Neural Computing and Applications (2023) 35:887–906
https://doi.org/10.1007/s00521-022-07841-x

ORIGINAL ARTICLE

Forecasting of solar radiation using different machine learning approaches

Vahdettin Demir[1] · Hatice Citakoglu[2]

Received: 24 February 2022 / Accepted: 13 September 2022 / Published online: 23 September 2022

Abstract

In this study, monthly solar radiation (SR) estimation was performed using five different machine learning-based approaches. The models used are support vector machine regression (SVMR), long short-term memory (LSTM), Gaussian process regression (GPR), extreme learning machines (ELM) and K-nearest neighbors (KNN). Modeling of these approaches was carried out in two stages. In the first stage, VIF analysis was carried out to develop the model. Thus, the input parameters that decrease the performance of the model are removed. In the second stage, remaining input parameters such as meteorological data, station location data and spatial and temporal information were used in the forecasting modeling according to the correlation SR. In this study, the data set is divided into two parts as test and training. 30% was used in the testing phase, and 70% of the data was used in the training phase. When comparing models, the following error statistics were used: Nash–Sutcliffe efficiency coefficient (NSE), mean absolute error (MAE), mean absolute relative error (MARE), root-mean-square error (RMSE) and coefficient of determination (R^2). In addition, Taylor diagrams, violin plots, box error, spider plot and Kruskal–Wallis (KW) and ANOVA test were utilized to determine robustness of model's forecast. As a result of the study, the KW test and ANOVA test results showed that the data of many models were from the same population with observations, and it has proved that LSTM and GPR algorithms are applicable, valid and an alternative for SR forecasting in Turkey, which has arid and semi-arid climatic regions.

Keywords Solar radiation · Long short-term memory · Gaussian process regression · Support vector machine regression · Extreme learning machines and K-nearest neighbors · Turkey

1 Introduction

Solar radiation (SR) is the energy emitted by the sun [1]. The energy balances of several physical, chemical and biological processes are influenced by solar radiation reaching the Earth's surface [2, 3, 4]. Changes in solar radiation have a significant impact on heat fluxes, the hydrological cycle, terrestrial biological ecosystems and climate [5, 6]. In addition, solar energy emits significantly less pollution than traditional sources such as fossil fuels, and it is the most abundant of all renewable and sustainable energy resources at locations all over the world and can be used for commercial purposes through large solar power plants [7, 8, 9]. Thus, precise measurement and comprehension of solar radiation's spatial–temporal variability are critical for meteorological and hydrological processes as well as energy development and usage [10, 11].

Meteorology, hydrology and agricultural activities are used in several research to forecast SR [12] . For example, Ododo et al. [13] suggested temperature as a solar radiation metric. SR has substantial relationships with air temperatures, according to Bandyopadhyay et al. [14]. In order to forecast solar radiation, Ododo [15] used relative humidity and maximum temperature. Average air temperature measurements were utilized as input data by Rehman and Mohandes [16] to forecast solar radiation. Kisi et al. [17] suggested many meteorological parameters for the SR

✉ Vahdettin Demir
vahdettin.demir@karatay.edu.tr

Hatice Citakoglu
hcitakoglu@erciyes.edu.tr

[1] Civil Engineering Department, Faculty of Engineering and Natural Sciences, KTO Karatay University, Konya, Turkey

[2] Civil Engineering Department, Faculty of Engineering, Erciyes University, Kayseri, Turkey

forecast. In addition, current studies in the literature have revealed that station location information is used in the forecast of global solar radiation [18]. for example, Kumar et al. [19] reviewed different models for SR forecast with latitude, longitude and altitude data. Chabane et al. [20] estimated SR as a function of latitude and longitude coordinates.

Over 400 articles were found in Scopus' reported database for machine learning (ML) approach for SR forecasting. The VOSviewer technique was used to generate a list of important keywords for this research domain (Fig. 1a). Furthermore, when the adopted research is examined across time (Fig. 1a), it is clear that many studies were published in 2018 and beyond. These studies appear to be more interested in climate change, deep learning, new machine learning models such as SVM, ELM, climate change and the development of renewable energy generation. Figure 1b shows the main regions where solar radiation estimates have been investigated. It is the region of China with the most research (76), followed by the USA (63), India (51), Spain (25), Iran (22), France (21) and Turkey (18).

Some researchers have investigated SR modeling using different mathematical equations and ML approaches; for example, Kumar et al. [19] compared the regression model with the ANN models for SR prediction. Kisi et al. [17] employed wavelet transform approach with ANN ELM, radial basis function (RBF) and their hybrid variants. Rahimikhoob et al. [21] compared the ANN's and statistical methodologies for deriving SR from satellite images. Polo et al. [22] investigated the sensitivity of satellite-based approaches for calculating SR to various aerosol input and model choices. Ahmad and Tiwari [23] investigated various SR models and discovered that the Collares-Pereira and Rabl model, as modified by Gueymard, had the best accuracy for projecting mean hourly SR, and that the Ertekin and Yaldiz model performed best against measured data from Konya, Turkey. Sonmete et al. [24] compared 147 SR models available in the literature for monthly solar radiation estimation in Ankara (Turkey). Citakoglu [25] also compared the ANFIS, ANN and MLR models, and different empirical equations; the end results showed that when it came to estimating monthly SR in Turkey, the ANN model outperformed the ANFIS, MLR and empirical equations. Wang et al. [11] compared three different ANN methods (GRNN, RBNN and MLP models) for predicting the daily SR using meteorological variables such as air temperature, relative humidity and sunshine duration. To our knowledge, no research has been conducted to evaluate the performance of machine learning approach on SR prediction by examining optimum conditions such as optimization or training algorithms and reducing input parameters.

Solar radiation is studied widely around the world, particularly in solar-rich locations like the Mediterranean and the Middle East [26, 27]. Unfortunately, most sites lack access to and measurement of observed sun radiation values. The costs of obtaining, installing and maintaining devices, as well as issues calibrating radiation detecting equipment, are the main causes for the absence of trustworthy radiation data [17]. As a result, location-based models, temperature-based models, remote sensing-based approaches, temperature-based models, day and month-number-based models, cloudiness-based models, sunshine-based models and hybrid models are all commonly employed to estimate solar radiation [11, 25, 28–38]. However, due to intricate connections between independent and dependent variables, these models cannot always provide trustworthy estimates, particularly in humid places where solar radiation is heavily influenced by clouds [11].

The aim of this study is to investigate forecasting of SR with five different ML approaches, including long short-term memory (LSTM), support vector machine regression (SVMR), Gaussian process regression (GPR), extreme learning machines (ELM) and K-nearest neighbors (KNN). Geographical positions (latitude, longitude and elevation), the time information of the station measurements (months and years) and monthly observed meteorological measurements (temperature, evaporation, wind speed and relative humidity) of 163 meteorological stations of Turkey were used to estimate SR. This research will make a substantial contribution to the existing literature in the following ways:

(i) The majority of Turkish meteorological stations are used in the SR forecasting process. In addition, the data has a continuous and long-term recording period.
(ii) Five different models were utilized for SR forecasting, and the methods were compared. In the model comparisons, the sub-parameters and the number of inputs were differentiated, and the best result was determined for each model parameter and the number of inputs.
(iii) Variance inflation factor (VIF) analysis was performed to enhancing the SR forecasting accuracy, and the input parameters that reduced the accuracy of the model were excluded from the study.
(iv) Finally, the Kruskal–Wallis test and ANOVA test were used to detect whether data estimated and measured were from the same distribution.

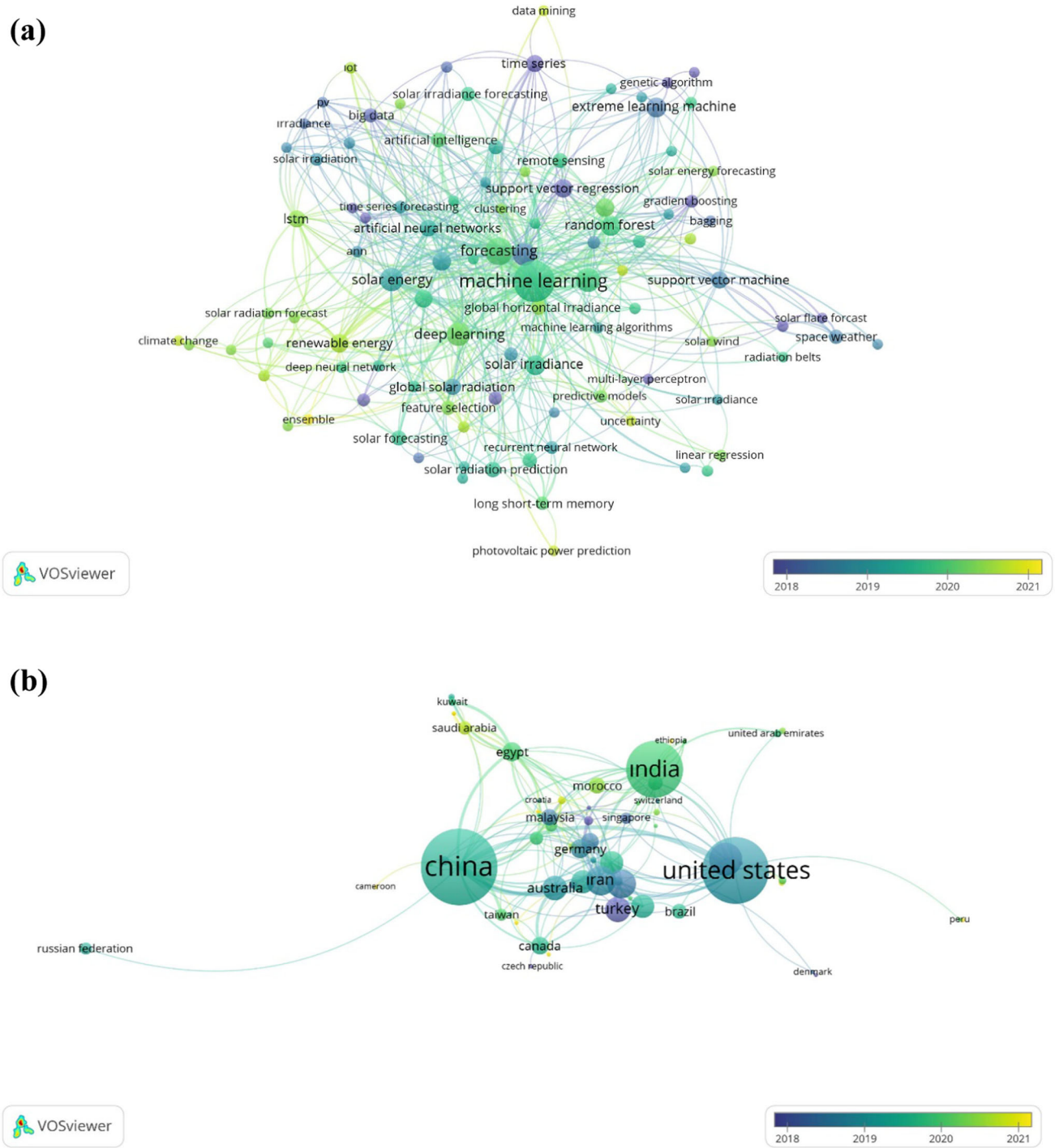

Fig. 1 Literature review keywords **a** for the SR forecasting using ML approach in research regions **b**

The next section of the paper presents the "Materials and Method." In this section, given study area description and data set, and theoretical framework of ML approaches, followed by this is performance metrics and application of variable selection. Then Sect. 5 presents the results and discussion, and Sect. 6 presents concluding remarks.

2 Materials and method

2.1 Study area description and data set

Turkey is bordered by the sea on three sides (north, south and west). Turkey is geographically located between the 36°–42° N and 26°–45° E meridians. The country has a

rectangular shape with a width of 1660 km. The actual coverage area of the country, including the lakes and islands, is 814,578 km^2, whereas the anticipated coverage area is 783,562 km^2. The vast gap between these two places is due to the country's steep and rugged terrain. The country's highest and most mountainous regions are largely found in the east. The interior of the country is primarily flat. Dry climate is seen in the interior of Turkey. Summers in Turkey are hot and dry, and winters are dry and chilly, especially in areas distant from the sea. Continental climate is seen in the mountainous parts of the Eastern Region, Southeast Region and Inner Region of Turkey. In this climate type, the annual temperature difference is huge and the winters are cold [39].

The data used in this study were obtained from the General Directorate of Meteorology (MGM). In total, 163 meteorological stations were used. Monthly solar radiation (SR, MJ/m^2), max. temperature (Tmax, °C), avg. temperature (Tavg °C), min. temperature (Tmin, °C), avg. wind speed (WSavg, m/s), elevation (m), year, longitude (°), month, latitude (°), max. relative humidity (RHmax, %) and min. relative humidity (RHmin, %) data were supplied from MGM. The data covers the years 1967–2020, the stations in the region are located between 2 and 1777 m, the highest temperatures are 46.40 °C, the relative humidity reaches a maximum of 110%, and the maximum solar radiation is read as 31.54 MJ/m^2 at the regional stations. It is understood that under the title of months, there is a continuous and 12-month periodic component. In the modeling phase, these parameters were introduced as input data, respectively. While deciding the order of the input data in the models, the correlation coefficient from strong to weak between SR and parameters was considered. The parameters used in the SR estimation are $T_{\max}$, T_{avg}, $T_{\min}$, W_{Savg}, elevation, year, longitude, month, latitude, $\text{RH}_{\max}$ and $\text{RH}_{\min}$. The correlation between these data and SR is given in Fig. 2.

In Fig. 2, a strong correlation is observed between T and SR, while negative correlations are observed between relative RH and SR. The data was randomly divided into two parts: training and test data, before the modeling of the study. Of the 163 stations' data, 70% was used in the training phase and the remaining 30% was used in the testing phase to compare the performance of the models while training data is used to construct the model. The training and test rates used are frequently used and recommended in the literature [11, 40, 41, 42]. Figure 3 shows the stations that were utilized.

Although the aim of the study is to distribute the stations regionally homogeneously, it is seen that training stations are not found in some regions, especially in the northern regions, in random selection, while in some regions, especially in the inner parts, there are no training stations. This is the result of completely random selection. The distance and independence of the training and test stations, as shown in Fig. 3, show that a solution is being sought for a difficult problem. Statistical information about the training and test stations is given in Table 1.

While the stations were separated during the training and testing phases, the station balancing was not performed after the data were separated according to the training or test rate. For this reason, the data of some stations in the entire recording period were used in the training phase; for example, while the previous years were used in the training phase in some stations, the data of some years were transferred to the testing phase. For this reason, Table 1 shows that the data in the year title are at least 1967 and at most 2020. In Table 1, the data are distributed homogeneously. For example, maximum temperatures are around 45–46 °C and SR values are about 31 MJ/m^2.

2.2 Long short-term memory (LSTM)

LSTM was first presented by Hochreiter and Schmidhuber [43] based on recurrent neural networks (RNN). It was created to solve vanishing and exploding gradient difficulties. Using its unique structure, gates and cell state, it can also maintain dependencies over lengthy periods of time. To ensure the integrity of this work, a brief review of the LSTM unit is offered below. The fundamentals of RNN and LSTM were extensively defined in [44]. LSTM is a superior evolution of recurrent neural networks (RNNs) that tackle the drawbacks of RNNs. In addition, LSTM technology is unusual in that it stores information for a lengthy period of time. Furthermore, the LSTM is made up of four layers that are linked together through various communication protocols. The fact that its entire network is built up of memory blocks is the next feature. These blocks are also known as cells. Information is stored in one cell and then sent to the next using gate controls. With the help of these gates, it becomes much easier to precisely examine data [45, 46]. Figure 4 shows the construction of the LSTM. LSTM equations are listed below:

$$i_{\text{t}} = \sigma(W_{\text{i}}x_{\text{i}} + U_{\text{i}}h_{t-1} + b_{\text{i}}) \tag{1}$$

$$f_t = \sigma(W_{\text{f}}x_t + U_{\text{f}}h_{t-1} + b_{\text{f}}) \tag{2}$$

$$o_t = \sigma(W_{\text{o}}x_t + U_{\text{o}}h_{t-1} + b_0) \tag{3}$$

$$\tilde{C}_t = \tanh(W_{\text{c}}x_t + U_{\text{c}}h_{t-1} + b_{\text{c}}) \tag{4}$$

$$C_t = f_t \otimes C_{t-1} + i_{t-1} \otimes \tilde{C}_t \tag{5}$$

$$h_{\text{t}} = o_{\text{t}} \otimes \tanh(C_{t-1}) \tag{6}$$

In the equations, i_{t}, f_{t} and o_{t} are the entrance, forgetting and exit, respectively; W_{i}, W_{f} and W_{o} show the weights

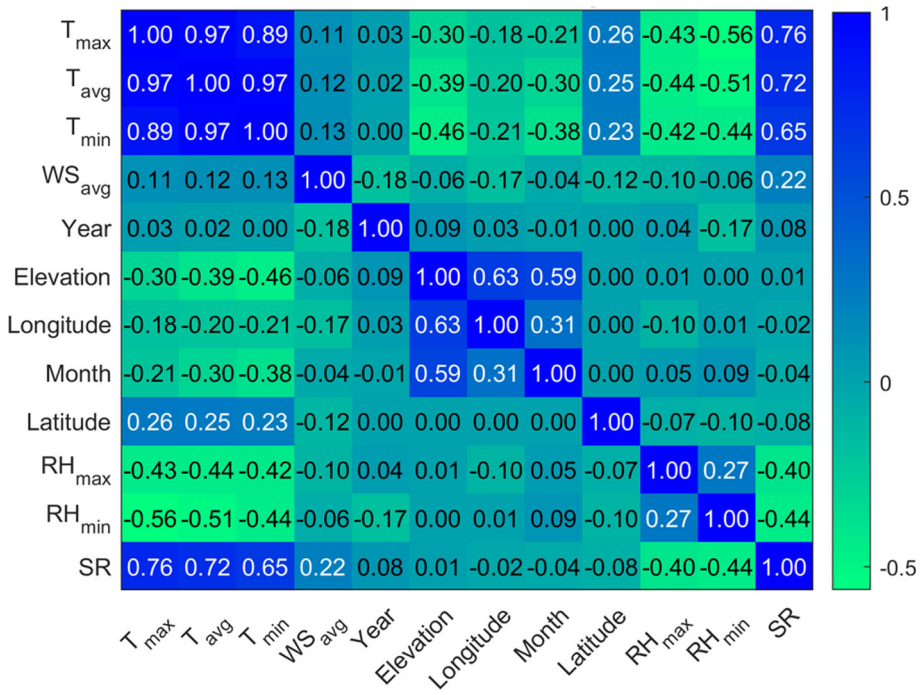

Fig. 2 Correlation matrix between SR and each input variables

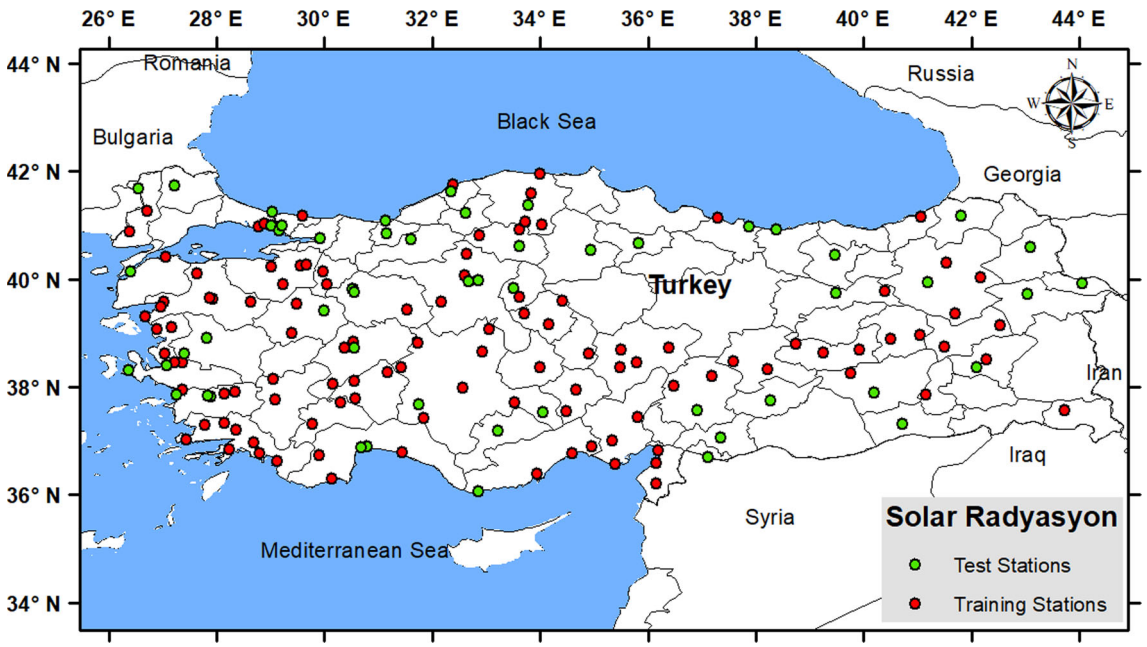

Fig. 3 Spatial distribution of training and testing stations

connecting the input, forget and output gates to the input, respectively; U_i, U_f and U_o represent the weights from the entry, forget and exit gates to the hidden layer in order; b_i, b_f and b_o indicate the input, forget and output gate bias vectors, respectively; $\widetilde{C}_t$ is the previous state of the cell; C_t is the current state of the cell; h_{t-1} refers to the cell's output at the previous time point; and h_t stands for the output of the cell [47]. In this study, SR was estimated using meteorological, spatial and temporal input parameters. The application of the LSTM model was carried out with the help of codes written in MATLAB. In this study, adaptive moment estimation (Adam), stochastic gradient descent with momentum (SGDM) and root-mean-square propagation (RMSProp) optimization algorithms were used for the training of the model and the forecasting performances were compared. For details of optimization algorithms, Pandey & Srivastava [48] can be examined.

Table 1 Information about the test and training data

	Parameters	Min	Mean	Max	Std	C_S	C_k
Training	T_{max} (°C)	− 5.80	25.85	45.50	9.02	− 0.37	− 0.57
	T_{avg} (°C)	− 19.50	13.62	33.30	9.18	− 0.47	− 0.06
	T_{min} (°C)	− 39.80	1.39	24.00	9.92	− 0.57	0.55
	WS_{avg} (m/s)	0.15	1.53	5.09	0.63	0.85	0.81
	Elevation (m)	2.00	669.10	1777.0	533.60	0.38	-0.65
	Year (1967–2020)	1967	1993	2020	12.60	− 0.15	− 0.60
	Longitude (°)	26.37	34.36	44.05	4.93	0.28	− 0.95
	Month (1–12)	1	6.50	12	3.45	0.00	− 1.22
	Latitude (°)	36.07	39.35	41.74	1.62	− 0.36	− 1.20
	RH_{max} (%)	37.00	94.23	104.00	5.83	− 3.26	15.26
	RH_{min} (%)	0	26.51	76.00	12.48	0.50	-0.07
	SR (MJ/m^2)	0.79	14.31	31.81	6.50	0.13	− 1.10
Testing	T_{max} (°C)	− 1.00	26.07	46.40	8.94	− 0.34	− 0.75
	T_{avg} (°C)	− 16.25	13.89	33.70	9.07	− 0.35	− 0.50
	T_{min} (°C)	− 37.90	1.71	27.00	9.79	− 0.37	− 0.02
	WS_{avg} (m/s)	0.00	1.59	5.16	0.64	0.84	1.16
	Elevation (m)	1967	1995	2020	11.31	− 0.20	− 0.12
	Year (1967–2020)	2.00	689.92	1890.00	542.85	0.00	− 1.33
	Longitude (°)	26.39	33.28	44.05	4.81	0.47	− 0.81
	Month (1–12)	36.07	38.68	41.96	1.33	0.26	− -0.49
	Latitude (°)	1.00	6.50	12.00	3.45	0.00	− 1.22
	RH_{max} (%)	37.00	93.72	110.00	6.84	− 3.18	13.38
	RH_{min} (%)	0	25.41	79.00	11.98	0.74	0.47
	SR (MJ/m^2)	0.03	14.86	31.54	6.48	0.11	− 1.10

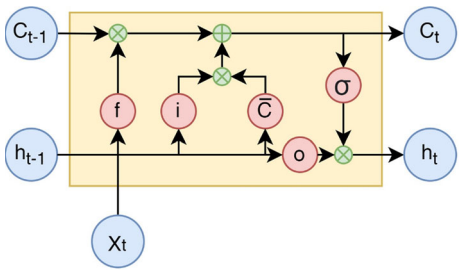

Fig. 4 LSTM structure

The equations for Adam are as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{7}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{8}$$

$$m_t^{'} = \frac{m_t}{1-\beta_1^t}, \quad v_t^{'} = \frac{v_t}{1-\beta_2^t} \tag{9}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t^{'} + \in}} m_t^{'} \tag{10}$$

The equations for RMSPRop are as follows:

$$E\left[g^2\right]_t = 0.9E\left[g^2\right]_{t-1} + 0.1 g_t^2 \tag{11}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \in}} g_t \tag{12}$$

$$g_t = \nabla_{\theta t} J(\theta_t) \tag{13}$$

The equations for SGDM are as follows:

$$v_{t+1} = \gamma v_t + \eta \nabla_\theta J \tag{14}$$

$$\theta_{t+1} = \theta_t - v_{t+1} \tag{15}$$

$\theta \in \mathbb{R}^d$: model parameters; η: learning coefficient; $\nabla_\theta J(\theta_t; x^{(i)}; y^{(i)})$: the slope of the target function depending on the parameters; $G_{t,ii:}$ each diagonal element is the sum of the squares of the slope values calculated up to t. iterations, according to parameter θ_i; and ε: the constant value assigned to prevent the learning coefficient from dividing by 0. [49].

2.3 Support vector machine regression (SVMR) model

Support vector machine (SVM) was first proposed by Vapnik [50] in 1995. The concept of SVM is based on statistical learning theory and the principle of structural risk minimization [50]. Smola [51], devised a form of regression model called support vector machine regression

(SVMR). SVMR models were created by merging regression functions with SVM to handle forecasting, prediction and regression problems [52, 53]. The SVMR model's main goal is to discover a function with the least "ε" deviation and that is as linear as possible for all training data points and target vectors [51]. The SVMR model's structural configuration is shown in Fig. 5. The SVMR regression function's summary is as follows [54]:

$$f(x) = w \times \phi(x) + b \tag{16}$$

In the equation, w is the weight vector, b is the deviation, and ϕ is the transfer function.

Optimal conditions are obtained with the Lagrangian multipliers and kernel function in SVMR. Linear, polynomial, radial basis function (RBF) and sigmoid functions are examples of kernel functions [39, 55, 56]. Application of the SVMR model was carried out with the help of codes written in MATLAB. The linear, polynomial, radial basis function were used for the training of the model and the forecasting performances were compared. The following technical report contains more information on SVM and SVMR approaches: Classification and regression with support vector machines [57].

2.4 Gaussian process regression (GPR) model

GPR is a probabilistic nonparametric approach. Both estimations and confidence intervals are calculated with GPR a probabilistic nonparametric model. GPR is a significant extension of the Gaussian probability distribution. The probability of a Gaussian distribution is calculated using the input vectors. Each input data vector's probability is determined. As a result, the GPR model computes a mean and variance–covariance vector [58, 59]. The SVMR regression function is:

$$f \approx GPR(m(x), k(x, x')) \tag{17}$$

where x is the vector of input variables; $m(x)$ is the average function of input variables; and $k(x, x')$ is the variance–covariance matrix. The shape of a multi-variate Gauss distribution is defined by the variance–covariance matrix.

Kernel (with ardmatern32, ardmatern52, squaredexponential, ardsquaredexponential, matern32, matern52 covariance function) and basis (with constant, none, linear, pureQuadratic, squaredexponential covariance function) were used in this study because they performed better in forecasting studies than the others. In the SR estimation, the function that gives the least error according to the error criteria in the next section is used. For details of covariance function, Rasmussen and Williams (2006) [58] and Neal [60] can be examined.

2.5 Extreme learning machines (ELM)

Extreme learning machine (ELM) was first presented by Huang et al. in 2006. ELM is a single hidden layer feed-forward neural network training algorithm that converges significantly faster than traditional ANN methods and produces promising results [61, 62]. This is because the input weights are created at random, resulting in a unique least-squares solution for the output weights, which is solved by the Moore–Penrose function [63]. Because the

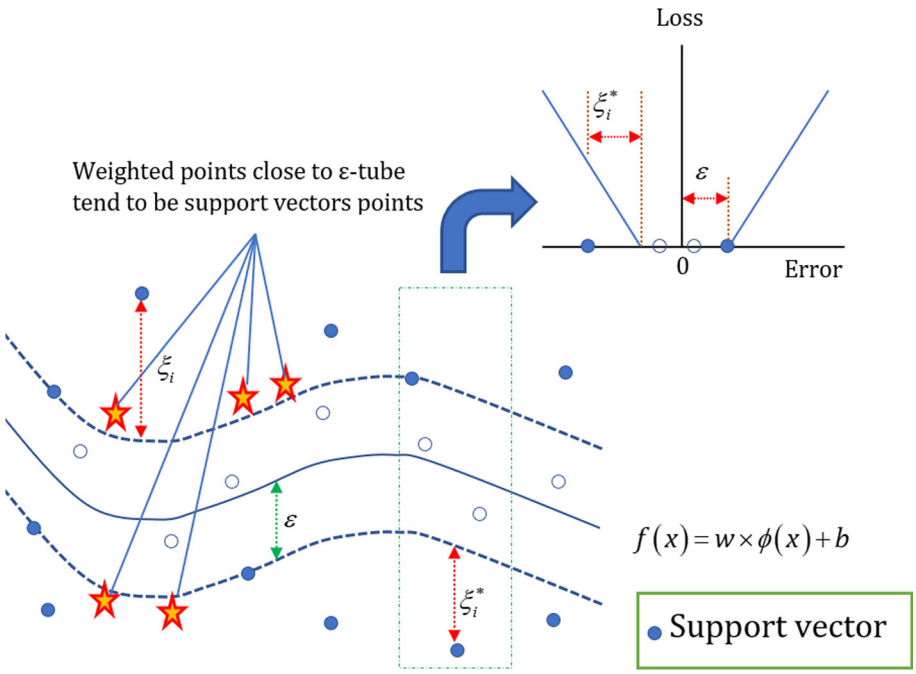

Fig. 5 Nonlinear support vector regression configuration

randomly initiated hidden neurons in ELM's underlying theory are fixed, ELM is extraordinarily efficient at achieving a global optimum solution using universal approximation capabilities. Slow convergence, poor generalization, local minima difficulties, overfitting and the necessity for iterative tweaking are key drawbacks of the ANN model, all of which point to ELM's superiority over ANN [63, 64, 65]. The ELM model's general structural configuration is given Fig. 6.

The SVMR regression function's summary is as follows:

$$\sum_{i=i}^{L} B_i g_i(\alpha_i x_t + \beta_i) = z_t \tag{18}$$

In Eq. 18, L is the hidden nodes number, $g_i\ (\alpha_i\ x_t + \beta_i)$ is the hidden layer output function, α_i and β_i is hidden node parameters, B_i is the weight factor connecting the ith hidden nodes and output node and z_t is ELM model output.

The application of the ELM model was carried out with the help of codes written in MATLAB. The number of input neurons was tried from 1 to 300, and training ratio was chosen 0.7 in this study. The input parameters in Fig. 6 are defined separately to the ELM model according to the correlation order expressed under the data set title (see Fig. 2).

2.6 K-nearest neighbors (KNN)

The KNN is a nonparametric classification method invented by Evelyn Fix and Joseph Hodges in 1951 [66] and expanded by Altman [67]. Data categorization and regression are both done with KNN. In both circumstances, the input is a data set with the k closest training samples. The KNN approach searches through a database for data that is comparable to the observed data. These data are referred to as the present data's nearest neighbors [68]. In this paper, KNN is used to forecast mostly related testing stations with the training station. The KNN regression function's summary is as follows:

$$f_{\mathrm{KNN}}(x') = \frac{1}{K} \sum_{i \in N_K(x')} y_i \tag{19}$$

For an unknown pattern x', KNN regression computes the mean of the function values of its K-nearest neighbors with set $N_{\mathrm{K}(x')}$ containing the indices of the K-nearest neighbors of x'. The notion of localization of functions in data and label space underpins the idea of averaging in KNN. In local neighborhoods of x_i, patterns x' are expected to have similar continuous labels $f(xi)$ like y_i [69]. The application of the KNN model was carried out with the help of codes written in MATLAB. The kdtree and exhaustive nearest neighbor search method were used for the training of the model and the forecasting performances were compared. The study flowchart of this study is given in Fig. 7.

3 Performance metrics

The accuracy of the models proposed in this research was evaluated using widely known performance metrics [70]. MAE, MARE, RMSE, R^2 and NSE were used in model evaluations. Low MAE, MARE and RMSE values, as well as R^2 values near 1, suggest accurate and dependable estimations. NSE values range from $-\infty$ and 1 [71].

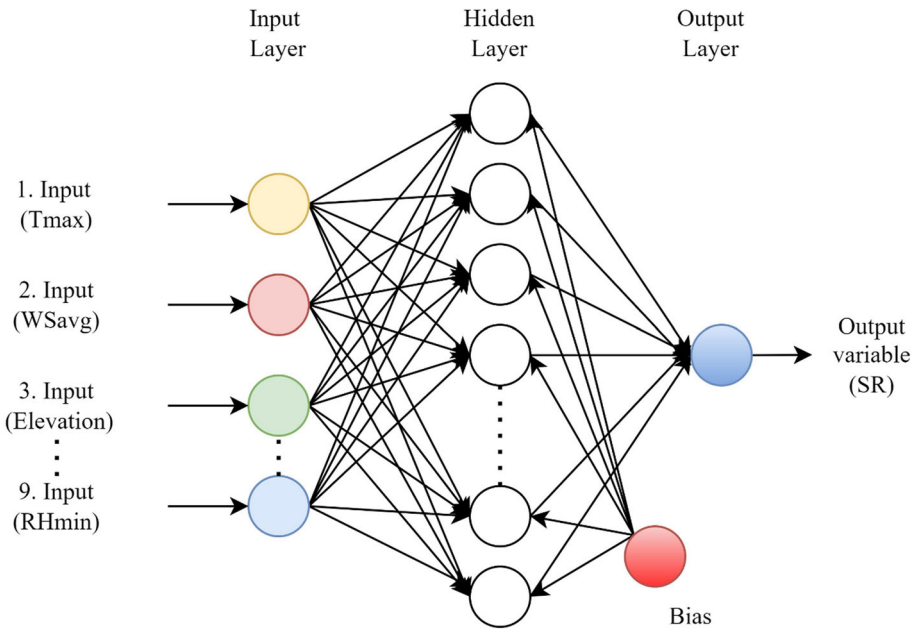

Fig. 6 ELM structure

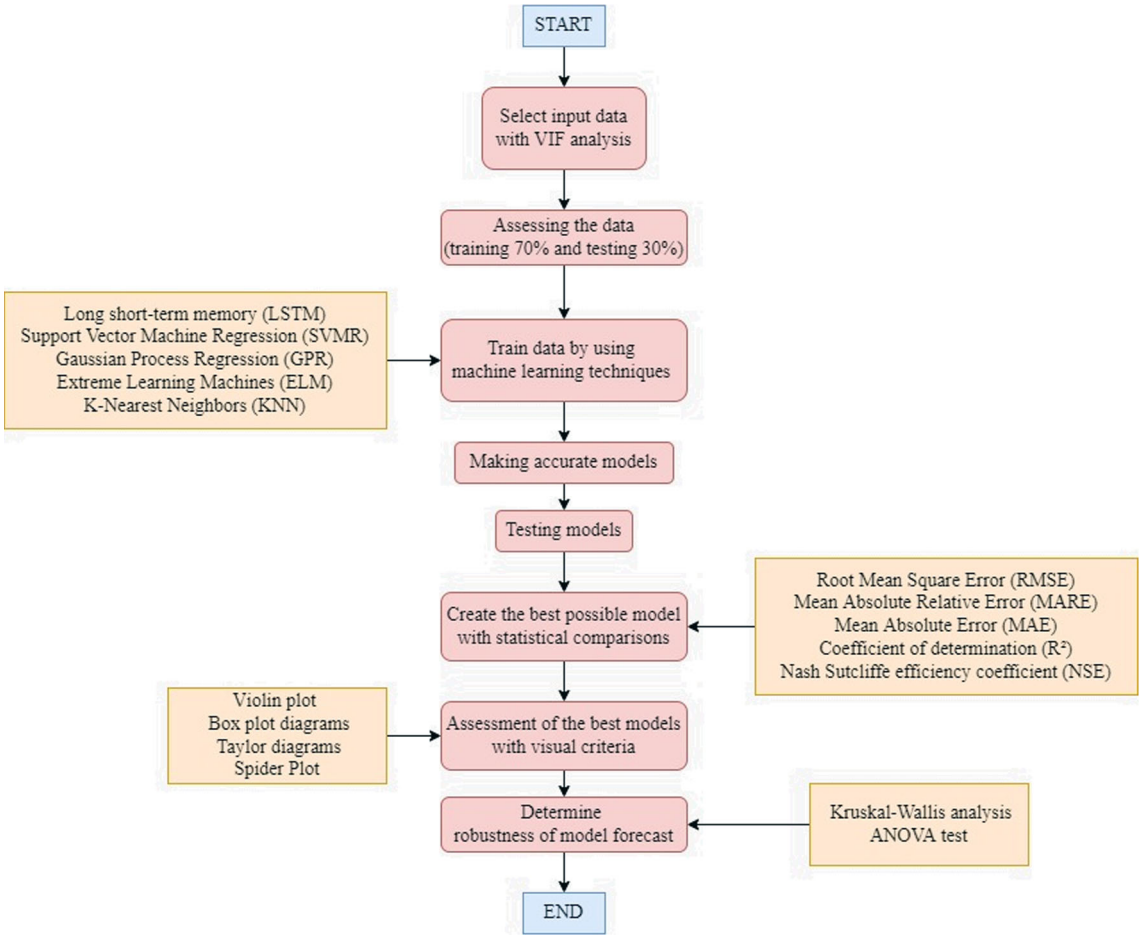

Fig. 7 Study flowchart

$$\mathrm{RMSE} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(\mathrm{SR}_{\mathrm{predicted}} - \mathrm{SR}_{\mathrm{measured}}\right)^2} \tag{20}$$

$$\mathrm{MARE} = 100\frac{1}{n}\sum_{i=1}^{n}\frac{\left|\mathrm{SR}_{\mathrm{predicted}} - \mathrm{SR}_{\mathrm{measured}}\right|}{\mathrm{SR}_{\mathrm{predicted}}} \tag{21}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{SR}_{\mathrm{predicted}} - \mathrm{SR}_{\mathrm{measured}}\right| \tag{22}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(\mathrm{SR}_{i\mathrm{measured}} - \overline{\mathrm{SR}_{\mathrm{imeasured}}}\right)^2.\left(\mathrm{SR}_{\mathrm{ipredicted}} - \overline{\mathrm{SR}_{\mathrm{predicted}}}\right)^2}{\sum_{i=1}^{n}\left(\mathrm{SR}_{\mathrm{imeasured}} - \overline{\mathrm{SR}_{\mathrm{imeasured}}}\right)^2.\sum_{i=1}^{n}\left(\mathrm{SR}_{\mathrm{ipredicted}} - \overline{\mathrm{SR}_{\mathrm{predicted}}}\right)^2} \tag{23}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(\mathrm{SR}_{\mathrm{predicted}} - \mathrm{SR}_{\mathrm{measured}}\right)^2}{\sum_{i=1}^{n}\left(\mathrm{SR}_{\mathrm{measured}} - \overline{\mathrm{SR}_{\mathrm{measured}}}\right)^2} \tag{24}$$

where SRmeasured is SR variables measured by MGM; $\mathrm{SR}_{\mathrm{predicted}}$ is SR variables predicted by approaches; $\overline{SR_{\mathrm{measured}}}$ is average of SR variables; and n is amount of data. In this study, Taylor diagram, violin and box error plot were used to compare LSTM, SVMR, GRP, ELM and KNN approaches. These diagrams graphically summarize how close the models are to the observations [72, 73, 74]. Comparisons in the Taylor diagram were made using model correlations and root-mean-square deviation (RMSD). On the other hand, many statistical parameters such as mean median standard deviation etc. are used in the violin diagram. In addition, for the final evaluation of the performance of the models, the spider graph of the methods of the input combinations that gave the best results was also given and more than one evaluation criteria were evaluated on a single figure [75].

4 Application of variable selection

In this study, model development was realized by reducing the input parameters. Models' variance inflation factors (VIFs) were calculated in three steps, and significant variables were selected from among many potential variables. Table 2 shows the computed VIFs for each phase. In Table 2, VIFs greater than 5.0 are written using bold definitions. In the first stage, the VIFs of the T_{avg} and T_{min} variables are all greater than 5.0, as shown in Table 2. The

Table 2 Summary of the results of the VIF analysis

	Term	Coef	T-Value	VIF
Step 1	Constant	− 45.05	− 13.8	
	Tmax (°C)	0.65152	237.62	2.12
	Tavg (°C)	**1.08002**	**112.09**	**27.95**
	Tmin (°C)	**− 0.41543**	**− 46.9**	**27.46**
	WSavg (m/s)	0.9467	32	1.13
	Elevation (m)	0.02596	16.65	1.1
	Year	0.004763	84.8	2.88
	Longitude (°)	− 0.08565	− 17.31	1.87
	Month	− 0.0971	− 4.94	1.66
	Latitude (°)	− 0.57637	− 107.85	1.12
	RHmax (%)	− 0.01723	− 6.16	1.33
	RHmin (%)	0.04232	21.5	1.62
	R^2 = 0.7562	Tcri = 1.96		
Step 2	Constant	0.09	0.67	
	Tavg (°C)	**1.1697**	**109.64**	**19.88**
	Tmin (°C)	**− 0.64791**	**− 65.49**	**19.88**
	R^2 = 0.5796	Tcri = 1.96		
Step 3	Constant	− 45.65	− 13.82	
	Tmax (°C)	0.65152	237.62	2.12
	WSavg (m/s)	1.0239	34.36	1.12
	Elevation (m)	0.02747	17.42	1.09
	Year	0.004237	79.12	2.56
	Longitude (°)	− 0.05686	− 11.61	1.79
	Month	− 22.84%	− 1182.00%	1.56
	Latitude (°)	− 0.56625	− 104.98	1.12
	RHmax (%)	− 0.02814	− 10.04	1.31
	RHmin (%)	0.04858	24.56	1.59
	R^2 = 0.7505	Tcri = 1.96		

value of 5.0 of VIF is the critical value, and the parameters exceeding this value represent the parameters that should be excluded from the modeling [25]. T_{avg} and T_{min} of t values are also less than t_{cri}. As a result, the variables T_{avg} and T_{min} are no longer included in the models. In the second step, a smaller number of variables are purposefully chosen, and after witnessing high VIF's for T_{avg} and T_{min} variables, T_{avg} and T_{min} are deleted for good, as T_{avg} and T_{min} are highly connected. In the last stage, the VIFs of all coefficients are less than 5.0, and the t values of all variables are greater than t_{cri}. T_{max} (°C), WS_{avg} (m/s), elevation (m), year, longitude (°), month, latitude (°), RH_{max} (%) and RH_{min} (%) are selected as valid input variables as a consequence of this study.

The performance of the LSTM, SVMR, GRP, ELM, KNN models was checked using the suggested 9-input parameters (for example T_{max} as 1st input, W_{Savg} as second input, …, RH_{min} as lastly input) during both training and testing phases. The results are given next section.

5 Results and discussion

The LSTM, SVMR, GRP, ELM and KNN techniques were utilized in this study to create models for forecasting SR in Turkey's using meteorological parameters, location and spatial and temporal information.

The LSTM model was used to estimate SR data in the first part of the research. Several trials were undertaken throughout the creation phase of LSTM models by adjusting the number of neurons in the hidden layer. In the LSTM models, tanh was used as "state activation function" and sigmoid was used as "gate activation function" [39]. Also, in LSTM models, Adam, SGDM and RMSProp optimization algorithms were used for network training. Trials were conducted with a single hidden layer, between 10 and 30 neurons, and 50 to 300 iterations in the LSTM model architecture. Initial learning rate coefficient was set to 0.05, learning rate reduction factor was set to 0.2, and learning rate reduction time was set at 125 for the other parameters of the LSTM model. The selection of LSTM model parameters was inspired by [39]. The best outcome for each output value is provided in Table 3 as a consequence of the trials undertaken during the LSTM modeling phase. The SVMR model was utilized to estimate SR in the current study's second phase. The "kernel function" was used to create estimating models in the SVMR technique. The common nonlinear radial basis function (RBF), linear and polynomial were utilized in this study because they performed better in estimate studies than the other kernel functions. The lowest values of alpha ($\alpha_i - \alpha^*_i$) and bias (b) parameters, representing the difference between two Lagrange multipliers, were obtained with sequential minimal optimization (SMO). The GPR model was utilized to estimate SR in the third phase of the current investigation. Estimation models were created using kernel and basis functions in the GPR technique. In this study, many kernel functions have also been tested. In order to obtain the best performance value, mater32, matern52, ardmatern32, ardmatern52, ardsquaredexponential and squaredexponential covariances were tried in this study. Similarly, many basis functions have been tried. Functions tried are constant, none, linear and pureQuadratic, respectively. The function that gave the least error in the training phase was used in the testing phase. "Subset of regressors approximation" and "fully independent conditional approximation" were used to determine beta and sigma parameters used in GPR approach [39]. In the fourth phase of the present study, ELM model was used for estimation of SR. The ELM allows to train a single hidden layer. The ELM uses feedforward network for estimation with the Moore–Penrose pseudoinverse of matrix [62]. In ELM approach, estimation models were developed with the use of different

Table 3 Comparison of the model results of the testing phase after reducing the number of inputs

Method	Criterion	Inputs									Avarage
		1	2	3	4	5	6	7	8	9	
LSTM-ADAM	RMSE	2.447	2.487	2.367	2.239	2.462	2.575	2.424	2.465	2.365	2.426
	MARE	16.718	16.843	15.965	14.291	16.062	17.299	15.369	15.360	15.652	15.951
	MAE	1.897	1.926	1.816	1.705	1.890	1.991	1.819	1.842	1.804	1.855
	R^2	0.864	0.856	0.872	0.882	0.858	0.848	0.866	0.861	0.873	0.864
	NSE	0.858	0.854	0.868	0.882	0.857	0.843	0.861	0.856	0.868	0.861
LSTM-RMSProp	RMSE	2.532	2.679	2.424	2.357	2.512	2.417	2.333	2.342	2.333	2.436
	MARE	16.522	17.851	16.426	16.179	17.423	17.706	14.272	14.174	14.247	16.089
	MAE	1.961	2.068	1.863	1.818	1.950	1.894	1.764	1.722	1.774	1.868
	R^2	0.849	0.831	0.865	0.872	0.853	0.869	0.882	0.871	0.883	0.864
	NSE	0.848	0.830	0.861	0.869	0.851	0.862	0.871	0.870	0.871	0.859
LSTM-SGDM	RMSE	2.491	2.523	2.308	2.286	2.333	2.192	2.117	2.153	2.266	2.296
	MARE	16.686	17.570	15.255	15.252	16.450	14.236	12.912	13.782	14.365	15.168
	MAE	1.936	1.959	1.769	1.756	1.799	1.673	1.588	1.635	1.713	1.759
	R^2	0.854	0.857	0.877	0.878	0.874	0.887	0.896	0.891	0.880	0.877
	NSE	0.853	0.850	0.874	0.876	0.871	0.886	0.894	0.890	0.879	0.875
SVMR-Linear	RMSE	4.198	4.116	3.764	3.741	3.743	3.309	3.300	3.297	3.268	3.637
	MARE	34.437	33.896	28.910	28.907	28.963	25.119	24.952	24.873	24.783	28.316
	MAE	3.415	3.339	3.046	3.025	3.028	2.625	2.616	2.615	2.594	2.923
	R^2	0.590	0.606	0.671	0.676	0.675	0.748	0.749	0.750	0.755	0.691
	NSE	0.417	0.401	0.335	0.331	0.331	0.259	0.257	0.257	0.253	0.316
SVMR- Polynomial	RMSE	4.261	3.771	3.483	3.430	3.448	2.325	2.199	2.226	2.248	3.043
	MARE	28.399	28.897	25.287	25.250	25.335	16.175	14.531	14.678	14.776	21.481
	MAE	3.358	3.011	2.781	2.727	2.728	1.772	1.658	1.677	1.689	2.378
	R^2	0.595	0.669	0.715	0.725	0.723	0.877	0.887	0.885	0.883	0.773
	NSE	0.429	0.336	0.287	0.278	0.281	0.128	0.114	0.117	0.120	0.232
SVMR-RBF	RMSE	3.952	3.712	3.522	3.451	3.430	2.303	2.323	2.312	2.297	3.034
	MARE	30.014	28.189	25.105	25.123	24.921	16.089	15.456	15.213	14.899	21.668
	MAE	3.158	2.960	2.789	2.722	2.707	1.763	1.742	1.725	1.705	2.363
	R^2	0.637	0.679	0.709	0.721	0.724	0.878	0.877	0.877	0.878	0.776
	NSE	0.369	0.326	0.293	0.282	0.278	0.125	0.128	0.126	0.125	0.228
GPR kernel	RMSE	3.931	3.711	3.549	3.504	3.593	2.461	2.379	2.381	2.344	3.095
	MARE	30.976	28.690	25.482	25.215	25.243	16.589	15.228	14.981	14.778	21.909
	MAE	3.185	2.991	2.826	2.791	2.845	1.871	1.785	1.770	1.755	2.424
	R^2	0.640	0.678	0.702	0.710	0.696	0.857	0.867	0.866	0.870	0.765
	NSE	0.635	0.674	0.702	0.710	0.695	0.857	0.866	0.866	0.870	0.764
GPR basis	RMSE	3.930	3.710	3.546	3.515	3.633	2.461	2.654	2.725	2.481	3.184
	MARE	30.963	28.673	25.462	25.269	25.253	16.610	16.219	17.325	15.985	22.418
	MAE	3.184	2.989	2.826	2.796	2.875	1.883	1.959	2.027	1.879	2.491
	R^2	0.640	0.678	0.703	0.708	0.691	0.857	0.845	0.834	0.856	0.757
	NSE	0.635	0.675	0.703	0.708	0.688	0.857	0.833	0.824	0.854	0.753
ELM	RMSE	4.369	3.893	3.713	3.563	3.545	2.9714	3.038	3.245	3.261	3.511
	MARE	30.138	29.451	26.504	24.498	25.656	22.1985	22.801	25.174	24.194	25.624
	MAE	3.485	3.131	2.955	2.859	2.820	2.3174	2.396	2.510	2.611	2.787
	R^2	0.616	0.650	0.694	0.712	0.707	0.796	0.787	0.769	0.749	0.720
	NSE	0.549	0.642	0.674	0.700	0.703	0.7912	0.782	0.751	0.749	0.705
KNN-Exhaustive	RMSE	5.358	5.385	4.964	4.909	4.823	3.315	3.364	3.482	3.514	4.346
	MARE	34.837	34.218	31.840	30.621	30.371	20.567	20.190	20.906	21.440	27.221

Table 3 (continued)

Method	Criterion	Inputs									Avarage
		1	2	3	4	5	6	7	8	9	
	MAE	4.185	4.196	3.862	3.716	3.673	2.444	2.465	2.581	2.623	3.305
	R^2	0.453	0.442	0.506	0.513	0.524	0.754	0.753	0.736	0.730	0.601
	NSE	0.321	0.314	0.417	0.430	0.450	0.740	0.733	0.713	0.708	0.536
KNN-Kdtree	RMSE	5.358	5.383	4.964	4.892	4.823	3.279	4.600	3.207	3.297	4.422
	MARE	34.837	34.206	31.837	30.505	30.369	19.846	28.949	18.937	19.554	27.671
	MAE	4.185	4.194	3.861	3.723	3.672	2.374	3.465	2.339	2.409	3.358
	R^2	0.453	0.442	0.506	0.515	0.524	0.758	0.558	0.773	0.759	0.588
	NSE	0.321	0.315	0.418	0.434	0.450	0.746	0.500	0.757	0.743	0.520

the number of cells in the interlayer, the standardization equation (Eq. 25) used while introducing the data to the model and the training ratio of 0.7 [62]. While estimating with ELM, the number of hidden layers was tried from 1 to 300 and the error criteria were obtained during the test phase by taking note of the number of hidden layers that gave the least RMSE error.

$$y = \frac{xi - \overline{x}}{\sigma} \tag{25}$$

The KNN model was utilized to estimate SR in the current study's final phase. To determine the k-nearest neighbors, estimate models were created using exhaustive and kdtree functions in the KNN technique. In this study, many distance metrics functions have also been tested. In order to obtain the best performance value, seuclidean, cosine, hamming, correlation, mahalanobis, jaccard, spearman distance metrics were tried in the exhaustive function. Similarly, many kdtree distance functions have been tried. Functions tried are euclidean, cityblock, minkowski and chebychev, respectively. The function that gave the least error in the training phase was used in the testing phase.

A direct comparison of the approaches is made in Table 3 for testing phase. The input parameters were introduced to the models by considering the correlation size between SR. The input parameters used for SR forecasting are T_{max}, W_{Savg}, elevation, year, longitude, month, latitude, RH_{max} and RH_{min}, respectively.

It can be noted that the LSTM model outperformed the GPR SVMR, KNN and ELM models in terms of each average performance metrics at the testing phase. MARE values (%) varied between 15.17 and 28.31, MAE values between 1.759 and 3.358 and RMSE values between 2.297 and 4.422. In the testing phase, the best input combination was observed in the SGDM optimization algorithm of LSTM, in which 7 input parameters were used. When the models are compared with the best in themselves, the kernel function is superior to the basis function in GPR. Similarly, SGDM is the LSTM optimization algorithm that gives the least error metric, followed by Adam and PMSEProp. In SVMR, on the other hand, while polynomial is given the least faulty function, it is followed by RBF and linear functions. The lowest error in KNN was observed in kdtree function, followed by exhaustive the function. The optimum sets of model inputs for each of the investigated predictive modeling strategies were not the same, demonstrating that each model type reacts differently to distinct input variable sets and data patterns/attributes in the input data [76]. Overall, the best accurate input combinations for the LSTM, SVMR, GPR, ELM and KNN were based on models 7, 7, 9, 6 and 8. Evaluation of different modeling approaches (LSTM, SVMR, GPR, ELM and KNN) with different sets of input variables (i.e., 1–9) shows that the most accurate predictions depend on the model used and the optimization of the model.

NSE values of less than one are ideal, as this indicates a 100 percent success rate. Low estimation success is indicated by NSE values between 0.3 and 0.5, acceptable estimation success is shown by NSE values between 0.5 and 0.7, great estimation success is indicated by NSE values between 0.7 and 0.9, and outstanding estimation success is indicated by NSE values between 0.9 and 1 [71]. In Table 3, mean NSE values ranged between 0.228 and 0.875. These values indicate that the models in which some inputs are used show low estimation success, but the best model shows great estimation success. For example, according to the NSE criterion, the most successful method was obtained in the modeling using 7 input combinations in the SGDM architecture of the LSTM approach. The estimating power of the LSTM with SGDM model was fairly good. The current findings demonstrated that the LSTM model could overcome nonlinear relationships between variables, indicating that it performed well.

The scatter-plots for LSTM, SVMR, GPR, ELM and KNN models are shown in Fig. 8. Figure 8 shows the regression coefficient R^2 and the regression equation ($y = ax + b$). With a best R^2 value of 0.8957, the LSTM model was able to obtain the best fit line between observed and anticipated SR values using the 7-input combination. The SVMR, GPR, ELM and KNN approaches had R^2 = 0.8871, 0.8701, 0.796 and 0.773, respectively.

Figure 8 shows the observed and estimated SR values for the four models during the testing phase. This is an indication of the variation of underestimated or overestimated SR values. As shown in the figure, low SR values are too high, and values are slightly overestimated. (This

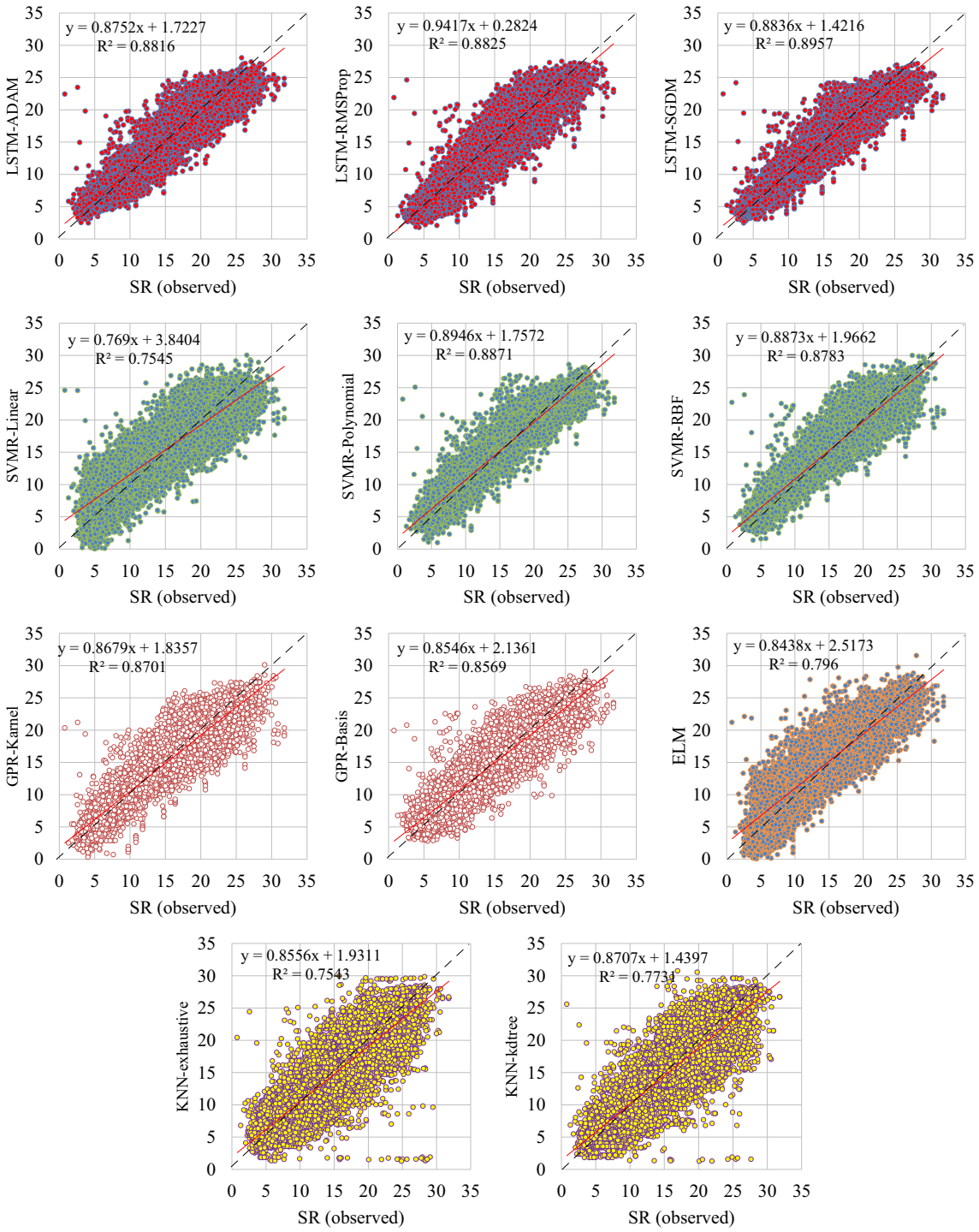

Fig. 8 Scatterplot comparison of measured and models predicted SR values

can be observed by following the dashed black line, 1:1 or the best line, $y = x$.) The relatively weak performance for these extreme values of SR indicates that the model is likely to fall short on the training data set used to estimate its parameters. When the relationship between the scatter-plots of the model results and the best line is examined, it is observed that the models forecast low SR values (< 10 MJ/m^2) higher and higher values (> 20 MJ/m^2) low. This can be observed at the intersection of the best line and the model line (red) and is a disadvantage observed in all approaches. Although all approaches better predict intermediate SR values (10–20 MJ/m^2), some deviations were observed when estimating high and low values. It is observed that the convergence to the best line is mostly in the LSTM-SGDM approach, and some low values are estimated with quite outlier (with high values) values in the KNN approach.

In previous statements, LSTM was considered as the best model for SR forecast since it had the least RMSE, MAE and MARE and highest R^2 and NSE values. All data were distributed around the regression line in scatter plots. In these plots, it was discovered that all models essentially followed the same regression lines. Although the MSE, MAE and RMSE error criteria indicated the correctness of the forecasted variables, they do not offer information about the models' distribution [39]. Therefore, violin plot (Fig. 9), box error plot (Fig. 10) and Taylor diagram (Fig. 11) were used for comparison.

The conformity of estimation data with observed data was examined using the violin plot. Further statistical comparisons of the models were conducted using the violin plot. Figure 9 shows a violin plot for the best outcome of the LSTM, SVMR, GPR, ELM and KNN techniques. Differences are seen in each ML approach based on the errors presented by the box plots (Fig. 10), with smaller error values circled for the GRP, LSTM, SVMR and ELM models. The error graph was obtained by subtracting the predicted values from the observed values by absolute value [77]. The Taylor graph in Fig. 11 is the graphical representation of Eq. 20 (RMSD) between the model and the observed values and the correlation between these two values [78].

The five best models in Fig. 9 were very similar to each other; however, LSTM was distinguishable from the other four approaches in Fig. 10's box plot diagrams and errors diagrams. The extreme error values of these models are almost at the same level; however, the KNN method differs from other methods based on excessive errors for its estimations. When the error graph is examined, it is seen that in particular the KNN model overestimates, while the ELM model underestimates. When using the Taylor, the LSTM model produced SR estimates that were quite similar to observed values. The Taylor diagram also demonstrated that the LSTM technique outperformed the other models.

It was quite difficult to identify the superior method for SR estimation in the study. For this reason, many statistical and graphical methods have been used. Finally, spider plots, which are used to evaluate all error criteria of the best approaches, were drawn in the study. Figure 12 shows the spider plot.

Thanks to the spider graph, it can be easily seen that LSTM is less than other approaches according to the RMSE, MARE and MAE criteria, while the exact opposite NSE and R^2 values are better than other methods. In addition, it has been determined that the least successful method is the KNN technique.

Finally, statistical significance comparisons between the results of the five approaches and the observed data were

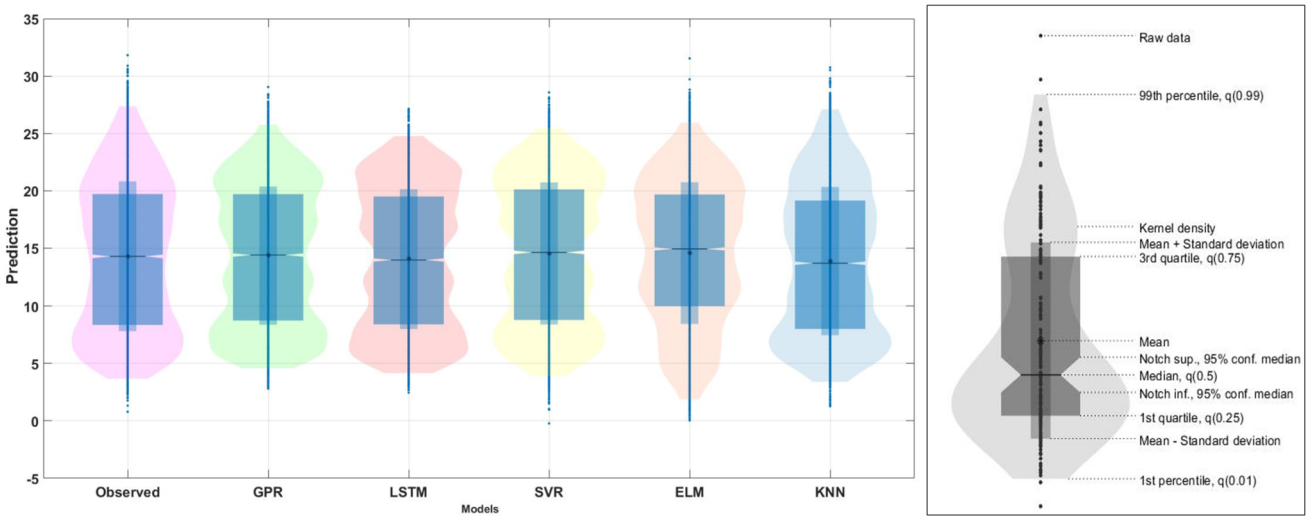

Fig. 9 Violin plot for GPR, LSTM, SVMR, ELM and KNN approaches

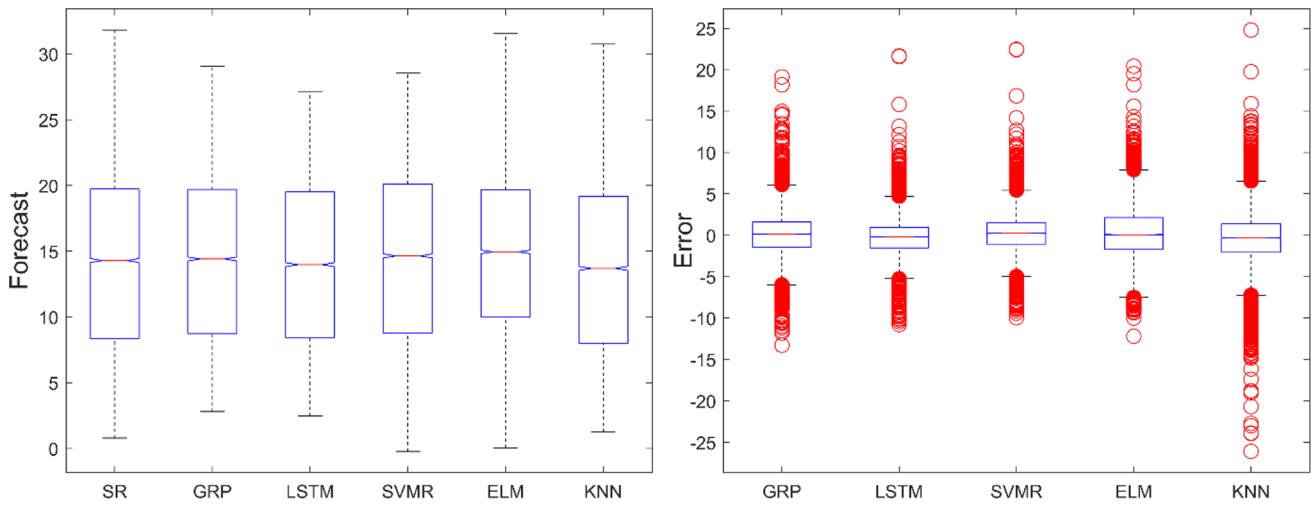

Fig. 10 Box plot diagrams and errors diagrams for GPR, LSTM, SVMR, ELM and KNN approaches

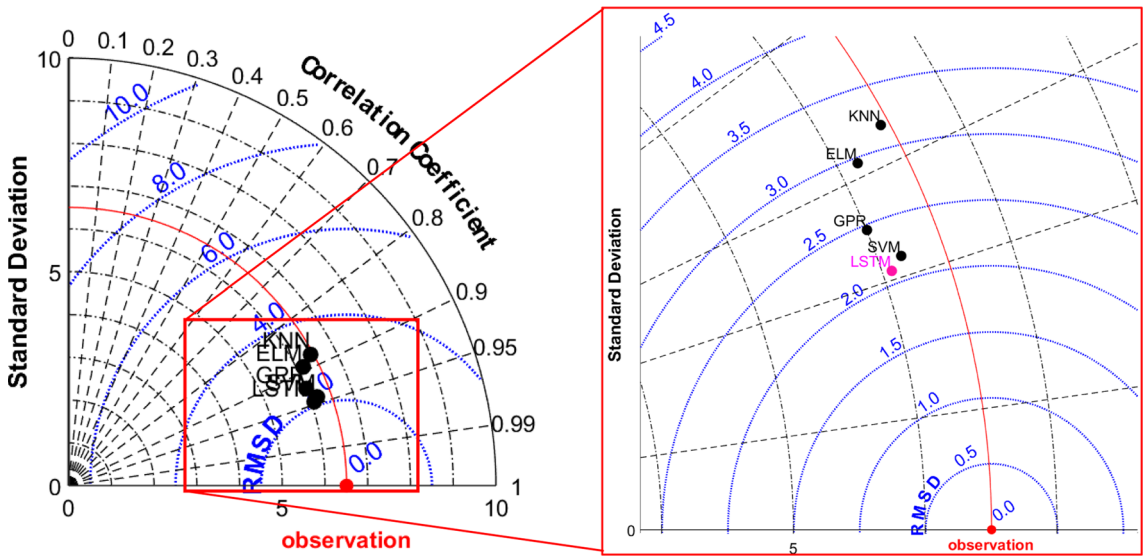

Fig. 11 Taylor diagrams for GPR, LSTM, SVMR, ELM and KNN approaches

made. Firstly, the Kruskal–Wallis test was employed to see whether the distributions of the estimated and measured data were identical. [39, 79]. In the estimations of the three approaches (GPR, LSTM and KNN) in Table 4, the H_0 hypothesis is rejected. In other words, it demonstrates that the means of the anticipated and observed data are not significantly different. Other models, on the other hand, have a considerable difference, and it is likely that the model findings are not from the same field as the actual data. The KW test was performed at 95 percent of the confidence interval.

With the KW test in Table 4, it has been seen that the models have less errors, which does not indicate that the technique is fully appropriate. This result shows that these models do not always provide reliable SR estimates due to the complex connections between independent and dependent variables. In particular, the large number of data and the inability to predict the extreme values well cause the H_0 hypothesis to be accepted in the KW test [80]. In Table 4, the GPR, LSTM and KNN approaches passed the KW test, meaning that the estimates given by these methods come from the same mean as the measured SR. Then, test results of the applied models were also evaluated by one-way analysis of variance (ANOVA) for evaluating the robustness (the significance of differences between the measured and estimated SR values) of the different machine learning approaches [81]. The test was set at a 95% significance level. Table 5 gives the test statistics.

In Table 5, the GPR kernel has the lowest test value (0.51) with the highest significance level (0.4734) compared with the others. According to the ANOVA test, the GPR kernel model is more robust than the GPR basis and

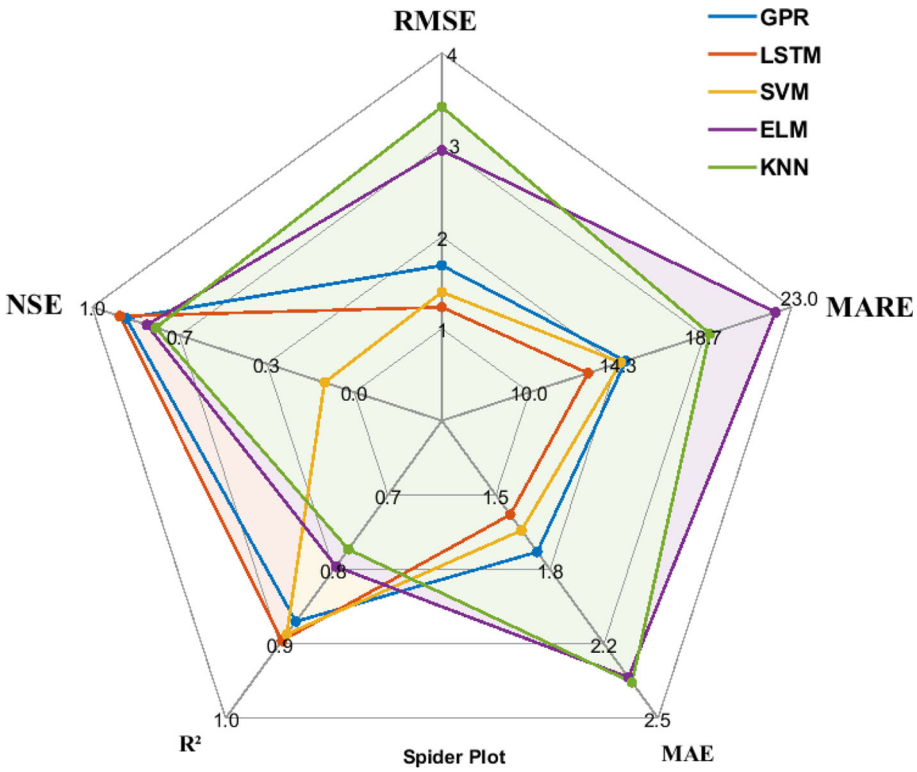

Fig. 12 Spider graph for best approaches

Table 4 KW test results

Method		*P* value	Critical value	Ho*
GPR	Kernel	0.7327	0.05	Reject
	Basis	0.0849	0.05	Reject
LSTM	ADAM	0.8302	0.05	Reject
	RMSProp	7.44×10^{-11}	0.05	Accept
	SGDM	0.0353	0.05	Accept
SVR	Linear	2.00×10^{-16}	0.05	Accept
	Polynomial	2.45×10^{-5}	0.05	Accept
	RBF	1.42×10^{-7}	0.05	Accept
ELM	–	1.23×10^{-6}	0.05	Accept
KNN	Exhaustive	0.1696	0.05	Reject
	Kdtree	5.46×10^{-7}	0.05	Accept

Ho*means that there are differences between the mean estimate and measurement values

LSTM-Adam models (the similarity between the measured SR values and the GPR kernel forecasts is significantly high) in modeling monthly SR. All other methods failed this test. Thus, unlike the ones stated above, it was decided that GPR kernel and LSTM ADAM were the most successful methods in this study.

6 Conclusion

The primary aim of the current research is to make SR prediction with different machine learning approaches. It was also to investigate the applicability and capacity of ML approaches by examining the effect of input parameters on forecast accuracy and removing parameters that decrease forecast accuracy to increase forecast performance. Finally, the most important findings of this study might be stated as follows:

1. VIF analysis was performed to develop the model. Thus, the input parameters that reduce the performance of the model are eliminated.
2. When the models are compared within themselves, kernel function is superior to the basis function in GPR, Polynomial is superior to the RBF and linear function in SVMR, SGDM is superior to the Adam and RMSProp optimization algorithm in LSTM and Kdtree function is superior to the exhaustive function in KNN.
3. The error criteria of MAE, MARE, RMSE, R^2 and NSE, the results were analyzed according to the Taylor, violin, box error and spider plots and it was decided that the method that best predicted the observed values was LSTM. It is followed by GRP, SVMR, ELM and KNN.
4. LSTM model average performance metrics at the testing phase. MARE values (%) varied between

Table 5 ANOVA test results of the LSTM, SVMR, GPR, ELM and KNN techniques in the testing phase

Method		*F*-statistics	Resultant singnificance level	Ho*
GPR	Kernel	0.51	0.4734	Accept
	Basis	0.53	0.4683	Accept
LSTM	ADAM	0.67	0.4135	Accept
	RMSProp	48.03	4.29×10^{-12}	Reject
	SGDM	10.1	0.0015	Reject
SVR	Linear	50.56	1.18×10^{-12}	Reject
	Polynomial	10.27	0.0014	Reject
	RBF	20.86	4.96×10^{-6}	Reject
ELM	–	13.23	0.0003	Reject
KNN	Exhaustive	2.93	0.0868	Accept
	Kdtree	26.82	2.52×10^{-7}	Reject

Ho* means that there are same between the mean estimate and measurement values

15.17 and 28.31, MAE values between 1.759 and 3.358, RMSE values between 2.297 and 4.422 and mean NSE values reached 0.875.

5. In addition, statistical significance test of the analysis results was performed with KW and ANOVA. It was concluded that the method that is more robust than other methods is GPR. This is followed by LSTM and KNN. With these tests, it was concluded that the predictions of the SVMR and ELM models were doubtful, while the predictions of the GPR, LSTM and KNN models could represent the mean.
6. Finally, these results proved that LSTM and GPR algorithms are applicable, valid and an alternative for SR estimation in Turkey, which has arid and semi-arid climatic regions.

The seven main limitations of this study can be mentioned as follows: (i) using data from 163 meteorological stations to represent Turkey, (ii) using data from 1967 to 2020, (iii) using VIF analysis for input selection, (iv) using different optimization techniques and five different machine learning methods, (v) using visual comparison criteria (violin, Taylor, spider and box plot) in addition to performance metrics and (vi) KW and ANOVA tests are used in the accuracy of the results.

This study is an effort to estimate SR in Turkey, which is of great importance in energy balances and production, biological processes, hydrological cycle, terrestrial biological ecosystems and climate. In future studies, the accuracy of the regional study can be increased by providing new machine learning methods. In addition to machine learning methods, it is considered to develop models that give equations using nature-inspired optimization algorithms and input parameters.

Funding No funding or support has been received for research from funding institutions.

Availability of data and material Climatic data and hydrometric data were provided by the General Directorate of State Meteorological Affairs (DMİ) and General Directorate of State Hydralics Works (DSİ).

Code availability Not applicable.

Declarations

Conflicts of interest The authors declare no conflicts of interest.

Ethics approval The author paid attention to the ethical rules in the study. There is no violation of ethics.
Consent for publication: If this study is accepted, it can be published in the **Neural Computing and Applications.**

References

1. Keller B, Costa AMS (2011) A Matlab GUI for calculating the solar radiation and shading of surfaces on the earth. Comput Appl Eng Educ 19:161–170. https://doi.org/10.1002/cae.20301
2. Beer C, Reichstein M, Tomelleri E, Ciais P, Jung M, Carvalhais N et al (2010) Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. Science (80-) 329:834–838. https://doi.org/10.1126/science.1184984
3. Islam MD, Kubo I, Ohadi M, Alili AA (2009) Measurement of solar energy radiation in Abu Dhabi. UAE Appl Energy 86:511–515. https://doi.org/10.1016/j.apenergy.2008.07.012
4. Khatib T, Mohamed A, Sopian K (2012) A review of solar energy modeling techniques. Renew Sustain Energy Rev 16:2864–2869. https://doi.org/10.1016/j.rser.2012.01.064
5. Meza F, Varas E (2000) Estimation of mean monthly solar global radiation as a function of temperature. Agric For Meteorol 100:231–241. https://doi.org/10.1016/S0168-1923(99)00090-8
6. Kisi O (2014) Modeling solar radiation of Mediterranean region in Turkey by using fuzzy genetic approach. Energy 64:429–436. https://doi.org/10.1016/j.energy.2013.10.009
7. Panwar NL, Kaushik SC, Kothari S (2011) Role of renewable energy sources in environmental protection: a review. Renew Sustain Energy Rev 15:1513–1524. https://doi.org/10.1016/j.rser.2010.11.037
8. Park J-K, Das A, Park J-H (2015) A new approach to estimate the spatial distribution of solar radiation using topographic factor and

sunshine duration in South Korea. Energy Convers Manag 101:30–39. https://doi.org/10.1016/j.enconman.2015.04.021
9. Purohit I, Purohit P (2015) Inter-comparability of solar radiation databases in Indian context. Renew Sustain Energy Rev 50:735–747. https://doi.org/10.1016/j.rser.2015.05.020
10. Wild M (2009) Global dimming and brightening: a review. J Geophys Res 114:D00D16. https://doi.org/10.1029/2008JD011470
11. Wang L, Kisi O, Zounemat-Kermani M, Salazar GA, Zhu Z, Gong W (2016) Solar radiation prediction using different techniques: model evaluation and comparison. Renew Sustain Energy Rev 61:384–397. https://doi.org/10.1016/j.rser.2016.04.024
12. Ndulue E, Onyekwelu I, Ogbu KN, Ogwo V (2019) Performance evaluation of solar radiation equations for estimating reference evapotranspiration (ETo) in a humid tropical environment. J Water L Dev 42:124–135. https://doi.org/10.2478/jwld-2019-0053
13. Ododo JC, Sulaiman AT, Aidan J, Yuguda MM, Ogbu FA (1995) The importance of maximum air temperature in the parameterisation of solar radiation in Nigeria. Renew Energy 6:751–763. https://doi.org/10.1016/0960-1481(94)00097-P
14. Bandyopadhyay A, Bhadra A, Raghuwanshi NS, Singh R (2008) Estimation of monthly solar radiation from measured air temperature extremes. Agric For Meteorol 148:1707–1718. https://doi.org/10.1016/j.agrformet.2008.06.002
15. Ododo JC (1997) Prediction of solar radiation using only maximum temperature and relative humidity: south-east and north-east Nigeria. Energy Convers Manag 38:1807–1814
16. Rehman S, Mohandes M (2008) Artificial neural network estimation of global solar radiation using air temperature and relative humidity. Energy Policy 36:571–576. https://doi.org/10.1016/j.enpol.2007.09.033
17. Kisi O, Alizamir M, Trajkovic S, Shiri J, Kim S (2020) Solar radiation estimation in Mediterranean climate by weather variables using a novel Bayesian model averaging and machine learning methods. Neural Process Lett 52:2297–2318. https://doi.org/10.1007/s11063-020-10350-4
18. Alsafadi M, Filik ÜB (2020) Hourly global solar radiation estimation based on machine learning methods in Eskişehir. Eskişehir Tech Univ J Sci Technol A Appl Sci Eng 21:294–313. https://doi.org/10.18038/estubtda.650497
19. Kumar R, Aggarwal RK, Sharma JD (2015) Comparison of regression and artificial neural network models for estimation of global solar radiations. Renew Sustain Energy Rev 52:1294–1299. https://doi.org/10.1016/j.rser.2015.08.021
20. Chabane F, Arif A, Benramache S (2020) Prediction of the solar radiation map on algeria by latitude and longitude coordinates. Tec Ital J Eng Sci 64:213–215. https://doi.org/10.18280/ti-ijes.642-413
21. Rahimikhoob A, Behbahani SMR, Banihabib ME (2013) Comparative study of statistical and artificial neural network's methodologies for deriving global solar radiation from NOAA satellite images. Int J Climatol 33:480–486. https://doi.org/10.1002/joc.3441
22. Polo J, Antonanzas-Torres F, Vindel JM, Ramirez L (2014) Sensitivity of satellite-based methods for deriving solar radiation to different choice of aerosol input and models. Renew Energy 68:785–792. https://doi.org/10.1016/j.renene.2014.03.022
23. Ahmad MJ, Tiwari GN (2011) Solar radiation models—a review. Int J Energy Res 35:271–290. https://doi.org/10.1002/er.1690
24. Sonmete MH, Ertekin C, Menges HO, Hacıseferoğullari H, Evrendilek F (2011) Assessing monthly average solar radiation models: a comparative case study in Turkey. Environ Monit Assess 175:251–277. https://doi.org/10.1007/s10661-010-1510-8
25. Citakoglu H (2015) Comparison of artificial intelligence techniques via empirical equations for prediction of solar radiation. Comput Electron Agric 118:28–37. https://doi.org/10.1016/j.compag.2015.08.020
26. Yacef R, Mellit A, Belaid S, Şen Z (2014) New combined models for estimating daily global solar radiation from measured air temperature in semi-arid climates: application in Ghardaïa, Algeria. Energy Convers Manag 79:606–615. https://doi.org/10.1016/j.enconman.2013.12.057
27. Citakoglu H, Babayigit B, Haktanir NA (2020) Solar radiation prediction using multi-gene genetic programming approach. Theor Appl Climatol 142:885–897. https://doi.org/10.1007/s00704-020-03356-4
28. Ozoegwu CG (2019) Artificial neural network forecast of monthly mean daily global solar radiation of selected locations based on time series and month number. J Clean Prod 216:1–13. https://doi.org/10.1016/j.jclepro.2019.01.096
29. Citakoglu H, Demir V (2021) Forecasting solar radiation using deep learning: the case of Turkey. In: International World energy conference, pp 167–175
30. Cano D, Monget JM, Albuisson M, Guillard H, Regas N, Wald L (1986) A method for the determination of the global solar radiation from meteorological satellite data. Sol Energy 37:31–39. https://doi.org/10.1016/0038-092X(86)90104-0
31. Al-Mostafa ZA, Maghrabi AH, Al-Shehri SM (2014) Sunshine-based global radiation models: a review and case study. Energy Convers Manag 84:209–216. https://doi.org/10.1016/j.enconman.2014.04.021
32. Badescu V, Dumitrescu A (2016) Simple solar radiation modelling for different cloud types and climatologies. Theor Appl Climatol 124:141–160. https://doi.org/10.1007/s00704-015-1400-7
33. Samuel CN (2017) A comprehensive review of empirical models for estimating global solar radiation in Africa. Renew Sustain Energy Rev 78:955–995. https://doi.org/10.1016/j.rser.2017.04.101
34. Fan J, Chen B, Wu L, Zhang F, Lu X, Xiang Y (2018) Evaluation and development of temperature-based empirical models for estimating daily global solar radiation in humid regions. Energy 144:903–914. https://doi.org/10.1016/j.energy.2017.12.091
35. Fan J, Wu L, Zhang F, Cai H, Zeng W, Wang X et al (2019) Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: a review and case study in China. Renew Sustain Energy Rev 100:186–212. https://doi.org/10.1016/j.rser.2018.10.018
36. Sharafati A, Khosravi K, Khosravinia P, Ahmed K, Salman SA, Yaseen ZM et al (2019) The potential of novel data mining models for global solar radiation prediction. Int J Environ Sci Technol 16:7147–7164. https://doi.org/10.1007/s13762-019-02344-0
37. Tao H, Ewees AA, Al-Sulttani AO, Beyaztas U, Hameed MM, Salih SQ et al (2021) Global solar radiation prediction over North Dakota using air temperature: development of novel hybrid intelligence model. Energy Rep 7:136–157. https://doi.org/10.1016/j.egyr.2020.11.033
38. Guermoui M, Melgani F, Gairaa K, Mekhalfi ML (2020) A comprehensive review of hybrid models for solar radiation forecasting. J Clean Prod. https://doi.org/10.1016/j.jclepro.2020.120357
39. Citakoglu H (2021) Comparison of multiple learning artificial intelligence models for estimation of long-term monthly temperatures in Turkey. Arab J Geosci 14:2131. https://doi.org/10.1007/s12517-021-08484-3
40. Ly HB, Nguyen TA, Pham BT (2021) Estimation of soil cohesion using machine learning method: a random forest approach. Adv Civ Eng. https://doi.org/10.1155/2021/8873993
41. Heddam S, Kisi O (2017) Extreme learning machines: a new approach for modeling dissolved oxygen (DO) concentration with

and without water quality variables as predictors. Environ Sci Pollut Res 24:16702–16724. https://doi.org/10.1007/s11356-017-9283-z
42. Naghibi SA, Moghaddam DD, Kalantar B, Pradhan B, Kisi O (2017) A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. J Hydrol 548:471–483. https://doi.org/10.1016/j.jhydrol.2017.03.020
43. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
44. Hrnjica B, Mehr AD (2020) Energy demand forecasting using deep learning. In: EAI/Springer ınnovations in communication and computing, pp 71–104. https://doi.org/10.1007/978
45. Sattari MT, Apaydin H, Band SS, Mosavi A, Prasad R (2021) Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. Hydrol Earth Syst Sci 25:603–618. https://doi.org/10.5194/hess-25-603-2021
46. Rahimzad M, Moghaddam Nia A, Zolfonoon H, Soltani J, Danandeh Mehr A, Kwon HH (2021) Performance comparison of an LSTM-based deep learning model versus conventional machine learning algorithms for streamflow forecasting. Water Resour Manag 35:4167–4187. https://doi.org/10.1007/s11269-021-02937-w
47. Liu M, Huang Y, Li Z, Tong B, Liu Z, Sun M et al (2020) The applicability of LSTM-KNN model for real-time flood forecasting in different climate zones in China. Water (Switzerland) 12:1–21. https://doi.org/10.3390/w12020440
48. Pandey MK, Srivastava PK (2021) A probe into performance analysis of real-time forecasting of endemic ınfectious diseases using machine learning and deep learning algorithms, pp 241–65. https://doi.org/10.1007/978-981-16-0538-3_12
49. Ser G, Bati CT (2019) Determining the best model with deep neural networks: Keras application on mushroom data. Yuz Yil Univ J Agric Sci 29:406–417. https://doi.org/10.29133/yyutbd.505086
50. Vapnik V (1995) The nature of statistical learning theory. Springer, New York
51. Smola A (1996) Regression estimation with support vector learning machines, Master's thesis, Tech Univ Munchen. Master's thesis, Tech Univ Munchen
52. Smola AJ, Scholkopf B (1998) A tutorial on support vector regression. R Hollow Coll London, UK, NeuroCOLT Tech,Technical Rep Ser
53. Smola AJ, Olkopf BSCH (2004) A tutorial on support vector regression. Kluwer Acad Publ Manuf Netherlands 14:199–222
54. Eldakhly N, Aboul-Ela MM, Abdalla A (2018) A novel approach of weighted support vector machine with applied chance theory for forecasting air pollution phenomenon in Egypt. Int J Comput Intell Appl 17(1). https://doi.org/10.1142/S1469026818500013
55. Khan MS, Coulibaly P (2006) Application of support vector machine in lake water level prediction. J Hydrol Eng 11:199–205. https://doi.org/10.1061/(asce)1084-0699(2006)11:3(199)
56. Yin Z, Wen X, Feng Q, He Z, Zou S (2017) Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area. Hydrol Res. https://doi.org/10.2166/nh.2016.205
57. Gunn S (1998) Support vector machines for classification and regression, Univ Southapt, Image Speech Intell Syst Res Group
58. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. vol. 7. The MIT Press, Massachusetts Institute of Technology
59. Kocijan J, Ažman K, Grancharova A (2007) The concept for Gaussian process model based system identification toolbox. In: Proceedings of the 2007 International Conference on Computer Computer Systems and Technologies—CompSysTech '07, New York, New York, USA: ACM Press; p. 1. https://doi.org/10.1145/1330598.1330647
60. Neal RM (1996) Bayesian learning for neural networks, vol 118. Springer, New York. https://doi.org/10.1007/978-1-4612-0745-0
61. Wang J, Lu S, Wang S, Zhang Y (2021) A review on extreme learning machine Jian. Multimed Tools Appl. https://doi.org/10.1007/s11042-021-11007-71181
62. Huang G, Member S, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern Part B (Cybern) 42:513–529
63. Huang G-B, Zhu Q-Y, Siew C-K (2006) Extreme learning machine: Theory and applications. Neurocomputing 70:489–501. https://doi.org/10.1016/j.neucom.2005.12.126
64. Huang G, Chen L (2007) Convex incremental extreme learning machine. Neurocomputing 70:3056–3062. https://doi.org/10.1016/j.neucom.2007.02.009
65. Yaseen ZM, Sulaiman SO, Deo RC, Chau KW (2019) An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. J Hydrol 569:387–408. https://doi.org/10.1016/j.jhydrol.2018.11.069
66. Fix E, Hodges JL (1951) Discriminatory analysis. Nonparametric discrimination: consistency properties. USAF School of Aviation Medicine, Randolph Field, Texas.. https://doi.org/10.2307/1403797
67. Altman NS (1992) An Introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46:175–185. https://doi.org/10.1080/00031305.1992.10475879
68. Luo X, Li D, Yang Y, Zhang S (2019) Spatiotemporal traffic flow prediction with KNN and LSTM. J Adv Transp. https://doi.org/10.1155/2019/4145353
69. Kramer O (2013) Dimensionality reduction with unsupervised nearest neighbors, vol 51. Springer, Berlin. https://doi.org/10.1007/978-3-642-38652-7
70. Voyant C, Notton G, Kalogirou S, Nivet M-L, Paoli C, Motte F et al (2017) Machine learning methods for solar radiation forecasting: a review. Renew Energy 105:569–582. https://doi.org/10.1016/j.renene.2016.12.095
71. Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—a discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6
72. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106:7183–7192. https://doi.org/10.1029/2000JD900719
73. Legouhy A (2021) al_goodplot—boxblot & violin plot. MATLAB Cent Mathworks. https://www.mathworks.com/matlabcentral/fileexchange/91790-al_goodplot-boxblot-violin-plot.
74. Uncuoğlu E, Latifoğlu L, Özer AT (2021) Modelling of lateral effective stress using the particle swarm optimization with machine learning models. Arab J Geosci 14:2441. https://doi.org/10.1007/s12517-021-08686-9
75. Moses (2022) Spider Plot. GitHub. https://github.com/NewGuy012/spider_plot/releases/tag/17.2. Accessed 4 Feb 2022
76. Yaseen ZM, Jaafar O, Deo RC, Kisi O, Adamowski J, Quilty J et al (2016) Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. J Hydrol 542:603–614. https://doi.org/10.1016/j.jhydrol.2016.09.035
77. Başakın EE, Ekmekcioğlu Ö, Çıtakoğlu H, Özger M (2022) A new insight to the wind speed forecasting: robust multi-stage ensemble soft computing approach based on pre-processing uncertainty assessment. Neural Comput Appl 34:783–812. https://doi.org/10.1007/s00521-021-06424-6
78. Uncuoglu E, Citakoglu H, Latifoglu L, Bayram S, Laman M, Ilkentapar M, Onera AA (2022) Comparison of neural network,

Gaussian regression, support vector machine, long short-term memory, multi-gene genetic programming, and M5 Trees methods for solving civil engineering problems. Appl Soft Comput 129:109623
79. Görkemli B, Citakoglu H, Haktanir T, Karaboga D (2022) A new method based on artificial bee colony programming for the regional standardized intensity–duration-frequency relationship. Arab J Geosci. https://doi.org/10.1007/s12517-021-09377-1
80. Citakoglu H, Demir V (2022) Developing numerical equality to regional intensity duration–frequency curves using evolutionary algorithms and multi-gene genetic programming. Acta Geophys. https://doi.org/10.1007/s11600-022-00883-8
81. Kisi O, Demir V, Kim S (2017) Estimation of long-term monthly temperatures by three different adaptive neuro-fuzzy approaches using geographical inputs. J Irrig Drain Eng. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001242