partial L}{\partial \eta_m} = 0 \rightarrow \alpha_m = \gamma\eta_m \tag{10}$$

$$\frac{\partial L}{\partial \alpha_m} = 0 \rightarrow \omega^T\phi(x_m) + c + \eta_m - y_m = 0 \tag{11}$$

By eliminating ω and η_m, the following equations can be obtained,

$$\begin{bmatrix} 0 & e_m^T \\ e_m & \text{Kernel} + \gamma^{-1}e \end{bmatrix}\begin{bmatrix} c \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{12}$$

where $e_m = [1, \ldots, 1]^T$, $\alpha = [\alpha_1, \ldots, \alpha_m]^T$, $y = [y_1, \ldots, y_m]^T$, and e is the unit matrix. $Kernel$ is the kernel functions, which is expressed as,

$$K(x_m, x_i) = \phi(x_m)\phi(x_i) \tag{13}$$

Radial basis functions (RBF) are used as the kernel function in this research, which is defined as:

$$K(x_m,x_i) = \exp\left(\frac{-||x_m, x_i||}{\delta^2}\right) \tag{14}$$

where δ is a constant parameter, which is determined by the ISA algorithm.

In this paper, the ISA algorithm was employed to optimize the LSSVM parameters (i.e., δ and γ). The proposed method is called LSSVM-ISA. Figure 6 demonstrates the flowchart of the LSSVM-ISA algorithm. The mathematical formulation and details of the MARS, MLRI, and GRNN models are available in Appendix 1. Also, the formulations of statistical metrics used in this study are provided in Appendix 2. Figure 7 illustrates the flowchart of the proposed procedure and a brief algorithm of each data-driven approach used in the daily GSR estimating process.

3 Result and discussion

For the development of predictive models, the employed influential input parameters at Ahvaz station include the day of the year [1 365] (366 for leap year), daily sunshine hours (S_h), mean daily air temperature (T_{ave}), mean daily wind velocity (W_s), daily relative humidity (R_h), and the daily global solar radiation (GSR) was considered as subjective models. In this study, a novel robust data-driven model, namely the LSSVM, together with improved simulated annealing (ISA) approach (LSSVM-ISA), is provided to predict the global solar radiation (GSR) accurately. Three other AI-based models validate the proposed method, namely MARS, GRNN, and MLRI.

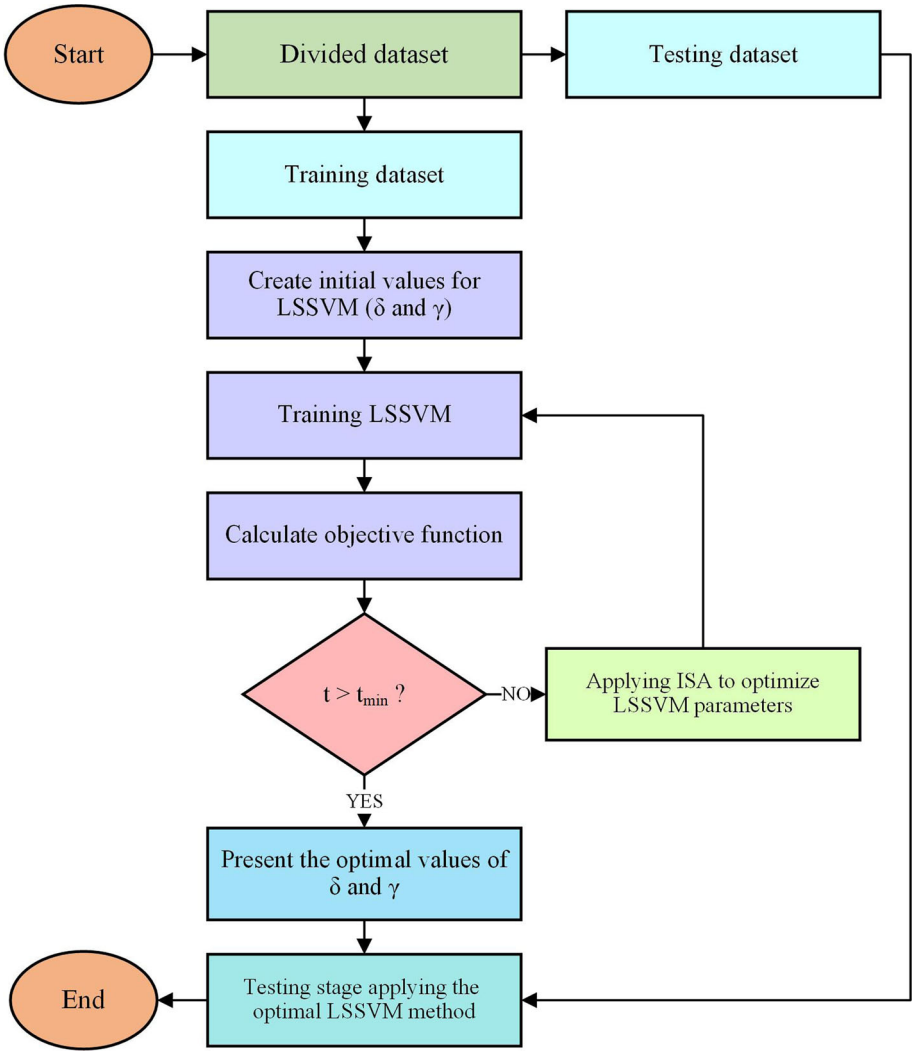

Fig. 6 Flowchart of the LSSVM-ISA method

To measure the best combination of predictor variables, a combination consisting of all influential variables and five combinations obtained by successively excluding the influence of each predictor were examined for each predictive model by four metrics consisting of the R, RMSE, MAPE, and NS. The best possible accuracy of the estimated datasets could be achieved when $R = 1$, RMSE $=$ 0, MAPE $< 10\%$, and NS $= 1$. According to the results of the first six combinations, the day and daily sunshine hours were utilized continuously for constructing other combinations due to having a significant impact on the results. According to the above explanations, a total of 10 combinations were examined for each data-driven model.

Exploration of the results of the ten combinations is tabulated in Table 3. As tabulated in Table 3, we can recognize that the LSSVM-ISA model with the combination of four factors consisting of day, S_h, $T_{ave,}$ and R_h has revealed the best performance in estimating the daily GSR with the highest R (0.973 and 0.980) and NS (0.947 and 0.957), and lowest *RMSE* (0.442 and 0.391 kWh/m^2/day) and *MAPE* (8.171% and 8.233%) for training and testing modes.

Besides, the implemented kernel functions in LSSVM-ISA consisting of two tuning parameters (λ, σ) are optimized for each combination by improved simulated annealing (ISA), which are tabulated in Table 4. An open MATLAB toolbox ARESLab was utilized to develop the MARS model in the current work. To verify the robustness and efficiency of the MARS performance model, tenfold cross-validation was considered. The maximum number of basis functions in the forward building stage was selected between 15 and 30 numbers by a trial-and-error procedure to provide the MARS model. After the backward step and removing the over-fitting model, some of the basis functions were eliminated. Table 5 shows the adjustment parameters of the MARS model.

Likewise, Table 6 sums the results achieved from the MARS model, which showed that the combination of 7 comprising day, $S_{h,}$ and R_h outperformed the other combinations in terms of R (0.968 and 0.974), RMSE (0.485 and 0.426 kWh/m^2/day), MAPE (9.596% and 9.003%), and NS (0.937 and 0.948) for training and testing modes, respectively.

The corresponding basis functions $B_{F,i}$ and the coefficients β_i of the MARS model for the optimum combination (No.7) are listed in Table 7.

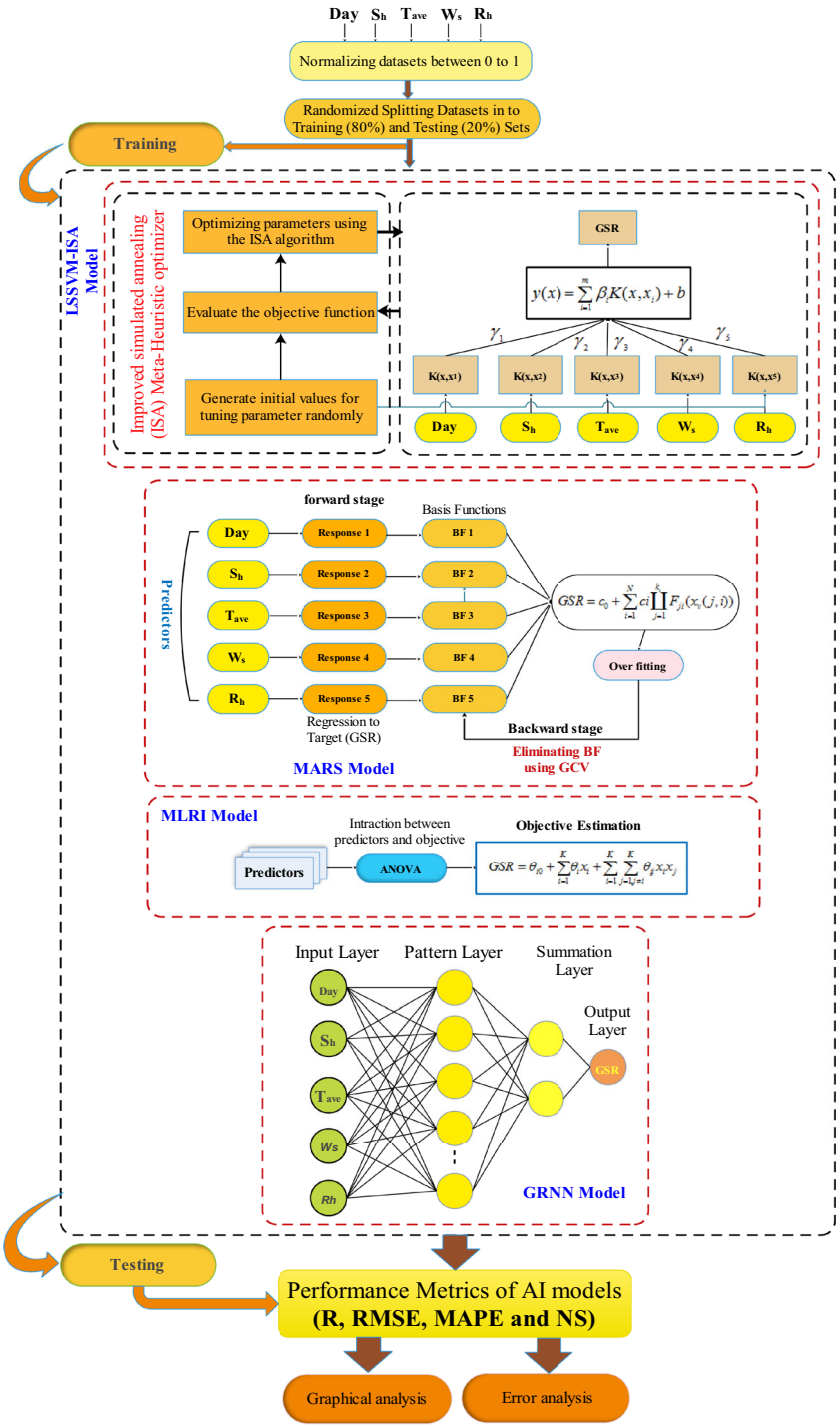

Fig. 7 The flowchart of data-driven models to estimate the daily GSR

In the GRNN model, the most significant criterion is the spread values obtained by a trial-and-error process for each combination. The spread values for each combination are summarized in Table 8. The statistical metrics obtained from the GRNN model to predict the daily GSR are tabulated in Table 9. In the GRNN model, the combination (No.10) including day and S_h provided more precise results in terms of R (0.964 and 0.971) and RMSE (0.526 and 0.469 kWh/m^2/day), MAPE (11.042% and 11.384%), and NS (0.926 and 0.938) in training and testing phases, respectively, than the other combinations in assessing the daily GSR.

In the last alternative, the MLRI model is surveyed in both training and testing procedures based on the correlation (R), RMSE, MAPE, and Nash–Sutcliffe efficiency (NS) for ten input combinations. As mentioned before, in the MLRI method, the relevant analysis of variance (ANOVA) is performed for each combination based on the number of predictor parameters and how they interact. Assessments of 10 combination schemes are given in Table 10. The statistical analysis demonstrates that in the MLRI model, the

Table 3 Performance of the estimating skills of the LSSVM-ISA model with various input variables over the training and testing phase

Model	Inputs	Train				Test			
		R	RMSE	MAPE%	NS	R	RMSE	MAPE%	NS
1	All	0.973	0.446	8.246	0.946	0.978	0.399	8.429	0.955
2	All-day	0.940	0.659	12.639	0.883	0.935	0.687	14.055	0.866
3	All-S_h	0.930	0.710	14.481	0.864	0.935	0.672	14.687	0.872
4	All-T	0.973	0.445	8.194	0.947	0.978	0.414	8.466	0.951
5	All-W_s	0.973	0.442	8.171	0.947	0.980	0.391	8.233	0.957
6	All-R_h	0.971	0.464	8.621	0.942	0.976	0.421	9.103	0.950
7	All-T-W_s	0.975	0.431	7.942	0.950	0.977	0.420	8.682	0.950
8	All-T-R_h	0.971	0.458	8.493	0.944	0.974	0.445	9.435	0.944
9	All-W_s-R_h	0.972	0.449	8.311	0.946	0.976	0.425	9.085	0.949
10	All-W_s-R_h-T	0.972	0.453	8.647	0.945	0.976	0.431	9.381	0.947
Optimum	All-W_s	0.973	0.442	8.171	0.947	0.980	0.391	8.233	0.957

Table 4 The tuning parameters of LSSVM for each combination

Model	Inputs	LSSVM-ISA	
		σ	γ
1	All	25.459	577.538
2	All-day	1.523	2.700
3	All-S_h	2.993	4.819
4	All-T	8.697	272.880
5	All-W_s	4.579	16.475
6	All-R_h	8.243	68.488
7	All-T-W_s	0.995	4.497
8	All-T-R_h	1.574	9.171
9	All-W_s-R_h	1.173	6.998
10	All-W_s-R_h-T	0.227	1.025

σ = kernel width; γ = loss-function parameter

Table 5 Setting parameter of the MARS model for estimating GSR

Parameter	Value
Maximum basis functions	15–30
Maximum self-interactions	0
Maximum degree of interactions	2–3
Threshold	0
Prune	Yes

combination (No.6) consists of all predictors excluding the relative humidity (R_h) achieved the best prediction by the highest R (0.945 and 0.958) and NS (0.892 and 0.902) in training and testing stages, respectively. The ANOVA associated with the optimum combination (No. 6) is given in Table 11, which reveals that the term (Day × W_s) regarding p-value and F-value is less effective than other terms and can be eliminated from the estimating process of daily GSR. Similarly, while the concept of this model is based on the linear dependence between predictors and target, the average daily temperature (T_{ave}) and sunshine hours (S_h) with the highest Pearson correlation coefficients ($r_p = 0.79$ and 0.77, respectively) have the most influence on obtained correlation as below equation:

$$\begin{aligned} GSR = &-2.1812 + 0.0053131Day + 0.36946S_h + 0.25816T_{ave} + 0.090826W_s \\ &- 0.0002626Day \times S_h - 0.00039114Day \times T_{ave} - 0.0049445S_h \times T_{ave} - 0.00017948Day \times W_s \\ &+ 0.022838S_h \times W_s - 0.0072242T_{ave} \times W_s \end{aligned} \quad (15)$$

Table 12 summarizes the best performance of each data-driven model among all input combinations, which demonstrated that the LSSVM-ISA outperformed the MARS, GRNN, and MLRI in estimating daily GSR in both the testing and training stage. Also, MARS and GRNN are the second and third predictive models for simulating daily GSR, respectively.

The results in Tables 3, 6, 9, and 10 indicate that the optimal input combination in each data-driven model is different due to the discrepancy in the performance mechanism of each of them. Figure 8 illustrates the bar plot of R, RMSE, MAPE, and NS values for

Table 6 Performance of the estimating skills of the MARS model with various input variables over the training and testing phase

Model	Inputs	Train				Test			
		R	RMSE	MAPE%	NS	R	RMSE	MAPE%	NS
1	All	0.968	0.487	9.480	0.936	0.973	0.433	9.409	0.947
2	All-day	0.928	0.717	14.118	0.862	0.930	0.692	14.109	0.864
3	All-S_h	0.920	0.754	15.908	0.847	0.930	0.691	15.011	0.865
4	All-T_{ave}	0.968	0.485	9.596	0.937	0.974	0.426	9.003	0.948
5	All-W_s	0.968	0.487	9.490	0.936	0.973	0.433	9.409	0.947
6	All-R_h	0.967	0.489	9.563	0.936	0.972	0.444	9.916	0.944
7	All- T_{ave} -W_s	0.968	0.485	9.596	0.937	0.974	0.426	9.003	0.948
8	All- T_{ave} -R_h	0.968	0.485	9.505	0.937	0.972	0.444	9.923	0.944
9	All-W_s-R_h	0.967	0.489	9.563	0.936	0.972	0.444	9.916	0.944
10	All-W_s-R_h- T_{ave}	0.968	0.485	9.505	0.937	0.972	0.444	9.923	0.944
Optimum	All- T_{ave} -W_s	0.968	0.485	9.596	0.937	0.974	0.426	9.003	0.948

Table 7 Basis functions and coefficients were obtained to estimate the daily GSR in the optimum combination (day, S_h, R_h) using the MARS model

Basic function	Equation	Coefficient β_i
Intercept	–	8.3675
$B_1(x)$	$B_{F1} = \max(0, \text{Day} - 181)$	− 0.028177
$B_2(x)$	$B_{F2} = \max(0, 181 - \text{Day})$	− 0.023802
$B_3(x)$	$B_{F3} = \max(0, S_h - 10.8)$	− 0.36892
$B_4(x)$	$B_{F4} = \max(0, 10.8 - S_h)$	− 0.78284
$B_5(x)$	$B_{F5}=B_{F2} \times \max(0, R_h - 31.23)$	− 3.8718E-05
$B_6(x)$	$B_{F6} = B_{F2} \times \max(0, 31.23 - R_h)$	0.00022022
$B_7(x)$	$B_{F7}=B_{F4} \times \max(0, R_h - 25.41)$	− 0.0014453
$B_8(x)$	$B_{F8} = B_{F4} \times \max(0, 25.41 - R_h)$	0.0086739
$B_9(x)$	$B_{F9}=B_{F4} \times \max(0, 317 - \text{Day})$	0.0026546
$B_{10}(x)$	$B_{F10} = B_{F3} \times \max(0, Day - 196)$	0.014655
$B_{11}(x)$	$B_{F11} = B_{F3} \times \max(0, 196 - Day)$	0.0079019
$B_{12}(x)$	$B_{F12} = B_{F4} \times \max(0, Day - 133)$	0.0030269
$B_{13}(x)$	$B_{F13} = B_{F4} \times \max(0, 133 - Day)$	− 0.0017473

Table 8 The spread values of GRNN for each combination

Inputs combination	All	All-day	All-S_h	All-T	All-Ws
Spread value	0.1	0.05	0.05	0.05	0.05
Inputs combination	All-R_h	All-T-W_s	All-T-R_h	All-W_s-R_h	All-W_s-R_h-T
Spread value	0.05	0.05	0.05	0.05	0.05

all input combinations and provided models. Based on the shown metric values in Fig. 8, the LSSVM-ISA has the best predictive performance in all combinations, whereas the MARS and GRNN models are ranked as the best second and third predictive models, respectively. However, MARS and GRNN in the combination No. 2 have similar performance in estimating the daily GSR. It can be mentioned that MLRI cannot perform better than the last model in terms of the accuracy of all the proposed data-driven models.

The performances of the LSSVM-ISA, MARS, GRNN, and MLRI models were compared by the daily observed and simulated GSR plots over the training and testing periods (Figs. 9, 10, 11, and 12). As can be seen in these figures, the GSR values predicted by the LSSVM-ISA model are better with the observed data than those of MARS, GRNN, and MLRI models in both testing and training stages. Moreover, the scatter plot of the GSR data points simulated by the LSSVM-ISA model is closer to the 1:1 line than those from the other developed data-driven models, which implies the LSSVM-ISA model's superiority in estimating daily GSR. Moreover, a closer look at the distribution of points at scatter plots reveals that for the range of GSR>4, all provided AI-based models have promising performance in the simulation of daily GSR, especially in testing mode.

For visual comparison between the results of the LSSVM-ISA and those of other data-driven approaches, Fig. 13 shows estimated and observed GSR values over the two selected intervals of (1 June 2013–31 December 2013) and (1 January 2017–1 July 2019) at the training and testing periods, respectively. It can be seen that the LSSVM-ISA performs better than other predictive models due to

Table 9 Performance of the estimating skills of the GRNN model with various input variables over the training and testing phase

Model	Inputs	Train				Test			
		R	RMSE	MAPE%	NS	R	RMSE	MAPE%	NS
1	All	0.974	0.436	8.652	0.949	0.967	0.495	10.792	0.931
2	All-day	0.967	0.494	8.543	0.934	0.927	0.724	14.443	0.851
3	All-S_h	0.948	0.615	11.790	0.898	0.917	0.752	16.142	0.840
4	All-T	0.990	0.275	4.992	0.980	0.967	0.484	9.781	0.933
5	All-W_s	0.985	0.329	6.001	0.971	0.970	0.471	9.740	0.937
6	All-R_h	0.986	0.322	6.116	0.972	0.970	0.474	10.239	0.936
7	All-T-W_s	0.965	0.511	10.518	0.930	0.966	0.488	10.714	0.933
8	All-T-R_h	0.966	0.515	10.832	0.929	0.970	0.477	11.384	0.935
9	All-W_s-R_h	0.965	0.510	10.338	0.930	0.970	0.481	11.289	0.934
10	All-W_s-R_h-T	0.964	0.526	11.042	0.926	0.971	0.469	11.384	0.938
Optimum	All-W_s-R_h-T	0.964	0.526	11.042	0.926	0.971	0.469	11.384	0.938

Table 10 Performance of the estimating skills of MLRI model with various input variables over the training and testing phase

Model	Inputs	Train				Test			
		R	RMSE	MAPE%	NS	R	RMSE	MAPE%	NS
1	All	0.949	0.612	11.231	0.899	0.956	0.584	11.567	0.903
2	All-day	0.911	0.796	15.209	0.829	0.919	0.772	15.525	0.831
3	All-S_h	0.883	0.918	19.692	0.773	0.883	0.902	19.784	0.769
4	All-T	0.901	0.841	15.087	0.810	0.899	0.849	15.360	0.796
5	All-W_s	0.947	0.625	11.435	0.895	0.955	0.599	12.080	0.898
6	All-R_h	0.945	0.634	11.585	0.892	0.958	0.587	12.706	0.902
7	All-T-W_s	0.896	0.861	15.323	0.801	0.895	0.867	15.417	0.787
8	All-T-R_h	0.800	1.163	21.328	0.636	0.847	1.016	21.315	0.707
9	All-W_s-R_h	0.943	0.642	11.735	0.889	0.957	0.596	12.794	0.899
10	All-W_s-R_h-T	0.771	1.234	22.510	0.590	0.834	1.043	21.313	0.691
Optimum	All-R_h	0.945	0.634	11.585	0.892	0.958	0.587	12.706	0.902

Table 11 ANOVA results identify the influence of predictor's interaction in the MLRI model for the combination No.6

Terms	Sum of squares	F-value	p-value
Day	535.37	2572.2	0
S_h	2222.6	10,672	0
T_{ave}	2616.2	12,561	0
W_s	6.635	31.858	1.818E-08
Day × S_h	21.044	101.04	2.16E-23
Day × T_{ave}	234.15	1124.2	0
T_{ave} × S_h	50.380	241.92	0
Day × W_s	1.244	5.972	0.0146
W_s × S_h	23.347	112.10	0
W_s × T_{ave}	13.397	64.321	1.510E-15

Table 12 Comparison between the outcomes from 4 data-driven models to estimate daily GSR

Model	Optimum combination	Train				Test			
		R	RMSE	MAPE%	NS	R	RMSE	MAPE%	NS
LSSVM-ISA	All-W_s	0.973	0.442	8.171	0.947	0.980	0.391	8.233	0.957
MARS	All-T-W_s	0.968	0.485	9.596	0.069	0.974	0.426	9.003	0.056
GRNN	All-W_s-R_h-T	0.964	0.526	11.042	0.926	0.971	0.469	11.384	0.938
MLRI	All-R_h	0.945	0.634	11.585	0.892	0.958	0.587	12.706	0.902

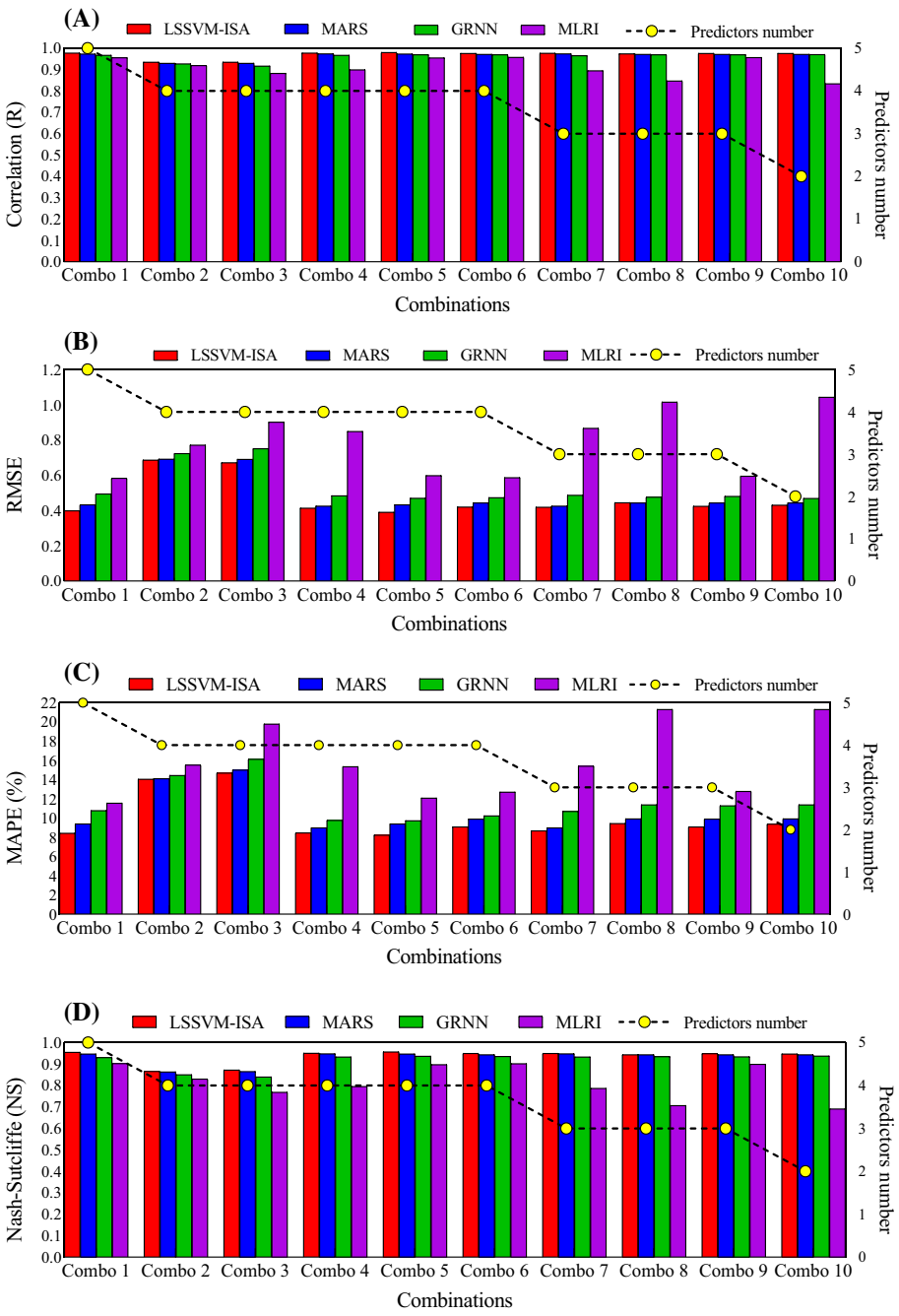

Fig. 8 The correlation (R), RMSE, MAPE, and NS values for all AI models in 10 combinations versus predictors number

having the best fit and lowest overestimation and underestimation. However, as the next alternatives, the MARS and GRNN models can simulate the daily GSR at an acceptable level of precision. The violin plot in Fig. 14 shows the distribution of standardized observed and estimating daily global solar radiation (GSR) using the LSSVM-ISA, MARS, GRNN, and MLRI models for testing and training modes in both testing and training modes of the best input combinations. Distribution of predicted GSR values in violin plot for training (Fig. 14A) and testing (Fig. 14B) stages reveals that all predictive models have relatively similar estimating performance in the median 50 percentile (thick solid line). In contrast, the LSSVM-ISA model is energetically superior to other data-driven models in upper and lower quartiles (dashed lined) of violin plots compared to observational daily GSR. Also, the MARS model with the second-best distribution consistency with observational values has acceptable performance in estimating daily GSR.

In the next validation graphical tools, the Taylor diagrams for evaluating the robustness of the developed predictive models in training and testing period of the simulation are plotted in Fig. 15. According to findings in Fig. 15, the locus of data obtained from the LSSVM-ISA model is in the closet distance to the target point compared to those of the other data-driven methods in both training and testing phases. This fact is due to having the highest correlation coefficients (R = 0.974 and 0.980), and the minimal

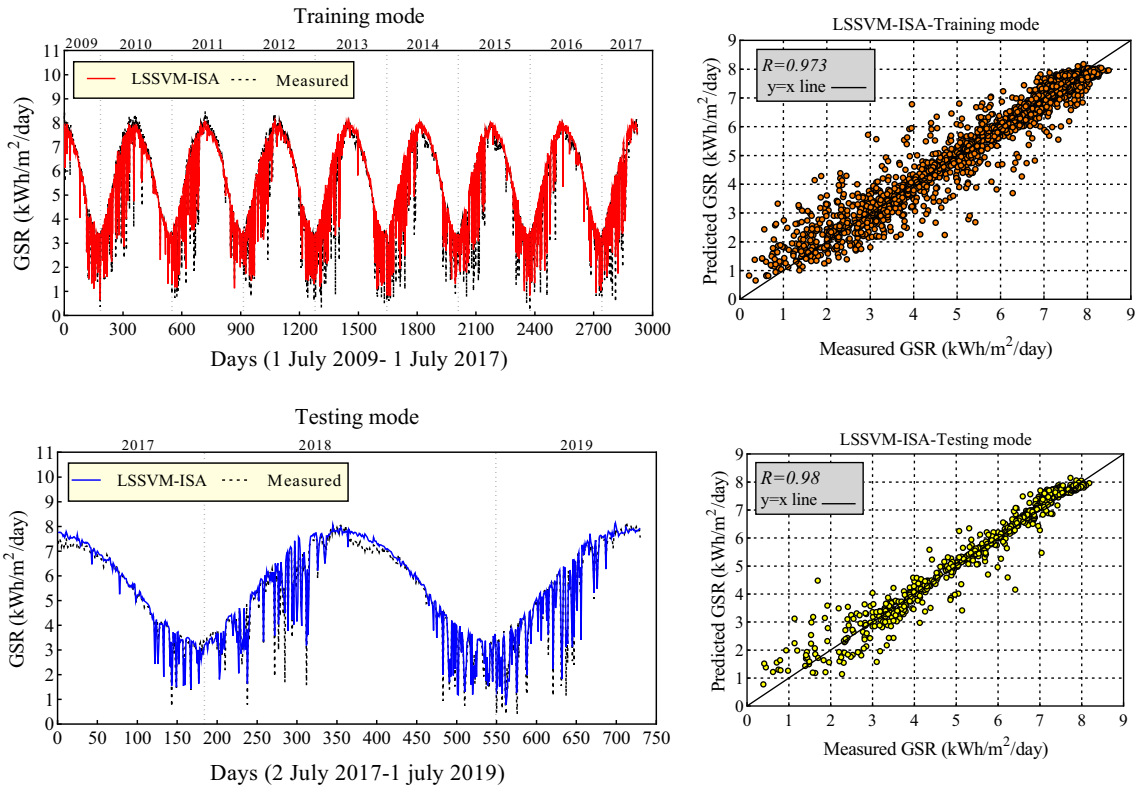

Fig. 9 The scatter plots of GSR (right); comparison of the measured and predicted GSR (left) using of LSSVM-ISA model for both training and testing datasets

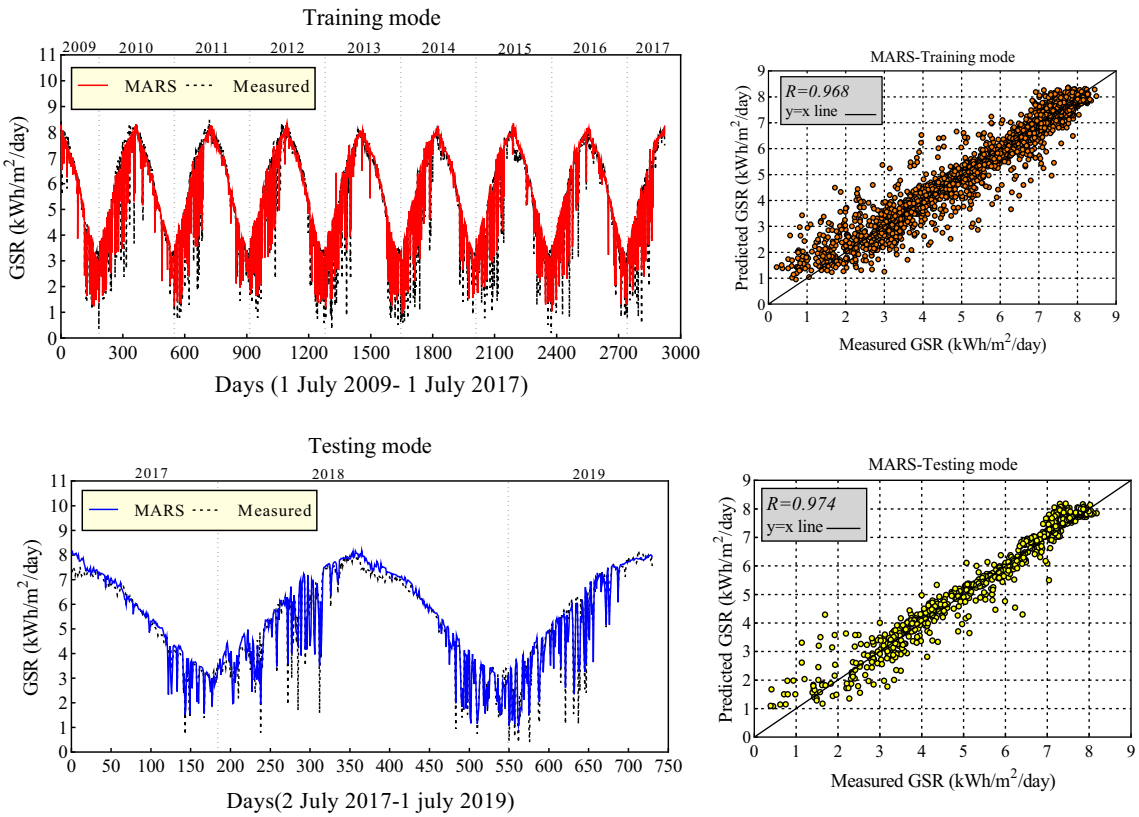

Fig. 10 The scatter plots of GSR (right); comparison of the measured and predicted GSR (left) using of MARS model for both training and testing datasets

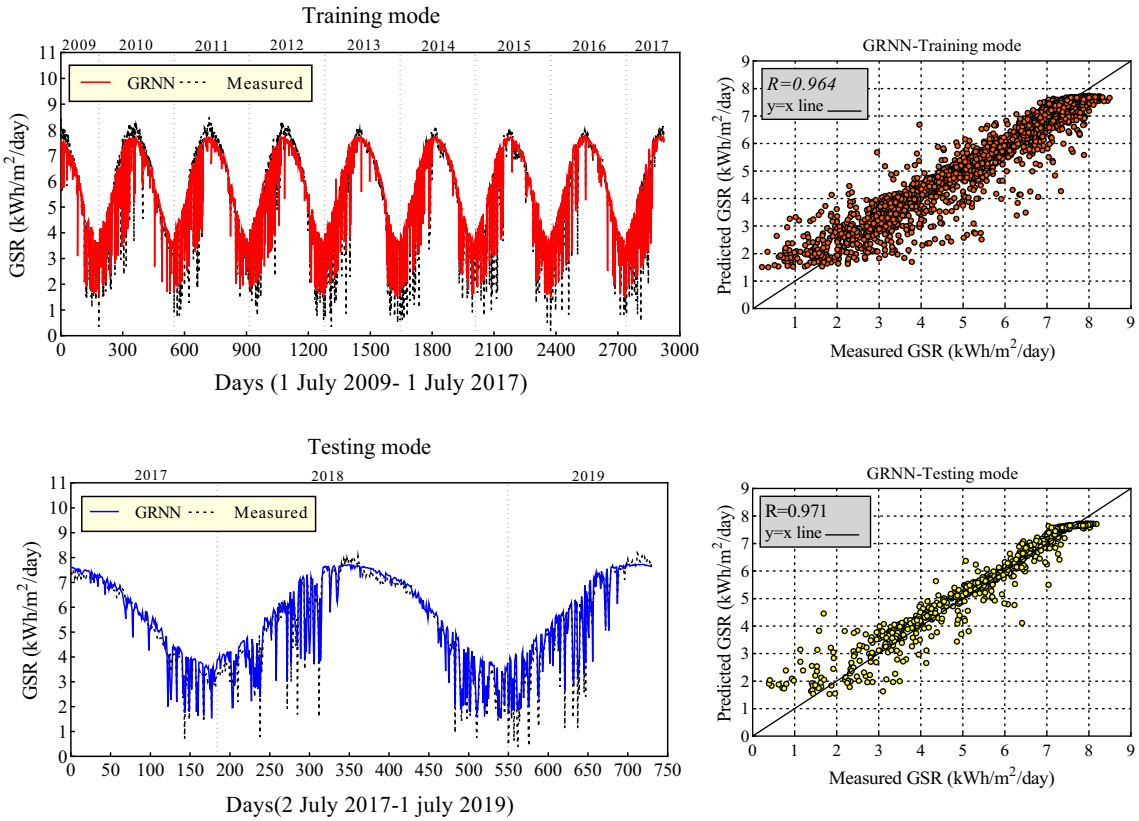

Fig. 11 The scatter plots of GSR (right); comparison of the measured and predicted GSR (left) using of GRNN model for both training and testing datasets

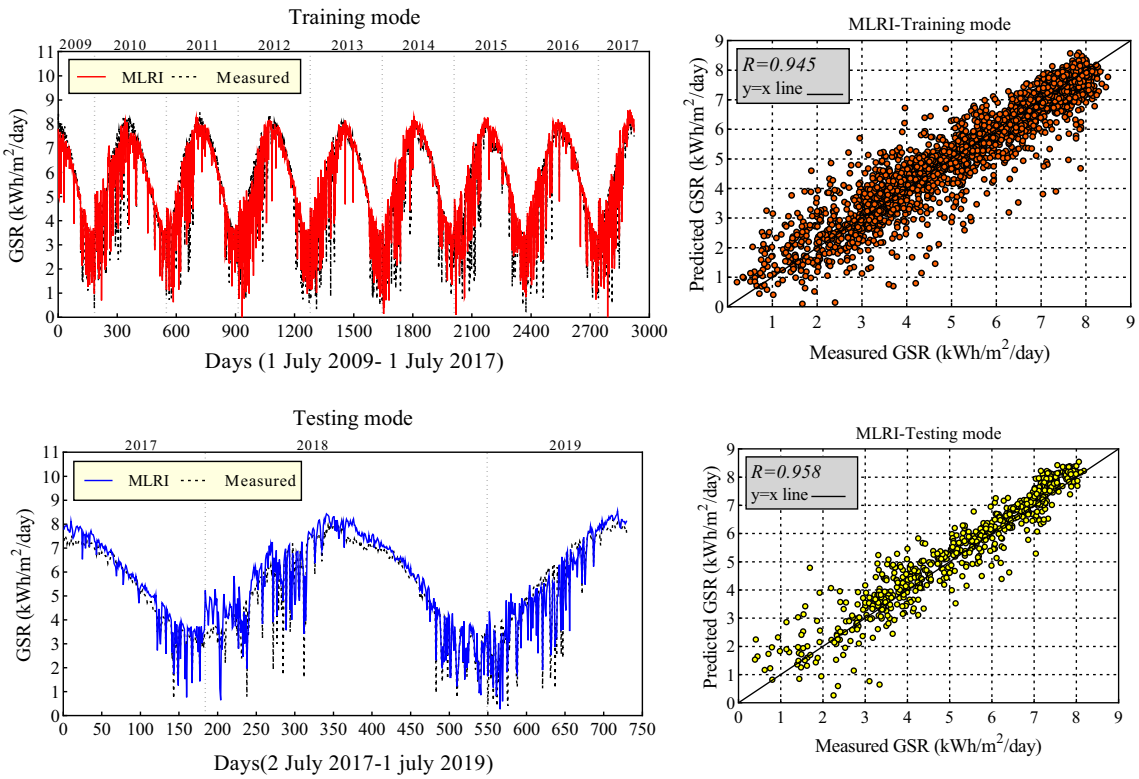

Fig. 12 The scatter plots of GSR (right); comparison of the measured and predicted GSR (left) using of MLRI model for both training and testing datasets

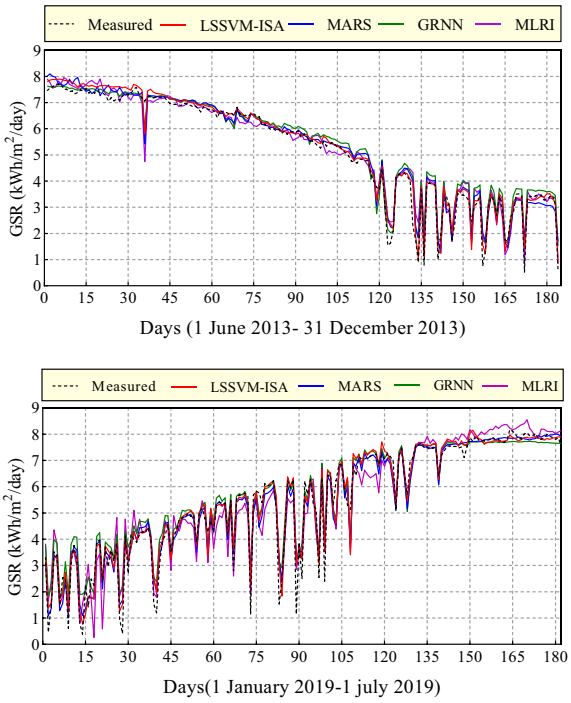

Fig. 13 Comparison of the predictive and measured daily solar radiation with all developed AI models throughout (1 June 2013–31 December 2013) and (1 January 2017–1 July 2019) at the training and testing datasets

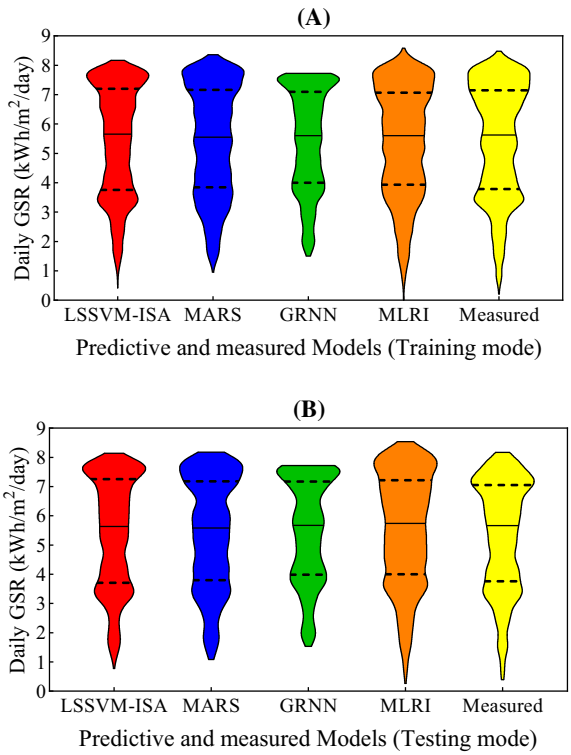

Fig. 14 The violin plots of the predicted and measured daily GSR values for the best combination of each data-driven model

difference of standard deviations ($S_d = 1.887$ and 1.873). The MARS and MLRI models have closer standard deviation to the target than the GRNN model. It is noteworthy that the standard deviation (S_d) of the observed GSR values for the training and testing datasets is 1.880 and 1.738, respectively.

In this part of the study, a comprehensive error analysis was conducted to evaluate the accuracy and reliability of current AI models. The relative deviation value by formula ($D_r = (\mathrm{GSR_i} - \mathrm{GSR_o})/\mathrm{GSR_o}$) was displayed versus observed GSR for the training and testing period of simulation for each model in Fig. 16. Figure 16 demonstrates that the range of the relative deviation for the LSSVM-ISA model by $-3.35 \leq D_r \leq 0.526$ is less than the MARS by $-5.80 \leq D_r \leq 0.582$, GRNN by $-6.20 \leq D_r \leq 0.550$,

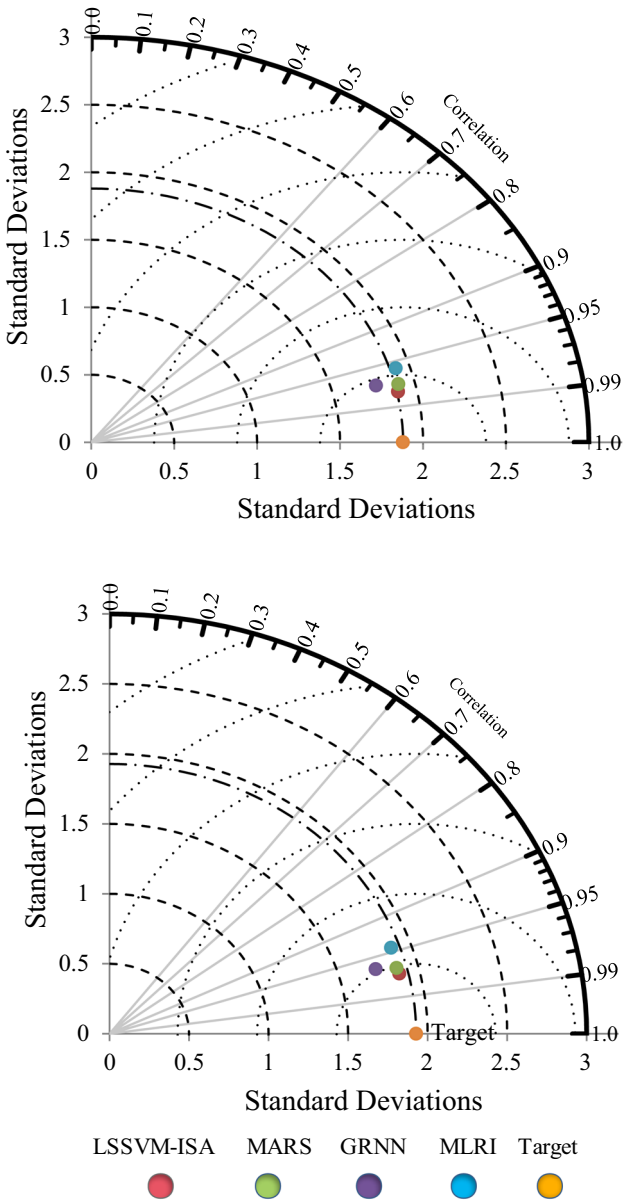

Fig. 15 The Taylor diagram of the training and testing modes for LSSVM-ISA, MARS, GRNN, and MLRI to estimate the GSR

and for MLRI by $-4.47 \leq D_r \leq 1.260$. Close examination of the error distribution plots vividly shows that for $GSR > 4$ all understudy predictive models have reliable performance and acceptable accuracy in estimating daily GSR and the relative deviation restricted in the range of $-1 \leq D_r \leq 0.520$.

Finally, a substantial error criterion is carried out to better understand the quantitative performance of all proposed models, which expresses the amount of cumulative absolute percent of error frequency (Fig. 17). LSSVM-ISA model can predict 75% of the daily GSR data with absolute relative error less than 6.1%, whereas 75% of the MARS, GRNN, and MLRI models as the second, third, and fourth most accurate approaches provided prediction errors less than or equal to 8.2%, 9.74%, and 11.56%, respectively. Also, all the equipped models can assess 90% of the data with an absolute relative error of less than 25%, which implies the reliability of all data-driven approaches for the prediction of daily GSR. It is vividly clear that the results of LSSVM-ISA are in the most satisfactory agreement with observed data sets in comparison with all the provided models.

4 Validation of the model with traditional approaches

In order to investigate the accuracy of the empirical equations in comparison with the AI-based methods used in the research, nine empirical equations were examined and evaluated, which are introduced in Table 1. The inputs of the selected equations are the

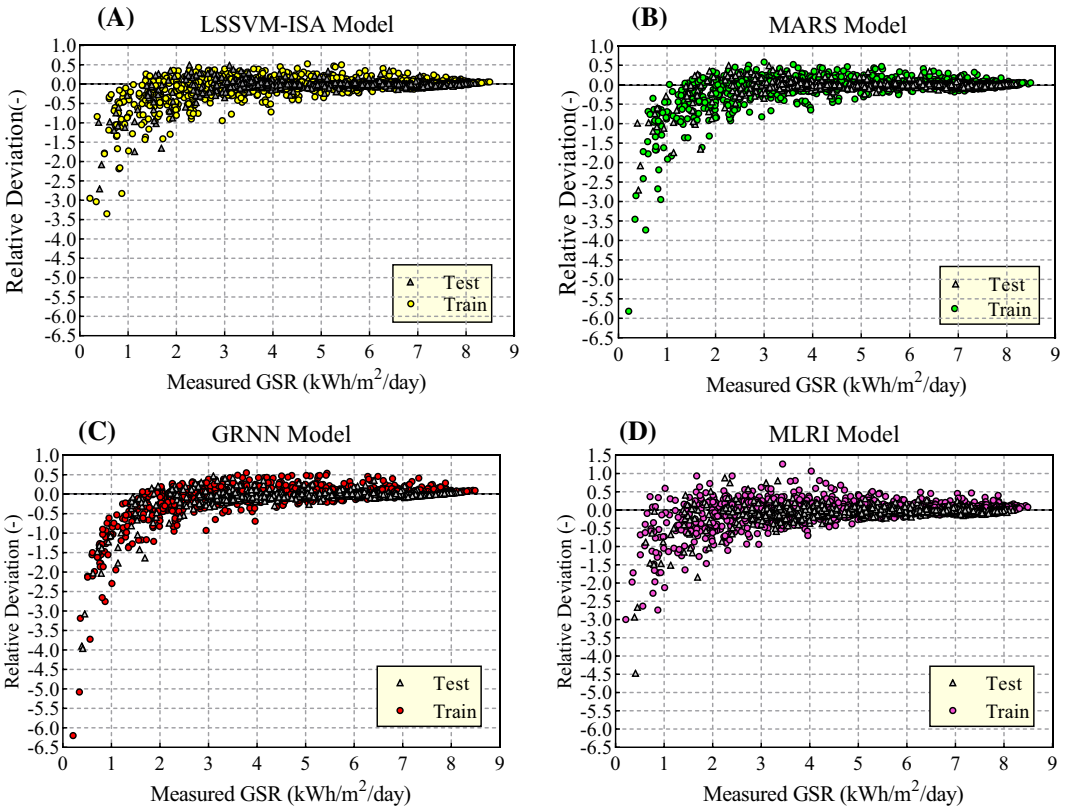

Fig. 16 The relative deviation plots of the developed data-driven models for predicting GSR: **a** LSSVM-ISA, **b** MARS, **c** GRNN, and **d** MLRI

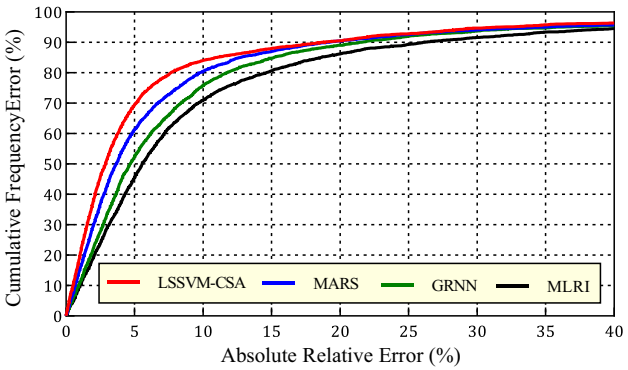

Fig. 17 Percentage of cumulative frequency of proposed AI models versus the absolute percentage of relative error in estimating GSR

same as the AI-based models utilized in this research. The coefficients of the empirical equations are calculated using the ordinary least square (OLS) method separately, and their values are given for each equation in Table 13. According to Table 13, the Bristow and Campbell [39] equation with a mean absolute percentage error (MAPE) of 14.811% provides more accurate results than other equations. The MAPE of Goodin et al. [44] and Hunt equations [43] is 14. 843% and 14.857%, respectively.

Table 14 shows a comparison between the best results of each model. According to Table 14, AI-based models have performed better than the empirical and MLRI models. LSSVM-ISA model with MAPE = 8.233% has the least error among models.

Table 15 shows MAPE percent error for different empirical equations used in this research and error improvement by the best LSSVM-ISA model. The results of Table 15 show that the best LSSVM-ISA model has reduced the MAPE error rate compared to empirical models significantly. Comparing the best practical model (Bristow and Campbell [39]) and the best LSSVM-ISA model shows a 6.578% reduction in MAPE percent.

Table 13 Coefficient and accuracy of empirical equations to predict GSR

Model	a	b	c	d	e	R	RMSE	MAPE%	NS
Swartman and Ogunlade [37]	6.08	1.8	− 0.065	–	–	0.814	1.115	21.775	0.662
Hargreaves [38]	0.158	–	–	–	–	0.919	0.754	17.011	0.845
Bristow and Campbell [39]	0.708	0.015	1.818	–	–	0.925	0.728	14.811	0.856
De Jong and Stewart [40]	0.162	0.490	− 0.010	− 0.010	- –	0.925	0.727	15.473	0.856
Allen [41]	0.158	–	–	–	–	0.919	0.754	17.012	0.845
Donatelli and Campbell [42]	0.665	0.006	3.526	–	–	0.911	0.791	15.501	0.830
Hunt [43]	0.159	− 0.053	–	–	–	0.919	0.754	16.984	0.845
Hunt [43]	0.144	0.012	− 0.135 –	0.002	0.079	0.927	0.719	14.875	0.859
Goodin, Hutchinson, Vanderlip, and Knapp [44]	0.681	0.011	2.846	–	–	0.921	0.751	14.843	0.847

Table 14 Comparison between the best empirical correlation and AI models

Model	R	RMSE	MAPE%	NS
Best LSSVM-ISA	0.980	0.391	8.233	0.957
Best MARS	0.974	0.426	9.003	0.948
Best GRNN	0.970	0.471	9.740	0.937
Best MLRI	0.956	0.584	11.567	0.903
Best Empirical	0.925	0.728	14.811	0.856

Table 15 Error improvement by best LSSVM-ISA model compared to empirical equations

Model	MAPE%	Error improvement by LSSVM-ISA %
Swartman and Ogunlade [37]	21.775	13.542
Hargreaves [38]	17.011	8.778
Bristow and Campbell [39]	14.811	6.578
De Jong and Stewart [40]	15.473	7.240
Allen [41]	17.012	8.779
Donatelli and Campbell [42]	15.501	6.968
Hunt [43]	16.984	8.751
Hunt [43]	14.875	6.642
Goodin, Hutchinson, Vanderlip, and Knapp [44]	14.843	6.610

5 Uncertainty

This section examines and evaluates the uncertainty analysis of the utilize models in predicting global solar radiation. The error value of each sample is calculated from the relation $e_j = GSR_p - GSR_o$. The mean value of the error of the relation is calculated using $\overline{e} = \sum_{j=1}^{n} e_j$, and finally, the standard deviation of the estimation error is calculated from the relation $S_{\text{de}} = \sqrt{\left(e_j - \overline{e}\right)^2 / n - 1}$. If the value of $\overline{e}$ is negative, it indicates that the model underestimates the predicted values, and its positive value indicates the overestimation of the model. High values of s_e indicate the high scattering of the model error, and increasing it will increase the uncertainty of the model prediction.

Confidence intervals for the estimation values of the models can be determined using the values $\overline{e}$ and S_e. Considering the $\pm 1.96 S_e$, values of 95% confidence intervals are obtained. Table 16 shows the mean values of estimation error, standard error deviation, and 95% error confidence intervals. In all five models used in the present study, except for the empirical equation, the mean value of the prediction error is positive, which indicates that the model is overestimating predicting values. Among the applied models, the best empirical equation has the lowest mean value of the prediction error with a value of -0.011. Among AI-based models, LSSVM-ISA and MARS have the lowest average error ($\overline{e} = 0.074$). The LMRI method has the highest average error value ($\overline{e} = 0.207$). The lowest value of the uncertainty band ± 0.792 is for the LSSVM-ISA model, which indicates that this model has a lower uncertainty. The highest uncertainty band ± 1.427 is for the Bristow and Campbell [39] equation, which indicates more uncertainty about the empirical model results than other models.

Table 16 Uncertainty estimates for GSR for best of each model

Model	Mean prediction error $\overline{e}$	S_e	Width of uncertainty band
Best LSSVM-ISA	0.074	0.404	±0.792
Best MARS	0.074	0.426	±0.835
Best GRNN	0.132	0.469	±0.919
Best MLRI	0.207	0.587	±1.150
Best Empirical	-0.011	0.728	±1.427

6 Conclusions

Given the importance of accurately estimating solar radiation in the design of solar energy systems, a novel hybrid data-driven approach comprised of LSSVM coupled with improved simulated annealing (ISA) approach (LSSVM-ISA) is developed to accurately predict the daily global solar radiation (GSR) over 10 years (from 01 July 2009 to 1 July 2019) at Ahvaz station, in Iran. In this research, the predictive input variable was the day of the year (1 to 366), average daily temperature (T_{ave}, °C), sunshine hours (S_h, hr), relative humidity ($R_{h,\%}$), average wind velocity (W_s, m/s) at 10 m and GSR (kWh/m^2/day) was considered as a target for ten input combinations. According to the obtained results, LSSVM-ISA model in the combination No. 5 (including day, S_h, $T_{ave,}$ and R_h) for the testing mode outperformed the MARS, GRNN, and MLRI model in estimating daily GSR. The MARS model, in combination No. 7 (including day, S_h, and R_h), was identified as the second accurate predictive approach for daily GSR estimating. Also, the GRNN and MLRI models as third best rank and the weakest model had the best their performances in combination No. 10 (including day and S_h) and combination No.6 (consisting of all predictors excluding the relative humidity), respectively. According to [70], for judgment of the accuracy of the GSR estimating procedure, the LSSVM-ISA and MARS models by having MAPE < 10% are identified as high accurate estimating tools, and GRNN and MLRI models by $10 \leq$ MAPE ≤ 20 are classified as good predictor data-driven models to assess the daily GSR. Besides, a comparison between all models in 10 combinations demonstrated that the day of the year and sunshine duration (S_h) due to existing in any the best combination performance are identified as the most effective predictive variables in the daily GSR estimating process. Besides, comparing the obtained results of the empirical equations with the AI models indicated that the AI models, especially the LSSVM-ISA method, can estimate the GSR more accurately than the empirical methods. Finally, the uncertainty analysis showed that the proposed LSSVM-ISA has the lowest uncertainty (±0.792) compared with the other AI methods. This demonstrates that the proposed method is more reliable and precise to estimate the GSR.

Authors' contribution Mehdi Jamei contributed to conceptualization, modeling, revision editing, and data formal analysis, wrote the main manuscript, provided software, and project leader. Iman Ahmadianfar wrote the main manuscript, reviewing and was involved in methodology and introduction, and Mozhdeh Jamei wrote the main manuscript and contributed to data visualization; Masoud Karbasi wrote the main manuscript and was involved in revision editing. Ali Asghar Heidari contributed to revision editing. Huling Chen was involved in supervision and revision editing.

Funding There is no funding.

Availability of data and materials Not applicable.

Declarations

Conflict of interest There is no conflict of interest.

Ethical approval and consent to participate Not applicable.

Consent for publication Not applicable.

Appendix 1

Multivariate adaptive regression spline (MARS)

Multivariate adaptive regression spline (MARS) scheme, as a nonlinear and nonparametric statistical regression model, is one of the most popular machine learning models, which was first introduced by Friedman (1991) [17]. The MARS model automatically is capable of mapping the intrinsic nonlinearities and interactions between predictors in data without an assumption about the relationships between dependent and predictor variables to predict continuous objective variables accurately [55]. The main concept

of MARS is splitting the predictor training data into several piecewise linear disjoint regions (or splines) by specific connection points (called 'knots'). Generally, the MARS model can adequately result in the flexible continuous modeling by fitting piecewise linear regression belonging to each segment (spline) and predicting the linear and nonlinear objective [17, 56]. Splines, which are known as the basis functions (BFs) by the smoothing process connection point, are capable of capturing the nonlinearities, curvatures, and threshold features based on the piecewise linear functions. MARS model involves two-step forward and backward pruning iterative procedures. During the stepwise forward process, the BFs were selected, and the location of potential knots was explored by an adaptive regression algorithm, which mostly leads to a very complex and over-fitted model [57, 58]. Mathematically, the MARS model can explain the relationship between predictors and output variable as follows:

$$Y = f(x) = \beta_0 + \sum_{i=1}^{N^*} \beta_i \prod_{j=1}^{K} B_{Fji}(x_{v(ji)}) \tag{16}$$

where Y is objective variable; β_0 is the intercept term; β_n is the coefficient vector of ith spline basis functions; $B_{F_{ji}}$ is the basis function; $x_{v(ji)}$ is the independent predictor of ith and jth products; K is the order of interaction limit, and N^* is the number of independent predictors. The piecewise linear basis functions implemented in the MARS model are generally expressed as [59]:

$$B^+_{Fji} = (x - s_{ji})_+ = \begin{cases} x - s_{ji} & x > s_{ji} \\ 0 & \text{therwise} \end{cases} \sqrt{2} \tag{17}$$

$$B^-_{Fji} = (s_{ji} - x)_+ = \begin{cases} s_{ji} - x & x < s_{ji} \\ 0 & \text{therwise} \end{cases} \tag{18}$$

where B^+_{Fji} and B^-_{Fji} are the positive and negative parts of spline functions, respectively, and s_{ji} denotes the knot of the spline (threshold value).

Likewise, in the backward stepwise process, the redundant BFs provided among the previous process were eliminated by implementing the generalized cross-validation (GCV) technique until the 'lack of fit' criterion is a minimum and estimating accuracy enhancement [60]. The GCV value is given as follows [57, 59]:

$$\text{GCVE} = \frac{\sum_{j=1}^{N} (Y_i - \widehat{Y}_i)^2}{N\left(1 - \frac{F+0.5d^*(F-1)}{N}\right)^2} \tag{19}$$

where Y_i is the observed value of output variable of ith; $\widehat{Y}_i$ is ith predicted value by MARS; N = number of datasets; F is the number of basis functions; and d^* is a penalty value for each basis function comprised into the developed model.

Generalization regression neural network (GRNN)

Generalized regression neural network (GRNN), as a probabilistic-based neural network, was firstly proposed by Specht (1991) [61]. GRNN model based on the nonparametric kernel regression network has been widely implemented in the field of classification and regression. It is adequately capable of handling the nonlinear fitting problems with large-scale training samples. Unlike the backpropagation and radial basis function ANN, GRNN has fewer adjustment parameters, and its learning algorithm rarely falls into the local minimum [62]. GRNN has four-layer comprising the input layer, radial layer, regression layer, and an output layer. The architecture of a GRNN model consists of precisely four layers, with a pattern (radial neurons) layer and a summation (regression) layer placed between the input and output layers [63], as illustrated in Fig. 7. The pattern layer contains the clustering of the input data in the training stage, and consequently, the neurons number in that is exactly equal to the number of data samples. Furthermore, the summation layer always has a new neuron in comparison with the output layer, which is devoted to calculating the probability density function, whereas rest of neurons are used for output calculation. Eventually, GRNN due to directly selecting an approximate function between predictors and output variables spends less time than other ANNs [63].

Multivariate linear regression with interactions (MLRI)

The multivariate linear regression with interactions (MLRIs) is an efficient data-driven model that can obtain a regression by considering the interaction between the predictors (x_i) on the independent variable (outcome) [64, 65]. An interaction effect happens when a predictor has a different impact on the dependent output variable, depending on the values of another predictor. MLRI can capture the regression relationship considering some interactions between a dependent variable (Y) and independent variables (x_i) by the following logic [64, 65]:

$$Y = \theta_{i0} + \sum_{i=1}^{K} \theta_i x_i + \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} \theta_{ij} x_i x_j + \varepsilon \tag{20}$$

where θ_{ij} is the interaction coefficient; ε is a random error; and i,j are the index of two considering predictors.

In the framework of the MLRI model, the significance of the interaction effect is specified using analysis of variance (ANOVA). In this process, the p-value or probability value indicates the importance of each integration term in regression by $\alpha = 0.05$, as a criterion, which lower values of p-values or its corresponding F-values show it is the more statistically significant. Whenever the p-values of each interaction term become less than α, interaction term will be preserved in outcome correlation (Y) and vice versa.

Appendix 2

Performance evaluation

To measure the degree of accuracy of the provided models, various visual efficient tools and statistical performance metrics were employed. The graphical tools include scattering plots, error analysis plots, and Taylor diagrams. Taylor diagram is employed for comparing the similarity patterns in polar space between predicted and observed values of GSR as the robust graphical tool. Basically, the Taylor diagram simultaneously demonstrates the correlation coefficient (R), standard deviation (S_d), and centered root mean square error (cRMSE) [66]. The performance criteria consist of the correlation coefficient (R), mean root square error (RMSE), mean absolute percentage error (MAPE), and Nash–Sutcliffe coefficient (NS) [67–69] which are expressed as following relationships:

$$R = \frac{\sum_{i=1}^{N}\left(\mathrm{GSR}_{p,i} - \overline{\mathrm{GSR}}_p\right)\cdot\left(\mathrm{GSR}_{o,i} - \overline{\mathrm{GSR}}_o\right)}{\sqrt{\sum_{i=1}^{N}\left(\mathrm{GSR}_{p,i} - \overline{\mathrm{GSR}}_p\right)^2 \sum_{i}^{N}\left(\mathrm{GSR}_{o,i} - \overline{\mathrm{GSR}}_o\right)^2}} \tag{21}$$

$$\mathrm{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{GSR_{o,i}} - \mathrm{GSR_{p,i}}\right)^2\right)^{0.5} \tag{22}$$

$$\mathrm{MAPE} = \left(\frac{100}{N}\right)\sum_{i=1}^{N}\left|\frac{\mathrm{GSR_{o,i}} - \mathrm{GSR_{p,i}}}{\mathrm{GSR_{o,i}}}\right| \tag{23}$$

$$\mathrm{NS} = 1 - \frac{\sum_{i=1}^{N}\left(\mathrm{GSR_{o,i}} - \mathrm{GSR_{p,i}}\right)^2}{\sum_{i=1}^{N}\left(\mathrm{GSR_{o,i}} - \overline{\mathrm{GSR}}_\mathrm{o}\right)^2} \tag{24}$$

$$\mathrm{cRMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}\left[\left(\mathrm{GSR_{o,i}} - \overline{\mathrm{GSR}}_\mathrm{o}\right) - \left(\mathrm{GSR}_{p,i} - \overline{\mathrm{GSR}}_\mathrm{p}\right)\right]\right)^2, \tag{25}$$

$$S_{\mathrm{do}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{GSR_{o,i}} - \overline{\mathrm{GSR}}_\mathrm{o}\right)^2, S_{\mathrm{dp}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{GSR}_{p,i} - \overline{\mathrm{GSR}}_\mathrm{p}\right)^2 \tag{26}$$

where $\mathrm{GSR_{o,i}}$ is the ith observed daily global solar radiation, $\mathrm{GSR_{p,i}}$ is the ith predicted daily global solar radiation, $\overline{\mathrm{GSR}}_\mathrm{o}$ is mean value of all the observed daily global solar radiation, $\overline{\mathrm{GSR}}_\mathrm{p}$ is mean value of the predicted daily global solar radiation, and N is the number of datasets.

References

1. WMO, *Guide to Meteorological Instruments and Methods of observation*. (World Meteorological Organization, 2014)
2. I. Moradi, Quality control of global solar radiation using sunshine duration hours. Energy **34**(1), 1–6 (2009)
3. R. Meenal, A.I. Selvakumar, Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. Renew. Energy **121**, 324–343 (2018)
4. N.A. Premalatha, Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms. J. Appl. Res. Technol. **14**(3), 206–214 (2016)
5. A. Rahimikhoob, Estimating global solar radiation using artificial neural network and air temperature data in a semi-arid environment. Renew. Energy **35**(9), 2131–2135 (2010)
6. S. Samadianfard et al., Daily global solar radiation modeling using data-driven techniques and empirical equations in a semi-arid climate. Eng. Appl. Comput. Fluid Mech. **13**(1), 142–157 (2019)
7. O. Kisi, S. Heddam, Z.M. Yaseen, The implementation of univariable scheme-based air temperature for solar radiation prediction: New development of dynamic evolving neural-fuzzy inference system model. Appl. Energy **241**, 184–195 (2019)
8. K. Mohammadi et al., Identifying the most significant input parameters for predicting global solar radiation using an ANFIS selection procedure. Renew. Sustain. Energy Rev. **63**, 423–434 (2016)
9. I.A. Basheer, M. Hajmeer, Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods **43**(1), 3–31 (2000)
10. M.J. Orr, *Introduction to radial basis function networks*. Technical Report, Center for Cognitive Science, (University of Edinburgh, 1996)
11. J.R. Koza, J.R. Koza, *Genetic Programming: on the Programming of Computers by Means of Natural Selection*, vol. 1 (MIT Press, Cambridge, 1992)
12. M. Pandey et al., Multiple linear regression and genetic algorithm approaches to predict temporal scour depth near circular pier in non-cohesive sediment. ISH J. Hydra. Eng. **26**(1), 96–103 (2020)
13. G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
14. C. Cortes, V. Vapnik, Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)

15. L. Breiman, Random forests. Mach. Learn. **45**(1), 5–32 (2001)
16. T. Chen et al., Xgboost: extreme gradient boosting. R Package Version 0.4–2 **1**(4), 1–4 (2015)
17. J.H. Friedman, Multivariate adaptive regression splines. The Annals of Statistics 1–67 (1991)
18. M. Pal, S. Deswal, M5 model tree based modelling of reference evapotranspiration. Hydrol. Process.: Int. J. **23**(10), 1437–1443 (2009)
19. R.E. Wright, Logistic regression. (1995)
20. J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
21. M. Behrang et al., The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Sol. Energy **84**(8), 1468–1480 (2010)
22. E.S. Mostafavi et al., A hybrid computational approach to estimate solar global radiation: an empirical evidence from Iran. Energy **49**, 204–210 (2013)
23. O. Kisi, Modeling solar radiation of Mediterranean region in Turkey by using fuzzy genetic approach. Energy **64**, 429–436 (2014)
24. H. Citakoglu, Comparison of artificial intelligence techniques via empirical equations for prediction of solar radiation. Comput. Electron. Agric. **118**, 28–37 (2015)
25. S. Shamshirband et al., Retracted article: Application of extreme learning machine for estimation of wind speed distribution. Clim. Dyn. **46**(5–6), 1893–1907 (2016)
26. S. Belaid, A. Mellit, Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate. Energy Convers. Manage. **118**, 105–118 (2016)
27. I.A. Ibrahim, T. Khatib, A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers. Manage. **138**, 413–425 (2017)
28. J. Fan et al., Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy Convers. Manage. **164**, 102–111 (2018)
29. D.H. Li et al., Estimation of hourly global solar radiation using Multivariate adaptive regression spline (MARS)–A case study of Hong Kong. Energy **186**, 115857 (2019)
30. A.E. Gürel, Ü. Ağbulut, Y. Biçen, Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation. J. Cleaner Prod. **277**, 122353 (2020)
31. J. Fan et al., Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. Renew. Energy **145**, 2034–2045 (2020)
32. M. Alizamir et al., A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions. Energy **197**, 117239 (2020)
33. H.O. Menges, C. Ertekin, M.H. Sonmete, Evaluation of global solar radiation models for Konya Turkey. Energy Convers. Manage. **47**(18–19), 3149–3173 (2006)
34. X. Liu et al., Evaluation of temperature-based global solar radiation models in China. Agric. For. Meteorol. **149**(9), 1433–1446 (2009)
35. IRIMO. Available from: http://www.irimo.ir/.
36. J. Prescott, Evaporation from a water surface in relation to solar radiation. Trans. Roy. Soc. S. Aust. **46**, 114–118 (1940)
37. R. Swartman, O. Ogunlade, Solar radiation estimates from common parameters. Sol. Energy **11**(3–4), 170–172 (1967)
38. G. Hargreaves, Responding to tropical climates. In: The 1980–81 Food and Climate Review. The Food and Climate Forum, Aspen Institute for Humanistic Studies. Boulder, USA., **81**, 29–32 (1981)
39. K.L. Bristow, G.S. Campbell, On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric. For. Meteorol. **31**(2), 159–166 (1984)
40. R.D. Jong, D. Stewart, Estimating global solar radiation from common meteorological observations in western Canada. Can. J. Plant Sci. **73**(2), 509–518 (1993)
41. R.G. Allen, Self-calibrating method for estimating solar radiation from air temperature. J. Hydrol. Eng. **2**(2), 56–67 (1997)
42. M. Donatelli, and G. Campbell, A simple model to estimate global solar radiation, in *Proceedings of the 5th European Society of Agronomy Congress (Zima M, Bartosova M eds), Nitra, Slovak*. (1998)
43. L. Hunt, L. Kuchar, C. Swanton, Estimation of solar radiation for use in crop modelling. Agric. For. Meteorol. **91**(3–4), 293–300 (1998)
44. D.G. Goodin et al., Estimating solar irradiance for crop modeling using daily air temperature data. Agron. J. **91**(5), 845–851 (1999)
45. N.A. Elagib, M.G. Mansell, New approaches for estimating global solar radiation across Sudan. Energy Convers. Manage. **41**(5), 419–434 (2000)
46. R. Chen et al., Validation of five global radiation models with measured daily data in China. Energy Convers. Manage. **45**(11–12), 1759–1769 (2004)
47. D. George, *SPSS for Windows Step by Step: A Simple Study Guide and Reference, 17.0 update, 10/e*. Pearson Education India (2011)
48. G. Chen et al., The genetic algorithm based back propagation neural network for MMP prediction in CO2-EOR process. Fuel **126**, 202–212 (2014)
49. R.C. Deo, X. Wen, F. Qi, A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl. Energy **168**, 568–593 (2016)
50. S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
51. M. Ji, Z. Jin, H. Tang, An improved simulated annealing for solving the linear constrained optimization problems. Appl. Math. Comput. **183**(1), 251–259 (2006)
52. H. Han et al., Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features. Appl. Therm. Eng. **154**, 540–547 (2019)
53. B. Zhu et al., Achieving the carbon intensity target of China: A least squares support vector machine with mixture kernel function approach. Appl. Energy **233**, 196–207 (2019)
54. Naseri, A., Jamei, M., Ahmadianfar, I., Behbahani, M. Nanofluids thermal conductivity prediction applying a novel hybrid data-driven model validated using Monte Carlo-based sensitivity analysis. Eng. Comput. 1–25 (2020)
55. J. Leathwick et al., Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshw. Biol. **50**(12), 2034–2052 (2005)
56. Friedman, J.H. and C.B. Roosen, *An introduction to multivariate adaptive regression splines*. Sage Publications Sage CA: Thousand Oaks, CA (1995)
57. G. Zheng et al., Multivariate adaptive regression splines model for prediction of the liquefaction-induced settlement of shallow foundations. Soil Dyn. Earthq. Eng. **132**, 106097 (2020)
58. W. Zhang, A.T. Goh, Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci. Front. **7**(1), 45–52 (2016)
59. A. Mohanta, K. Patra, MARS for Prediction of Shear Force and Discharge in Two-Stage Meandering Channel. J. Irrig. Drain. Eng. **145**(8), 04019016 (2019)
60. L. Wang, et al., *Probabilistic stability analysis of earth dam slope under transient seepage using multivariate adaptive regression splines.* Bulletin of Engineering Geology and the Environment, p. 1–13 (2020)

61. D.F. Specht, A general regression neural network. IEEE Trans. Neural Networks **2**(6), 568–576 (1991)
62. X. Yu, Prediction of chemical toxicity to Tetrahymena pyriformis with four-descriptor models. Ecotoxicol. Environ. Saf. **190**, 110146 (2020)
63. P. Ramsami, V. Oree, A hybrid method for forecasting the energy output of photovoltaic systems. Energy Convers. Manage. **95**, 406–413 (2015)
64. C. Coulton, J. Chow, Interaction effects in multiple regression. J. Soc. Serv. Res. **16**(1–2), 179–199 (1993)
65. J. Jaccard, R. Turrisi, J. Jaccard, *Interaction Effects in Multiple Regression* (Sage, CA, 2003)
66. K.E. Taylor, Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmos. **106**(D7), 7183–7192 (2001)
67. J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. **10**(3), 282–290 (1970)
68. M. Jamei, I. Ahmadianfar, X. Chu, Z.M. Yaseen, *Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models* (Flow Meas, Instrum, 2020), p. 101878
69. M. Jamei, I.A. Olumegbon, M. Karbasi, I. Ahmadianfar, A. Asadi, M. Mosharaf-Dehkordi, On the thermal conductivity assessment of oil-based hybrid nanofluids using extended kalman filter integrated with feed-forward neural network. Int. J. Heat Mass Transf. **172**, 121159 (2021)
70. C.D. Lewis, Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting. Butterworth-Heinemann (1982)