Environmental Science and Pollution Research (2023) 30:12769–12784
https://doi.org/10.1007/s11356-022-22934-9

RESEARCH ARTICLE

Particle swarm optimization algorithm with Gaussian exponential model to predict daily and monthly global solar radiation in Northeast China

Yue Jia[1,2] · Hui Wang[1,2] · Pengcheng Li[1,2] · Yongjun Su[1,2] · Fengchun Wang[1,2] · Shuyi Huo[1,2]

Received: 4 April 2022 / Accepted: 4 September 2022 / Published online: 17 September 2022

Abstract

Reliable global solar radiation (R_s) information is crucial for the design and management of solar energy systems for agricultural and industrial production. However, R_s measurements are unavailable in many regions of the world, which impedes the development and application of solar energy. To accurately estimate R_s, particle swarm optimization (PSO) algorithm integrating Gaussian exponential model (GEM) was proposed for estimating daily and monthly global R_s in Northeast China. The PSO-GEM was compared with four other machine learning models and two empirical models to assess its applicability using daily meteorological data from 1997 to 2016 from four stations in Northeast China. The results showed that in different stations, the PSO-GEM with full climatic data as inputs showed the highest accuracy to estimate daily R_s with RMSE, RRMSE, MAE, R^2, and E_{ns} values of 1.045–1.719 MJ m^{-2} d^{-1}, 7.6–12.7%, 0.801–1.283 MJ m^{-2} d^{-1}, 0.953–0.981, and 0.946–0.977, respectively. The PSO-GEM showed the highest accuracy to estimate monthly R_s with RMSE, RRMSE, MAE, R^2, and E_{ns} values of 0.197–0.575 MJ m^{-2} d^{-1}, 1.5–7.0%, 0.137–0.499 MJ m^{-2} d^{-1}, 0.999–1, and 0.992–0.999, respectively. Overall, the PSO-GEM had the highest accuracy under different inputs and is recommended for modeling daily and monthly R_s in Northeast China.

Keywords Global solar radiation · Gaussian exponential model · Particle swarm optimization · Machine learning models · Empirical models

Introduction

Solar radiation (R_s) provides the essential energy for life on Earth (Wild et al. 2005) and is the foundation of global climate formation (Antonopoulos et al. 2019). Solar energy is one of the most advantageous energy sources, as it is clean, free, abundant, and inexhaustible (Khatib et al. 2012; Desideri et al. 2013; Jamil and Akhtar 2017; Zhang et al. 2019). As the global energy demand is gradually increasing, solar energy has attracted increasing attention. The application of solar energy systems depends on the amount and intensity of global R_s; thus, reliable information on R_s directly affects the development of solar energy (Citakoglu 2015; Zhang et al. 2019). Furthermore, the level of R_s is directly related to the characteristics of regional climate change and the layout of agricultural production, especially crop production (Bailek et al. 2018; Fan et al. 2019; Jiang et al. 2020; Wu et al. 2022a). The most accurate R_s data can be obtained by measurements (Fan et al. 2019). However, the high requirements and costs of the measuring devices have resulted in few measurements worldwide (Besharat et al. 2013; Oates et al. 2017; Feng et al. 2020). China has the largest energy demand in the world. Among the 752 national meteorological stations in China, only 122 stations have measured R_s data (Pan et al. 2013). Thus, using other commonly available climatic data to predict R_s is a feasible alternative.

Various climatic variables, such as precipitation (P), sunshine duration (n), air temperature, and relative humidity (H_r), are effective factors for R_s estimation (Katiyar and Pandey 2010; Jamil and Akhtar 2017; Jamil and Siddiqui

Responsible Editor: Philippe Garrigues

✉ Hui Wang
wanghuihebei@126.com

[1] Hebei University of Water Resources and Electric Engineering & Remote Sensing and Smart Water Innovation Center, Cangzhou 061001, China

[2] Center for Water Automation and Information Application Technology, Hebei University, Cangzhou 061001, China

2018; Kaba et al. 2018; Wu et al. 2022b). Thus, various types of models have been developed based on these climatic variables, including empirical models (Liu et al. 2009; Citakoglu 2015; Demircan et al. 2020; Feng et al. 2021a), machine learning models (Hossain et al. 2017; Fan et al. 2018; Feng et al. 2019c), and radiative transfer models (Gueymard 2001; Wu et al. 2020). Owing to the acceptable accuracy and low computational costs and input requirements, empirical models are the most widely applied models (Hassan et al. 2016), among which the Hargreaves–Samani (HS) model and Bristow–Campbell (BC) model are two well-known empirical models. Liu et al. (2009) modified the HS and BC models in different regions of China, and found that the accuracy of the models was improved by 4–7% after correction.

Because R_s has a nonlinear relationship with other climatic variables, as indicated by empirical models, machine learning models can improve the accuracy of R_s estimation and prediction (Chen et al. 2011). To date, many machine learning models have been extensively applied to estimate and simulate R_s (Katiyar and Pandey 2010; Jamil and Akhtar 2017; Kaba et al. 2018; Feng et al. 2019a), such as the adaptive neuro-fuzzy inference system (ANFIS) (Tabari et al. 2012), M5 model tree (Kisi 2016), random forests (Feng et al. 2017a), and gene expression programming (Shiri et al. 2014). Bueno et al. (2019) evaluated the performances of neural networks, support vector regression, and Gaussian processes for R_s prediction using satellite data as inputs, and reported that the three machine learning models provided reliable estimates. Zou et al. (2017) compared the ANFIS model with an improved BC model and Yang's model for R_s estimation, and found that machine learning models showed better results than the BC model and Yang's model. Fan et al. (2019) compared 12 machine learning models and 12 empirical models to estimate R_s. They showed that the ANFIS model, MARS model, and XGBoost model may be promising models in China.

Although machine learning models have improved the accuracy for estimating R_s, they still have some issues to deal with. The parameters random selection of traditional machine learning models can affect the calculation accuracy. The particle swarm optimization (PSO) algorithm can solve the limitations of parameters and improve the accuracy of traditional machine learning models. Gaussian exponential model (GEM) is a novel machine learning model that has not been applied to R_s estimation. To further improve the accuracy of GEM, the PSO algorithm was utilized and the PSO-GEM was developed in this paper. To confirm the accuracy of PSO-GEM and GEM, we compared the models with three traditional machine learning models (M5 model tree (M5T), support vector machine (SVM), random forest (RF)) and two empirical models (HS and BC). China consumes a large amount of energy, and a significant amount of energy is used for economic development every year (Liu et al. 2017; Fan et al. 2018). Clean solar energy is of great significance for energy conservation and emission reduction (Jin et al. 2005; Feng et al. 2021b). Northeast China, which is the main industrial production region, accounts for approximately 20% of China's energy consumption (Zheng et al. 2019). Therefore, determining an optimal R_s model for this region can provide scientific information for solar energy applications. However, the performance of different models in this region has not been well documented. Thus, in this paper, PSO-GEM and GEM were developed to estimate R_s in Northeast China with different climate data. The main purpose of this study was to examine the applicability of five machine learning models (M5T model, SVM model, RF model, GEM, PSO-GEM, HS, and BC) for R_s prediction in Northeast China.

Methods and materials

Study area and data collection

Northeast China generally consists of three provinces, including Liaoning, Jilin and Heilongjiang. In Liaoning province, the terrain is generally high in the north and low in the south. Mountains and hills are distributed on the east and west sides of Liaoning province. In Jilin province, the terrain is high in the southeast and low in the northwest. In Heilongjiang province, the terrain is higher in the northwest, northern and southeastern regions, and lower in the northeast and southwest. Northeast China has a temperate monsoon climate (Feng et al. 2018), where the average annual temperature is 6.6 °C, the annual relative humidity is 60%, the annual precipitation is 608.3 mm. In this study, long-term climatic data, including R_s, n, maximum and minimum air temperature (T_{max} and T_{min}, respectively), H_r, wind speed at 2 m height, and P, during 1997–2016 were collected from four stations located in Northeast China (Fig. 1). Extra-terrestrial solar radiation (R_a) calculated from geographic information and the day of the year (DOY) were also used for modeling. These data were provided and quality examined by the China Meteorological Administration. We further refined the data based on linear interpolation according to the rules: (1) missing measurements; (2) $T_{min} \geq T_{max}$; (3) $n > N$. Here N is the theoretical sunshine duration. Figure 2 shows the monthly variations in climatic variables. Table 1 shows the climatic conditions of the study region.

Gaussian exponential model

The GEM was proposed by Liu et al. (2014). The model is divided into three procedures. First, learning samples are

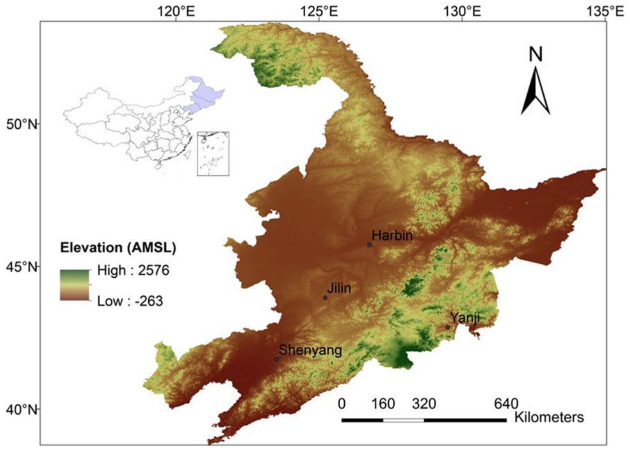

Fig. 1 The geographical distribution of the stations in Northeast China

clustered by the k-means algorithm as the most primitive allocation of samples. Second, the parameter estimates of the sample are calculated using the maximum likelihood estimation. Third, learning samples are regrouped according to the maximum posterior probability criterion. The model can be defined as follows:

$$f(n) = H_i \times \exp\left(-\frac{2(n - N_i)}{W_i^2}\right), i = 1, 2, \ldots, n \tag{1}$$

where H_i is the peak amplitude, N_i is the peak time position, and W_i is the half-width of the Gaussian wave.

Hybrid Gaussian exponential model and particle swarm optimization

The PSO algorithm has been widely used in model optimization, and has been proved its applicability (Yu et al. 2016; Zhu et al. 2020). In PSO, every particle has a fitness value. By calculating the fitness, the optimal output result is obtained.

In D-dimensional space, given a population with n particles (N_1, N_2, N_3, …, N_n). The position and velocity of every particle i are (n_{1i}, n_{2i}, n_{3i}, …, n_{ni}) and (v_{1i}, v_{2i}, v_{3i}, …, v_{ni}). Updating of the position and velocity can be expressed as:

$$N_{id} = N_{id} + N_{id}(d = 1, 2, \ldots, D, i = 1, 2, \ldots n) \tag{2}$$

$$V_{id}^{k+1} = \omega V_{id}^{k} + c_1 r_{1,i}^{k} (P_{id}^{k} - X_{id}^{k}) + c_2 r_{2,i}^{k} (P_{gd}^{k} - X_{id}^{k}) \tag{3}$$

where ω is the weight; k is the current iteration number; $c_1$and c_2 are the acceleration coefficients; $r_{1,i}{}^{k}$ and $r_{2,i}{}^{k}$ are the empirical parameters falling [0,1].

Although GEM has been proved to have high accuracy and computation speed (Jia et al. 2021), the PSO algorithm can further optimize the structure of GEM and improve the model accuracy.

M5 model tree

Quinlan (1992) first developed the M5 tree (M5T) model, which selects the expected standard deviation after scanning all the possible splits (Feng et al. 2019b). The

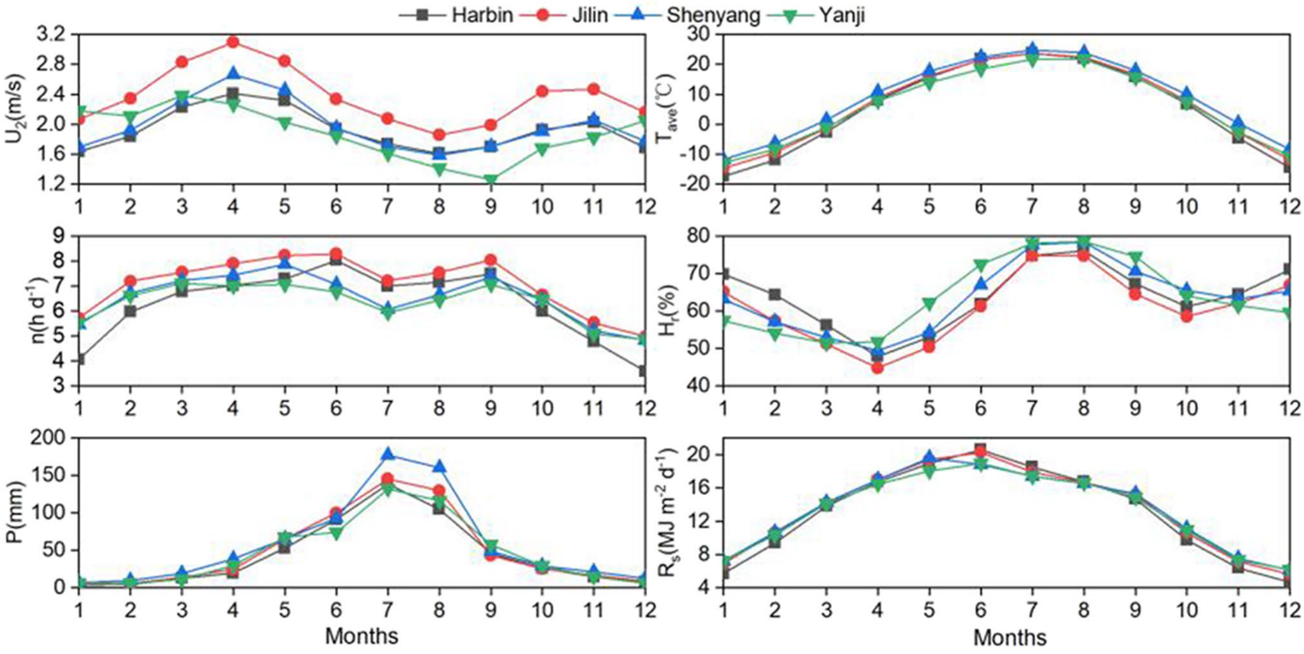

Fig. 2 Monthly variations of meteorological variables at the four stations in Northeast China

Table 1 Climatic conditions of the four stations in this study

Station	Longitude (°E)	Latitude (°N)	Variable	Max	Min	Average	S_x	C_v
Harbin	126.8	45.8	U_2 (m s^{-1})	7.6	0.0	1.9	0.9	0.5
			T_{ave} (°C)	30.9	−30.9	5.3	15.1	2.8
			n (h)	14.9	0.0	6.3	4.0	0.6
			H_r (%)	100	20	60	20	20
			P (mm d^{-1})	146.6	0.0	1.4	5.4	3.7
			R_s (MJ m^{-2} d^{-1})	33.3	0.0	13.0	7.2	0.6
Jilin	125.2	43.9	U_2 (m s^{-1})	10.5	0.0	2.4	1.2	0.5
			T_{ave} (°C)	30.4	−30.1	6.5	14.2	2.2
			n (h)	14.3	0.0	7.1	3.8	0.5
			H_r (%)	100	10	60	20	30
			P (mm d^{-1})	122.0	0.0	1.6	6.1	3.8
			R_s (MJ m^{-2} d^{-1})	39.6	0.0	13.5	7.4	0.5
Shenyang	123.5	41.7	U_2 (m s^{-1})	9.0	0.0	2.0	1.0	0.5
			T_{ave} (°C)	30.5	−26.8	8.6	13.3	1.5
			n (h)	13.9	0.0	6.5	3.9	0.6
			H_r (%)	100	10	60	20	30
			P (mm d^{-1})	145.7	0.0	1.9	7.6	4.0
			R_s (MJ m^{-2} d^{-1})	33.4	0.0	13.5	7.0	0.5
Yanji	129.5	42.9	U_2 (m s^{-1})	10.0	0.0	1.9	1.3	0.7
			T_{ave} (°C)	29.7	−23.7	6.0	12.8	2.1
			n (h)	14.0	0.0	6.3	3.7	0.6
			H_r (%)	100	10	60	20	30
			P (mm d^{-1})	124.6	0.0	1.5	5.5	3.6
			R_s (MJ m^{-2} d^{-1})	32.6	0.0	13.2	6.9	0.5

Max, Min, Average, S_x, and C_v denote the maximum, minimum, mean, standard deviation, and variation coefficient of each meteorological variable, respectively

procedure that makes up the model is divided into two parts. First, the data are divided into several subsets to create decision trees. The expected error of the subsets can be calculated by the model. The model accuracy can be defined as follows:

$$SDR = SD(Q) - \sum \frac{|Q_i|}{|Q|} SD(Q_i) \tag{4}$$

where *SD* and *SDR* are the standard deviations, *Q* is a set of samples that reach the target value, and Q_i is a subset of *Q*.

To improve the application efficiency of the model, it is necessary to traverse each node of the initial model tree through the pruning process to merge some subtrees and replace them with leaf nodes (Sattari et al. 2013). The detailed model procedure of the M5T model is described by Quinlan (1992).

Support vector machine

The SVM was first proposed by Vapink (1999). This model is considered the best theory for current small-sample statistical estimation and prediction learning (Belaid and Mellit 2016; Shamshirband et al. 2016). The model replaces traditional experience minimization with structural experience minimization, which can overcome many shortcomings of neural networks (Quej et al. 2017). The SVM function can be expressed as follows:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \kappa(x_i, y_i) + b \tag{5}$$

where $\kappa(x_i,x_j)$ is a higher-dimensional feature vector converted from the input vector x_i and x_j. y_i is the ordinate of the input vector, α_i is the weight of the input vector, and *b* is the bias.

Random forest model

The RF model was proposed by Breiman (2001). The model introduces random attribute selection during model training. The model extracts data based on randomness and difference, which can greatly improve decision accuracy. The procedures of the RF model are described by Buja et al. (2008).

Hargreaves–Samani model

The HS model only uses $T_{\max}$ and $T_{\min}$ data as inputs and is widely reported to have acceptable accuracy for R_s estimation. The model is as follows:

$$R_s = [C(T_{\max} - T_{\min})^{0.5}] \times R_a \tag{6}$$

where R_s is the global R_s (MJ m^{-2} d^{-1}), $T_{\max}$ and $T_{\min}$ are the $T_{\max}$ and $T_{\min}$, respectively (℃), C is the empirical coefficient, and R_a is the R_a (MJ m^{-2} d^{-1}).

Bristow–Campbell model

Bristow and Campbell (1984) developed the BC model, which only uses R_a and the diurnal temperature range (ΔT) as the input data. The model is defined as follows:

$$R_s = a[1 - \exp(-b\Delta T^c)] \times R_a \tag{7}$$

where ΔT is the ΔT (℃) and a, b, and c are empirical coefficients.

Model training and testing

Five input combinations of meteorological data were used to train the machine learning models. Details of the combinations are presented in Table 2. The dataset was divided into two parts, i.e., 1997–2011 and 2012–2016, for training and testing the machine learning models, respectively. The coefficients of the empirical models were locally calibrated at each station by the least square error method using the training data (data from 1997 to 2011). The model training/calibration and testing were performed in Matlab 2018a. The parameters of the machine models are presented in Table 3.

Statistical indicators

The root mean square error (RMSE), relative root mean square error (RRMSE), coefficient of determination (R^2), mean absolute error (MAE), and coefficient of efficiency (E_{ns}) were used to assess the R_s models (Feng et al. 2017b), as follows:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - X_i)^2} \tag{8}$$

$$RRMSE = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - X_i)^2}}{\overline{X}} \times 100\% \tag{9}$$

Table 2 Input combinations for training the machine learning models in this study

Input scenario	Model ID	Models	Model ID	Models	Model ID	Models	Model ID	Models	Model ID	Models	Inputs
		M5T		SVM		PSO-GEM		RF		GEM	
1	A	M5T1	E	SVM1	K	PSO-GEM1	P	RF1	U	GEM1	DOY, R_a
2	B	M5T2	G	SVM2	L	PSO-GEM2	Q	RF2	V	GEM2	DOY, R_a, $T_{\max}$, $T_{\min}$
3	C	M5T3	H	SVM3	M	PSO-GEM3	R	RF3	W	GEM3	DOY, $R_{a,}$ $T_{\max}$, $T_{\min}$, n
4	D	M5T4	I	SVM4	N	PSO-GEM4	S	RF4	X	GEM4	DOY, $R_{a,}$ n, $\ln(P+1)$
5	E	M5T5	J	SVM5	O	PSO-GEM5	T	RF5	Y	GEM5	DOY, $R_{a,}$ $T_{\max}$, $T_{\min}$, n, H_r, $\ln(P+1)$

Table 3 Parameters applied for different machine learning models in this study

Model	Key parameters
M5T model	Minimum leaf size = 10, minimum parent size = 20
SVM	Kernel function type = Gaussian function, gamma = 30, cost = 40
RF model	Maximum depth of the tree = 3, number of trees = 500
GEM	Kernel function type = Gaussian function, gamma = 2, cost = 10
PSO-GEM	Particle swarm number = 50, acceleration factor = 1.5, inertia weight = 0.5 Kernel function type = Gaussian function, gamma = 2, cost = 10

$$R^2 = \frac{\left[\sum_{i=1}^{m}(X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{\sum_{i=1}^{m}(X_i - \overline{X})^2 \sum_{i=1}^{m}(Y_i - \overline{Y})^2} \tag{10}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|Y_i - X_i| \tag{11}$$

$$E_{ns} = 1 - \frac{\sum_{i=1}^{m}(Y_i - X_i)^2}{\sum_{i=1}^{m}(X_i - \overline{X})^2} \tag{12}$$

where X_i and Y_i are the trained and estimated values, respectively, and $\overline{X}$ is the average value of X_i.

Owing to the excessive evaluation index, it is very difficult for a single evaluation index to compare different models. Therefore, the global performance indicator (GPI) was introduced to comprehensively evaluate the model simulation results (Despotovic et al. 2015), as follows:

$$GPI_i = \sum_{j=1}^{5} \alpha_j (g_j - y_{ij}) \tag{13}$$

where α_j is a coefficient that is equal to 1 for the RMSE, RRMSE, and MAE and equal to − 1 for E_{ns} and R^2; g_j represents the median of statistical indicator j, and y_{ij} represents the scaled value of the statistical indicator j. A higher GPI value indicates the better performance of the model.

Results and discussion

Results

Evaluation of the models on a daily basis

The statistical performance of the models at the four stations is presented in Table 4. At Harbin station, the PSO-GEM1 showed the highest accuracy under input

Table 4 Statistical performances of daily R_s of different models at the four stations. The best model in each station is marked in bold

Station	Indicators	M5T1	M5T2	M5T3	M5T4	M5T5	SVM1	SVM2	SVM3	SVM4	SVM5	PSO-GEM1	PSO-GEM2	PSO-GEM3	PSO-GEM4	PSO-GEM5	RF1	RF2	RF3	RF4	RF5	GEM1	GEM2	GEM3	GEM4	GEM5	HS	BC
Harbin	RMSE ($MJ\cdot m^{-2}\cdot d^{-1}$)	5.207	4.132	2.900	2.809	2.824	5.749	3.791	2.518	2.497	2.406	2.893	3.068	1.944	2.033	**1.719**	5.404	4.044	2.813	2.766	2.700	5.049	3.398	2.149	2.161	1.797	4.229	4.127
	RRMSE (%)	38.4	30.5	21.4	20.7	20.8	41.4	27.9	18.6	18.4	17.7	36.1	22.6	15.4	15.0	**12.7**	39.8	29.8	20.7	20.4	19.9	37.2	25.0	15.8	15.9	13.2	31.2	30.4
	R^2	0.516	0.705	0.904	0.914	0.911	0.424	0.746	0.919	0.923	0.929	0.569	0.842	0.952	0.948	**0.964**	0.483	0.721	0.913	0.919	0.924	0.541	0.803	0.941	0.942	0.960	0.689	0.705
	E_{ns}	0.505	0.688	0.847	0.856	0.854	0.397	0.738	0.884	0.886	0.894	0.563	0.828	0.931	0.925	**0.946**	0.467	0.702	0.856	0.860	0.867	0.535	0.789	0.916	0.915	0.941	0.674	0.689
	MAE ($MJ\cdot m^{-2}\cdot d^{-1}$)	4.015	3.096	2.237	2.164	2.187	4.718	2.791	1.892	1.869	1.798	3.742	2.305	1.456	1.525	**1.283**	4.110	3.067	2.178	2.149	2.108	3.886	2.524	1.604	1.632	1.353	3.270	3.187
Jilin	RMSE ($MJ\cdot m^{-2}\cdot d^{-1}$)	5.543	4.465	2.049	1.982	1.962	19.887	3.937	1.839	1.830	1.760	5.393	3.157	1.513	1.543	**1.245**	6.020	4.105	1.947	1.972	1.869	5.550	3.471	1.610	1.538	1.410	4.377	4.207
	RRMSE (%)	40.7	32.8	15.1	14.6	14.4	146.1	28.9	13.5	13.4	12.9	43.6	23.2	11.1	11.3	**9.1**	44.2	30.2	14.3	14.5	13.7	40.8	25.5	11.8	11.3	10.4	32.2	30.9
	R^2	0.474	0.661	0.929	0.934	0.933	0.147	0.734	0.942	0.943	0.947	0.503	0.835	0.961	0.959	**0.974**	0.384	0.712	0.935	0.934	0.941	0.473	0.795	0.956	0.960	0.966	0.672	0.696
	E_{ns}	0.472	0.658	0.928	0.933	0.934	0.105	0.734	0.942	0.943	0.947	0.501	0.829	0.961	0.959	**0.973**	0.378	0.711	0.935	0.933	0.940	0.471	0.793	0.956	0.959	0.966	0.671	0.696
	MAE ($MJ\cdot m^{-2}\cdot d^{-1}$)	4.125	3.062	1.427	1.377	1.358	15.663	2.714	1.249	1.256	1.183	4.030	2.211	1.013	1.072	**0.844**	4.462	2.888	1.337	1.392	1.271	4.147	2.403	1.077	1.067	0.945	3.248	3.069
Shenyang	RMSE ($MJ\cdot m^{-2}\cdot d^{-1}$)	5.697	4.407	2.619	2.539	2.484	5.994	3.983	2.274	2.387	2.119	5.355	3.227	1.876	2.010	**1.658**	5.919	4.283	2.465	2.526	2.347	5.586	3.565	2.019	2.088	1.754	4.557	4.479
	RRMSE (%)	40.1	31.0	18.4	17.9	17.5	42.2	28.0	16.0	16.8	15.5	37.7	22.7	13.2	14.1	**11.7**	41.6	30.1	17.3	17.8	16.5	39.3	25.1	14.2	14.7	14.1	32.1	31.5
	R^2	0.440	0.656	0.882	0.890	0.894	0.389	0.717	0.909	0.901	0.916	0.503	0.822	0.940	0.930	**0.953**	0.392	0.676	0.897	0.891	0.908	0.457	0.776	0.929	0.925	0.947	0.634	0.648
	E_{ns}	0.421	0.654	0.878	0.885	0.890	0.359	0.717	0.908	0.898	0.914	0.489	0.814	0.937	0.928	**0.951**	0.375	0.673	0.892	0.886	0.902	0.444	0.773	0.927	0.922	0.945	0.630	0.642
	MAE ($MJ\cdot m^{-2}\cdot d^{-1}$)	4.557	3.284	1.836	1.797	1.764	4.741	2.912	1.566	1.642	4.836	4.265	2.412	1.299	1.407	**1.147**	4.644	3.234	1.778	1.799	1.703	4.466	2.637	1.394	1.481	1.206	3.526	3.483
Yanji	RMSE ($MJ\cdot m^{-2}\cdot d^{-1}$)	5.402	3.849	1.746	1.675	1.743	6.879	3.454	1.509	3.521	1.434	5.030	2.809	1.226	1.259	**1.045**	5.571	3.767	1.686	1.671	1.682	5.209	3.118	1.274	1.319	1.149	4.063	3.885
	RRMSE (%)	39.3	28.0	12.7	12.2	12.7	50.0	25.1	11.0	25.6	10.4	36.6	20.4	8.9	9.2	**7.6**	40.5	27.4	12.3	12.2	12.2	37.9	22.7	9.3	9.6	8.4	29.6	28.3
	R^2	0.406	0.716	0.946	0.952	0.947	0.430	0.961	0.953	0.762	0.963	0.484	0.849	0.973	0.971	**0.981**	0.376	0.727	0.952	0.952	0.957	0.445	0.812	0.971	0.969	0.977	0.677	0.707
	E_{ns}	0.396	0.693	0.937	0.942	0.937	0.021	0.753	0.959	0.744	0.958	0.476	0.837	0.969	0.967	**0.977**	0.358	0.706	0.941	0.942	0.941	0.439	0.799	0.966	0.964	0.973	0.659	0.688
	MAE ($MJ\cdot m^{-2}\cdot d^{-1}$)	4.296	2.980	1.322	1.297	1.332	5.197	2.628	1.169	2.655	1.101	3.973	2.194	0.943	0.970	**0.801**	4.347	2.962	1.306	1.300	1.316	4.138	2.431	0.977	1.018	0.877	3.320	3.132

scenario 1 with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 2.893 MJ m^{-2} d^{-1}, 36.1%, 0.569, 0.563, and 3.742 MJ m^{-2} d^{-1}, respectively. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, considering the values of their evaluation indices. Under input scenario 3, the five machine learning models had higher accuracies than the models under input scenarios 1 and 2, with an RMSE value of less than 2.900 MJ m^{-2} d^{-1}, RRMSE value of less than 21.4%, R^2 value of greater than 0.952, E_{ns} value of greater than 0.931, and MAE of less than 2.237 MJ m^{-2} d^{-1}. This indicated that introducing climatic variables into the model training greatly improved the model performance. Among the models under input scenario 3, the PSO-GEM3 had the highest accuracy, considering the values of their evaluation indices. The PSO-GEM4 showed the highest accuracy under input scenario 4, considering the values of their evaluation indices. Under input scenario 5, the PSO-GEM5 showed the highest accuracy with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 1.719 MJ m^{-2} d^{-1}, 12.7%, 0.964, 0.946, and 1.283 MJ m^{-2} d^{-1}, respectively.

At Jilin station, the PSO-GEM1 showed the highest accuracy under input scenario 1 with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 5.393 MJ m^{-2} d^{-1}, 43.6%, 0.503, 0.501, and 4.030 MJ m^{-2} d^{-1}, respectively. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, considering the values of their evaluation indices. Under input scenario 3, the PSO-GEM3 had the highest accuracy, considering the values of their evaluation indices. Under input scenario 4, the PSO-GEM4 showed the highest accuracy. The five machine learning models under input scenario 5 showed the highest accuracies among the models under different input scenarios. The PSO-GEM5 showed the highest accuracy with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 1.245 MJ m^{-2} d^{-1}, 9.1%, 0.974, 0.973, and 0.844 MJ m^{-2} d^{-1}, respectively.

At Shenyang station, the PSO-GEM1 showed the highest accuracy under input scenario 1 with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 5.355 MJ m^{-2} d^{-1}, 37.7%, 0.503, 0.489, and 4.265 MJ m^{-2} d^{-1}, respectively. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, considering the values of their evaluation indices. Under input scenario 3, all the machine learning models had higher accuracies than the models under input scenarios 1 and 2, with an RMSE value of less than 2.619 MJ m^{-2} d^{-1}, RRMSE of less than 18.4%, R^2 of over 0.882, E_{ns} of over 0.878, and MAE of less than 1.836 MJ m^{-2} d^{-1}. The PSO-GEM3 had the highest accuracy. The PSO-GEM4 showed the highest accuracy under input scenario 4, considering the values of their evaluation indices. Under input scenario 5, the PSO-GEM5 showed the highest accuracy with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 1.658 MJ m^{-2} d^{-1}, 11.7%, 0.953, 0.951, and 1.147 MJ m^{-2} d^{-1}, respectively.

At Yanji station, the PSO-GEM1 showed the highest accuracy under input scenario 1 with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 5.030 MJ m^{-2} d^{-1}, 36.6%, 0.484, 0.476, and 3.973 MJ m^{-2} d^{-1}, respectively. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, considering the values of their evaluation indices. Under input scenario 3, the five models had higher accuracies than the models under input scenarios 1 and 2, with an RMSE of less than 1.746 MJ m^{-2} d^{-1}, RRMSE of less than 12.7%, R^2 of over 0.946, E_{ns} of over 0.937, and MAE of less than 1.322 MJ m^{-2} d^{-1}. The PSO-GEM3 had the highest accuracy. The PSO-GEM4 showed the highest accuracy under input scenario 4, considering the values of their evaluation indices. Under input scenario 5, the PSO-GEM5 showed the highest accuracy with RMSE, RRMSE, R^2, E_{ns}, and MAE values of 1.045 MJ m^{-2} d^{-1}, 7.6%, 0.981, 0.977, and 0.801 MJ m^{-2} d^{-1}, respectively.

As for the empirical models, the HS and BC models showed lower accuracies compared with those of the machine learning models with the same inputs (input scenario 3), with an RMSE of 3.885–4.557 MJ m^{-2} d^{-1}, R^2 of 0.634–0.707, RRMSE of 28.3–32.2%, E_{ns} of 0.630–0.696, and MAE of 3.069–3.526 MJ m^{-2} d^{-1}. The accuracy of the machine learning models considering n was significantly higher than that of the models without n input, with the RMSE reduced by 44.3–79.9%, RRMSE reduced by 44.2–91.2%, MAE reduced by 40.2–80.6%, R^2 increased by 67.7–95.6%, and E_{ns} increased by 67.4–124.9%.

The boxplots of the statistical indicators of daily R_s for different models in the study area are presented in Fig. 3. Under input scenario 1, the five machine learning models showed low prediction accuracies for the whole region, with average RMSE, RRMSE, MAE, and E_{ns} values of 4.668–9.627 MJ m^{-2} d^{-1}, 38.8–69.9%, 4.002–7.579 MJ m^{-2} d^{-1}, and 0.220–0.507, respectively. The PSO-GEM1 showed the highest accuracy among the five models. Under input scenario 2, the PSO-GEM2 was the best model, considering the values of their evaluation indices. The five models under input scenario 3 showed higher prediction accuracies than the models under input scenarios 1 and 2, which did not consider climatic variables as inputs. The PSO-GEM3 showed the best results, considering the values of their evaluation indices. Under input scenario 4, the PSO-GEM4 showed the highest accuracy, considering the values of their evaluation indices. Under input scenario 5, the PSO-GEM5 showed the highest accuracy with average RMSE, RRMSE, MAE, and E_{ns} values of 1.417 MJ m^{-2} d^{-1}, 10.26%, 1.019 MJ m^{-2} d^{-1}, and 0.962, respectively. The HS and BC models showed much lower prediction accuracies compared with those of the machine learning models, with average RMSE, RRMSE, MAE, and E_{ns} values of 4.306 MJ m^{-2} d^{-1} and 4.174 MJ m^{-2} d^{-1}, 31.23% and

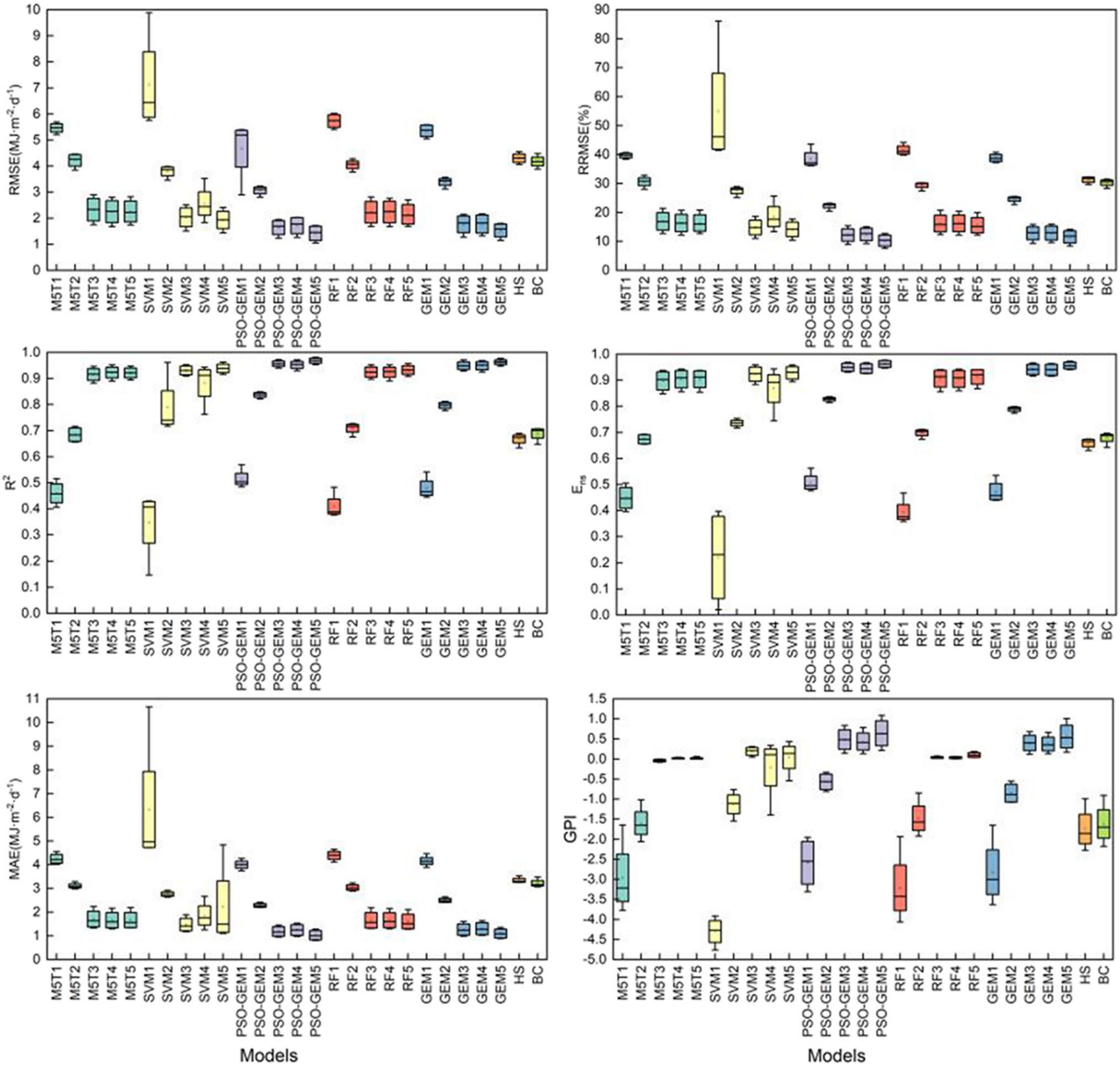

Fig. 3 Boxplots of the statistical indicators of daily R_s for different models

30.26%, 3.341 MJ m^{-2} d^{-1} and 3.218 MJ m^{-2} d^{-1}, and 0.658 and 0.679, respectively.

The GPI values of the different models at the four stations are presented in Fig. 4. The SVM1, M5T1, GEM1, RF1, and PSO-GEM1 models under input scenario 1 showed the lowest prediction accuracies compared with those of models under other input scenarios, with average GPI values of – 3.915, – 3.101, – 2.883, – 3.357, and – 2.163, respectively. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, followed by the GEM2, SVM2, RF2, and M5T2 models. Under input scenario 3, the PSO-GEM3 showed the highest accuracy, considering the values of their evaluation indices. The PSO-GEM4 was the best model under input scenario 4, followed by the GEM4, RF4, M5T4, and SVM4 models with average GPI values of 0.434, 0.375, 0.033, 0.019, and – 0.211, respectively. Under input scenario 5, the PSO-GEM5 and GEM5 showed much higher accuracies with average GPI values of 0.641 and 0.560, respectively. The accuracies of the HS and BC models were higher than those of the M5T1, SVM1, GEM1, RF1, and PSO-GEM1 models without climatic inputs, with average GPI values of – 1.745 and – 1.622, respectively. Relatively good estimates and high accuracies could be obtained from models with at least the DOY, R_a, and n as inputs, including models

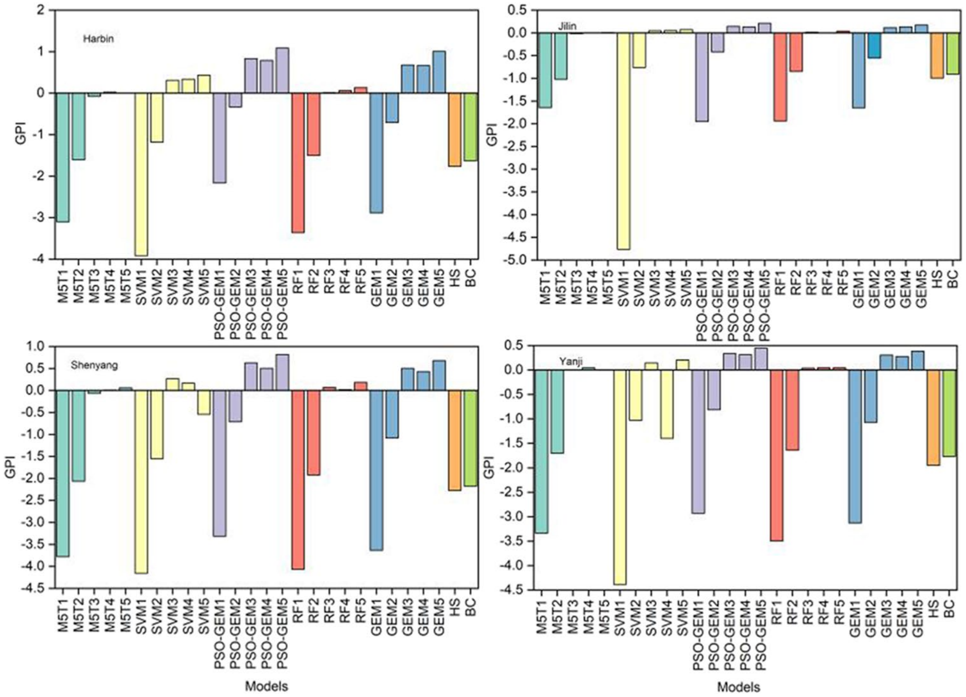

Fig. 4 GPI values of daily R_s of different models at the four stations in Northeast China

under input scenarios 3, 4, and 5. These results further confirm that *n* is the most important variable for estimating R_s.

Evaluation of the models on a monthly basis

The accuracy index of monthly R_s of different models in different stations is presented in Table 5. As shown in Table 5, in Harbin station, the PSO-GEM1 showed the highest accuracy under input scenario 1, with RMSE, RRMSE, R^2, E_{ns}, and MAE of 0.878 MJ m^{-2} d^{-1}, 13.5%, 0.984, 0.943, and 0.803 MJ m^{-2} d^{-1}, respectively. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, followed by the GEM2, considering the values of their evaluation indices. Under input scenario 3, the five models had higher accuracy than the models under input scenario 1 and scenario 2, with RMSE less than 0.825 MJ m^{-2} d^{-1}, RRMSE

Table 5 Statistical performances of monthly R_s of different models at the four stations. The best model in each station is marked in bold

Stations	Evaluation index	M5T1	M5T2	M5T3	M5T4	M5T5	SVM1	SVM2	SVM3	SVM4	SVM5	PSO-GEM1	PSO-GEM2	PSO-GEM3	PSO-GEM4	PSO-GEM5	RF1	RF2	RF3	RF4	RF5	GEM1	GEM2	GEM3	GEM4	GEM5	HS	BC
Harbin	RMSE	1.155	1.145	0.825	0.802	0.790	1.941	0.848	0.593	0.592	0.593	0.878	0.692	0.601	0.506	0.575	1.16	1.175	0.783	0.802	0.739	0.885	0.814	0.655	0.594	**0.456**	1.186	1.193
	RRMSE	13.5	13.4	8.5	8.3	8.2	14.3	13.3	8.3	8.6	8.3	13.5	11.1	7.5	7.2	**7.0**	12.6	14.7	8.2	8.3	7.8	12.5	11.0	7.5	7.8	7.1	8.8	8.8
	R^2	0.972	0.991	0.999	0.999	0.999	0.947	0.992	0.999	0.999	0.999	0.984	0.997	**1**	0.999	**1**	0.971	0.988	0.999	0.999	0.999	0.984	0.996	0.999	0.999	**1**	0.977	0.976
	E_{NS}	0.953	0.954	0.984	0.987	0.989	0.868	0.925	0.972	0.968	0.988	0.943	0.953	0.984	0.987	**0.989**	0.935	0.952	0.989	0.987	0.994	0.923	0.937	0.973	0.985	0.988	0.951	0.95
	MAE	1.523	1.502	0.545	0.525	0.513	1.586	1.552	0.621	0.664	0.616	0.803	0.702	0.467	0.463	0.499	1.236	1.113	0.608	0.627	0.603	1.494	1.371	0.568	0.548	**0.411**	0.955	0.951
Jilin	RMSE	0.495	0.542	0.41	0.359	0.407	1.078	0.772	0.307	0.317	0.312	0.69	0.525	0.256	0.245	**0.197**	0.685	0.547	0.394	0.36	0.377	0.692	0.648	0.264	0.221	0.242	0.672	0.643
	RRMSE	5.6	4.0	3.0	2.6	3.0	7.9	3.5	2.3	2.3	2.3	3.6	2.4	1.9	1.8	**1.5**	5.0	3.3	2.9	2.6	2.8	5.6	2.6	1.9	1.6	1.8	5.4	4.7
	R^2	0.984	0.983	0.995	0.996	0.994	0.956	0.984	0.997	0.997	0.999	0.954	0.907	0.998	0.998	**0.999**	0.918	0.976	0.995	0.996	0.995	0.844	0.897	0.998	0.998	0.998	0.872	0.991
	E_{NS}	0.871	0.989	0.994	0.995	0.997	0.856	0.892	0.996	0.996	0.996	0.891	0.926	0.998	0.998	**0.999**	0.902	0.973	0.994	0.995	0.997	0.891	0.915	0.997	0.998	0.998	0.971	0.984
	MAE	0.593	0.417	0.321	0.26	0.304	0.743	0.363	0.217	0.218	0.211	0.364	0.261	0.176	0.164	**0.137**	0.526	0.390	0.279	0.246	0.249	0.381	0.276	0.189	0.149	0.159	0.520	0.559
Shenyang	RMSE	1.225	0.828	0.606	0.633	0.612	1.415	0.705	0.376	0.398	0.318	0.932	0.579	0.340	0.341	**0.313**	1.227	0.919	0.617	0.629	0.614	0.962	0.687	0.347	0.32	0.351	0.856	0.962
	RRMSE	8.6	6.4	4.3	4.5	4.3	10.0	5.9	3.1	3.0	2.9	6.6	4.7	2.4	2.8	**2.2**	8.6	6.4	4.3	4.6	4.5	6.8	4.7	2.4	3.0	2.5	6.0	6.8
	R^2	0.981	0.987	0.995	0.995	0.996	0.953	0.973	0.997	0.996	0.998	0.971	0.986	0.999	0.998	**0.999**	0.972	0.988	0.996	0.994	0.996	0.979	0.986	0.999	0.998	0.998	0.939	0.926
	E_{NS}	0.938	0.964	0.985	0.983	0.985	0.917	0.993	0.994	0.993	0.996	0.964	0.984	0.995	0.994	**0.996**	0.938	0.964	0.984	0.983	0.983	0.962	0.984	0.995	0.995	0.995	0.970	0.962
	MAE	0.996	0.625	0.477	0.497	0.478	1.079	0.497	0.305	0.309	0.305	0.750	0.496	0.276	0.301	**0.256**	1.005	0.785	0.513	0.51	0.512	0.788	0.390	0.268	0.293	0.273	0.680	0.790
Yanji	RMSE	1.215	1.149	0.694	0.695	0.693	0.800	0.745	0.552	0.607	0.506	0.795	0.661	0.455	0.454	**0.463**	0.976	0.787	0.536	0.523	0.533	0.989	0.786	0.467	0.475	0.447	1.305	1.239
	RRMSE	10.7	8.4	5.1	5.2	5.3	8.8	7.4	4.0	4.8	4.1	8.1	4.8	3.6	3.3	**3.0**	9.1	7.9	5.4	5.3	5.4	9.0	6.7	3.4	3.6	3.3	9.5	9.0
	R^2	0.915	0.992	0.998	0.998	0.999	0.951	0.974	0.998	0.998	0.998	0.991	0.998	0.999	0.999	**1**	0.978	0.993	0.998	0.998	0.999	0.992	0.977	0.999	0.999	0.999	0.979	0.984
	E_{NS}	0.929	0.936	0.977	0.979	0.980	0.949	0.973	0.985	0.989	0.989	0.977	0.979	0.991	**0.992**	**0.992**	0.934	0.943	0.974	0.975	0.976	0.947	0.97	0.989	0.988	**0.992**	0.917	0.926
	MAE	1.198	1.039	0.649	0.648	0.642	0.701	0.608	0.514	0.526	0.522	0.627	0.603	0.408	0.409	**0.384**	0.864	0.78	0.523	0.533	0.528	0.804	0.713	0.436	0.438	0.407	1.133	1.092

less than 8.5%, R^2 over than 0.999, E_{ns} over than 0.972, MAE less than 0.621 MJ m^{-2} d^{-1}. The PSO-GEM3 had the highest accuracy. The PSO-GEM4 showed the best precision under input scenario 4, considering the values of their evaluation indices. Under input scenario 5, the PSO-GEM5 and GEM5 showed much higher accuracy among the five models, considering the values of their evaluation indices. HS and BC models showed much poorer prediction accuracy with RMSE of 1.186 and 1.193 MJ m^{-2} d^{-1}, with RRMSE of 8.8% and 8.8%, R^2 of 0.977 and 0.976, MAE of 0.955 and 0.951 MJ m^{-2} d^{-1}, and E_{ns} of 0.951 and 0.950, respectively.

In Jilin station, the PSO-ELM1 showed the highest accuracy under input scenario 1, with RMSE, RRMSE, R^2, E_{ns}, and MAE of 0.932 MJ m^{-2} d^{-1}, 6.6%, 0.971, 0.964, and 0.750 MJ m^{-2} d^{-1}. Under input scenario 2, the PSO-ELM2 had the best precision, considering the values of their evaluation indices. Under input scenario 3, the PSO-ELM3 had the highest accuracy, considering the values of their evaluation indices. Under input scenario 4, the PSO-ELM4 showed the best precision, with RMSE, RRMSE, R^2, E_{ns}, and MAE of 0.341 MJ m^{-2} d^{-1}, 2.8%, 0.998, 0.994, and 0.301 MJ m^{-2} d^{-1}. The five models under the input scenario 5 showed the

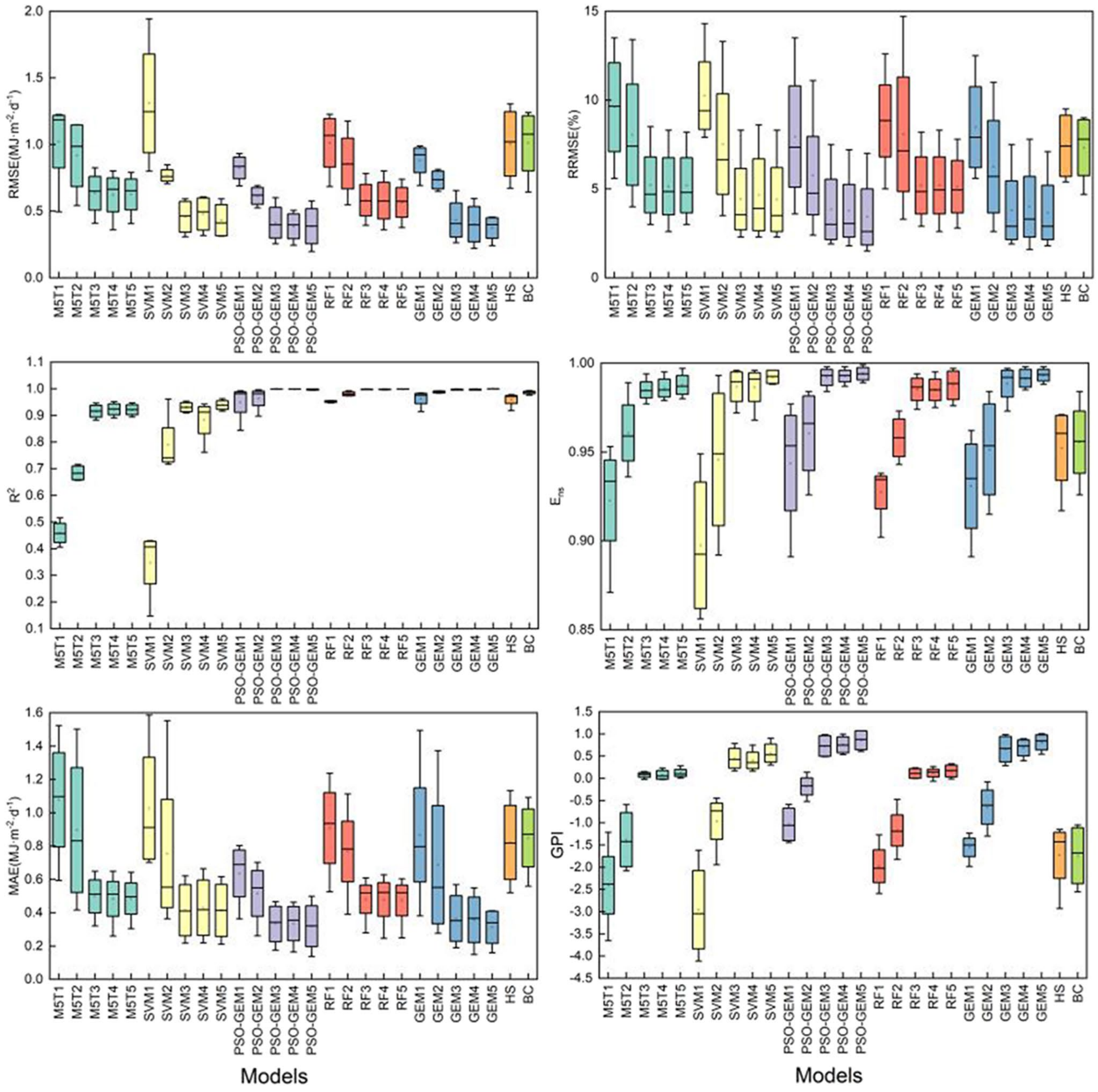

Fig. 5 Boxplots of the statistical indicators monthly R_s for different models

highest accuracy among the models under other inputs. The PSO-ELM5 showed the highest accuracy, followed by the GEM5, with RMSE, RRMSE, R^2, E_{ns}, and MAE of 0.197 and 0.242 MJ m^{-2} d^{-1}, 1.5% and 1.8%, 0.999 and 0.998, 0.999 and 0.998, and 0.137 and 0.159 MJ m^{-2} d^{-1}.

In Shenyang station, under input scenario 1, the PSO-GEM1 showed the highest accuracy, followed by GEM1, with RMSE, RRMSE, R^2, E_{ns}, and MAE of 0.932 and 0.962 MJ m^{-2} d^{-1}, 6.6% and 6.8%, 0.971 and 0.979, 0.964 and 0.962, and 0.750 and 0.788 MJ m^{-2} d^{-1}. Under input scenario 2, the PSO-GEM2 showed the best precision, considering the values of their evaluation indices. Under input scenario 3, the PSO-GEM3 and GEM3 model showed higher accuracy, considering the values of their evaluation indices. The PSO-GEM4, GEM4 and SVM4 models showed better precision under input scenario 4, considering the values of their evaluation indices. Under input scenario 5, PSO-GEM5 showed the highest accuracy, followed by GEM5, with RMSE, RRMSE, R^2, E_{ns}, and MAE of 0.313 and 0.351 MJ m^{-2} d^{-1}, 2.2% and 2.5%, 0.999 and 0.998, 0.996 and 0.995, and 0.256 and 0.273 MJ m^{-2} d^{-1}, respectively.

In Yanji station, the PSO-GEM1 showed the highest accuracy under input scenario 1, considering the values of their evaluation indices. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, with RMSE, RRMSE, R^2, E_{ns}, and MAE of 0.661 MJ m^{-2} d^{-1}, 4.8%, 0.998, 0.979, and 0.603 MJ m^{-2} d^{-1}. Under input scenario 3, the five models had higher accuracy than the models under input scenario 1 and scenario 2, with RMSE less than 0.694 MJ m^{-2} d^{-1}, RRMSE less than 5.4%, R^2 over than 0.998, E_{ns} over than 0.974, and MAE less than 0.649 MJ m^{-2} d^{-1}. The PSO-GEM3 had the highest accuracy, considering the values of their evaluation indices. The PSO-GEM4 model showed the best precision under input scenario 4, considering the values of their evaluation indices. Under input scenario 5, the PSO-GEM5 showed the highest accuracy, followed by the GEM5, considering the values of their evaluation indices.

The boxplots of the statistical indicators of monthly R_s for different models in the study area are presented in Fig. 5. Under input scenario 1, the five models showed lower prediction accuracy in the whole studied area, with RMSE, RRMSE, MAE, E_{ns} of 0.824–1.308 MJ m^{-2} d^{-1}, 8.0–10.3%, 0.636–1.077 MJ m^{-2} d^{-1}, 0.898–0.944, respectively. The PSO-GEM1 showed the highest accuracy among the five models. Under input scenario 2, the PSO-GEM2 model was the best, considering the values of their evaluation indices. The five models under the input scenario 3 showed higher accuracy than the models under the input scenarios 1–2. The PSO-GEM3 showed the best precision, followed by the GEM3, considering the values of their evaluation indices. Under input scenario 4, the PSO-GEM4 showed the highest accuracy. Under input scenario 5, the PSO-GEM5 and GEM5 showed higher accuracy, considering the values of their evaluation indices. HS model and BC model showed poorer prediction accuracy with RMSE of 1.005 and 1.009 MJ m^{-2} d^{-1}, with RRMSE of 7.3% and 7.4%, MAE of 0.822 and 0.848 MJ m^{-2} d^{-1}, and E_{ns} of 0.952 and 0.955, respectively.

GPI of monthly R_s of different models in the whole studied area is presented in Fig. 6. As shown in Fig. 6, SVM1, M5T1, RF1, GEM1 and PSO-GEM1 models under input

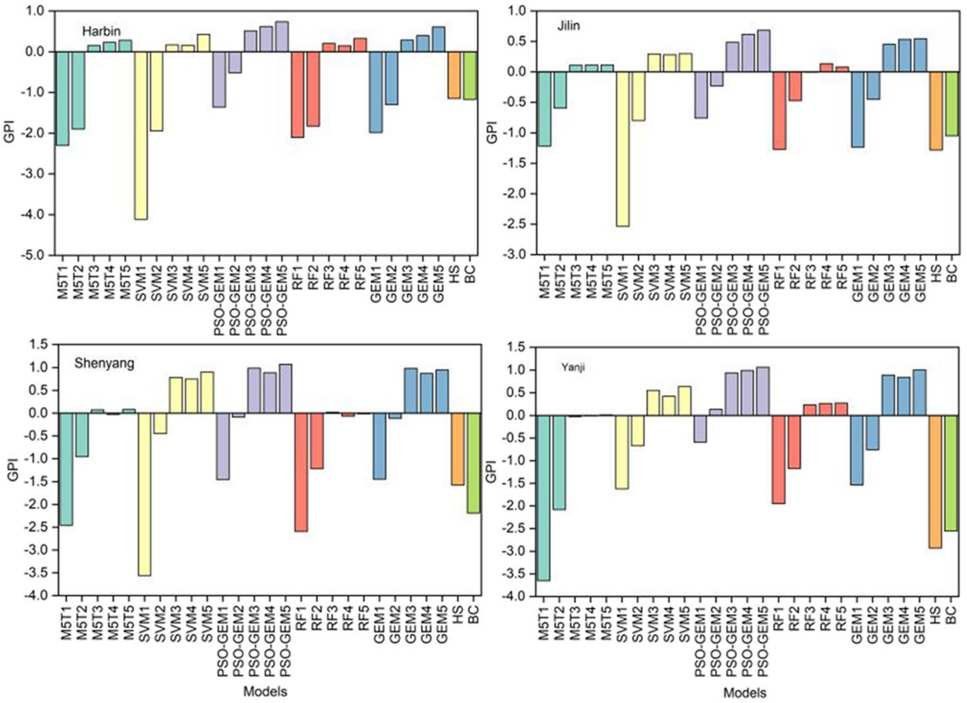

Fig. 6 GPI values of monthly R_s of different models at the four stations in Northeast China

scenario 1 showed the lowest prediction accuracy, with average GPI of – 2.957, – 2.406, – 1.979, – 1.553, and – 1.034, respectively. Under input scenario 2, the PSO-GEM2 showed the highest accuracy, considering the values of their evaluation indices. Under input scenario 3, the PSO-GEM3 showed the best precision, followed by the GEM3 model, considering the values of their evaluation indices. The PSO-GEM4 was the best model under input scenario 4, followed by the GEM4 model, with average GPI of 0.755 and 0.686, respectively. Under input scenario 5, the PSO-GEM5 showed the best precision, with average GPI of 0.855, respectively. The accuracy of HS and BC models was high than M5T1, SVM1, and RF1 models, with average GPI of – 1.734 and – 1.742 respectively. Machine learning models with complete data inputs had the highest precision. Meanwhile, the models which considered n, $T_{\max}$ and $T_{\min}$, n and P showed similar precision compared to the models as for seven-inputs. The models only considered DOY and Ra showed the lowest prediction accuracy, with GPI of – 4.114 to – 0.588. The accuracy of the monthly R_s models which considered DOY, R_a, $T_{\max}$, and $T_{\min}$ was higher than the models for two inputs, with GPI increased by 17.5–29.4%. The increase in accuracy was not significant. In the calculation of monthly R_s, sunshine duration was the most significant variable in the studied area.

Discussion

The PSO can further improve the accuracy of GEM, as PSO can improve the iteration rate of GEM and avoid the initialized

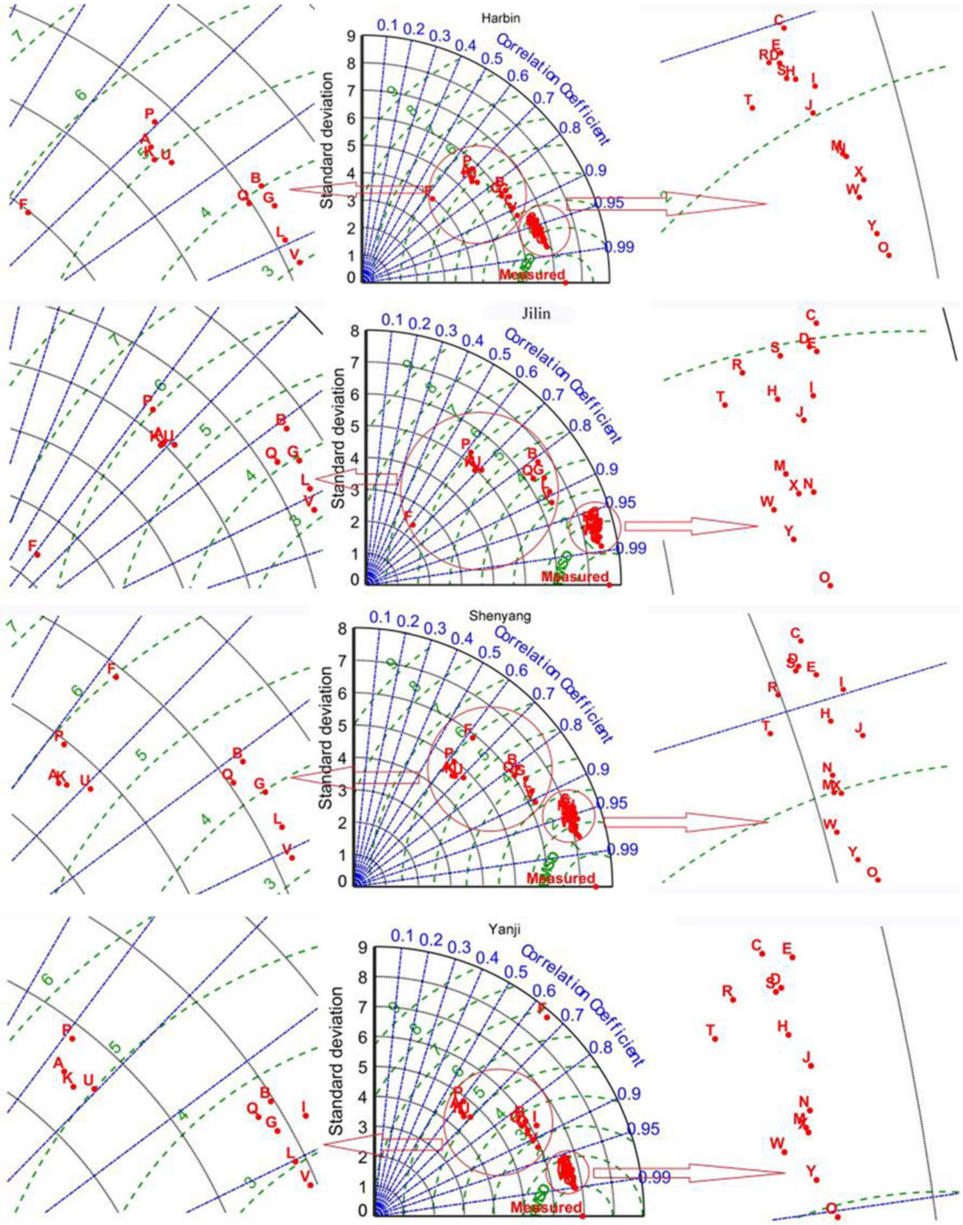

Fig. 7 Taylor diagrams of daily R_s of different machine learning models at different stations

weights. Under different input scenarios, the PSO-GEM showed the highest accuracy. The GEM can better reflect the nonlinear relationship between radiation and meteorological factors by calculating the Gaussian exponents. The accuracy of GEM has been proven (Lesser et al. 2011; Jia et al. 2021.). Wu et al. (2021) showed that the PSO can improve the accuracy of the extreme learning machine models and have better ability in optimizing the parameters. It confirmed generalizability and robustness of PSO-GEM. Machine learning models generally had a higher accuracy than the HS and BC models when climatic variables were included as inputs. The machine learning models that considered only the DOY and R_a showed the lowest accuracies at the four stations, especially the SVM1 and RF1 models. Fan et al. (2019) showed that in China, the SVM and RF models had worse rankings, which agrees with our conclusion.

To further confirm the reliability of PSO-GEM for R_s estimation, the Taylor diagrams of different models at four stations were analyzed. The standard deviation and correlation coefficient of the statistical indicators by the models over the stations are listed in Figs. 7 and 8. It was clear that PSO-GEM5 at different stations have the lowest standard deviation, the lowest mean square error and the highest correlation coefficient with

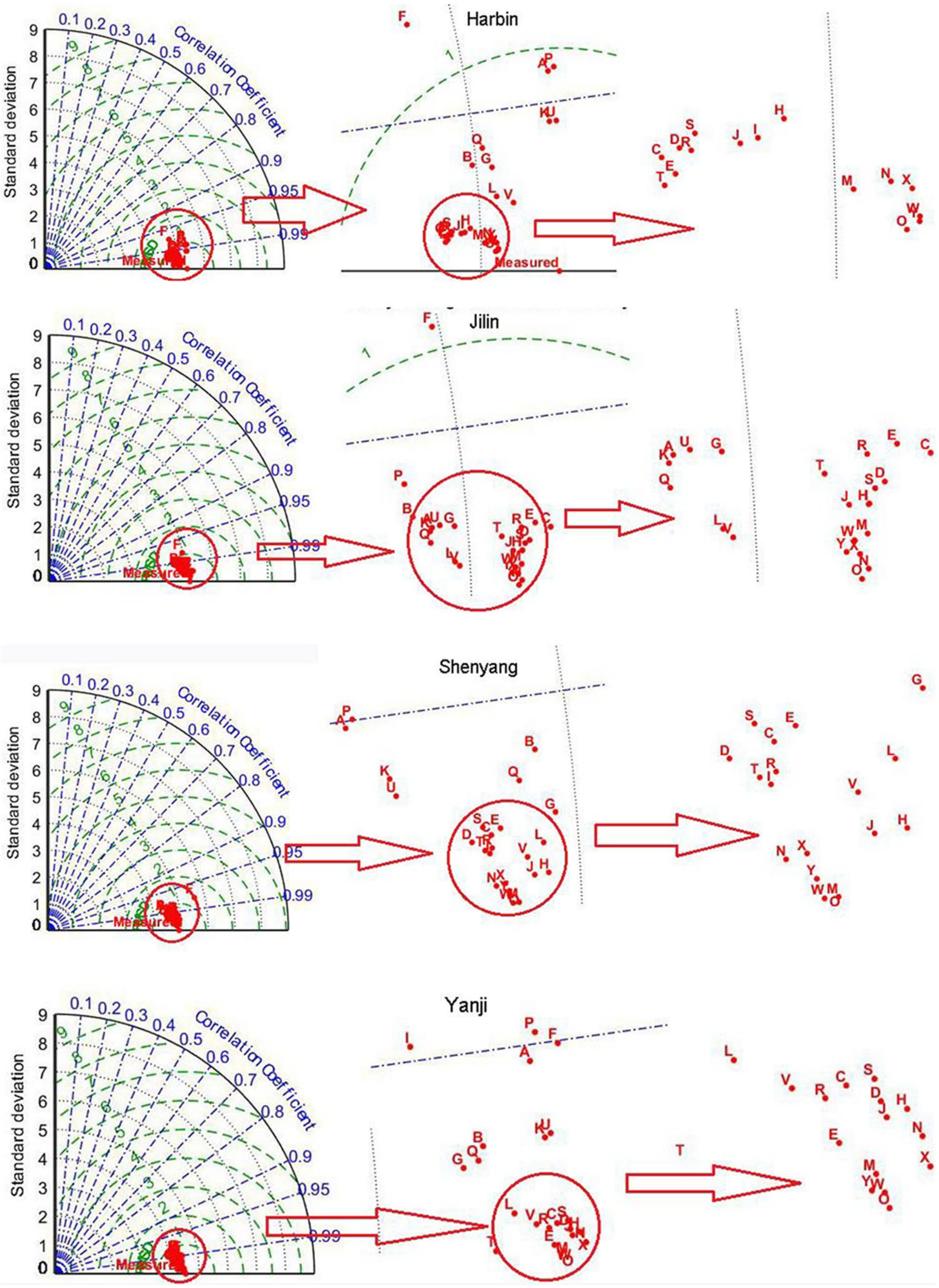

Fig. 8 Taylor diagrams of monthly R_s of different machine learning models at different stations

the standard values. These results further confirmed the performance of PSO-GEM5 at different stations in Northeast China.

The results of this study showed that the models with complete inputs had the highest accuracy. This indicated that the effect of each meteorological factor on R_s estimation was positive. However, the models with T_{max} and T_{min} as inputs showed lower accuracy, especially the HS and BC models. The models considering n (input scenarios 3, 4, and 5) showed a much higher accuracy, which revealed that n is the most important factor affecting R_s estimation in Northeast China. Mecibah et al. (2014) investigated the performance of different R_s models and found that the accuracy of models with n was much higher than that of models with air temperature. The same conclusion was also reported by Zhang et al. (2018) because the magnitude of n directly affects the R_s reaching the surface of the earth. The amount of solar radiation reaching the Earth's surface is closely related to sunshine duration. Clouds and the weather patterns are also the most important atmospheric phenomena limiting solar radiation on the Earth's surface. These are the main reasons for the higher accuracy of the models considering sunshine duration and precipitation. The solar radiation reaching the Earth's surface is absorbed by the atmosphere or emitted into the air in the form of long-wave radiation. The long-wave radiation absorbed by the atmosphere will increase the temperature. Thus, the temperature is also one of the important factors affecting solar radiation. But there are many factors affect the atmospheric temperature, the relationship between solar radiation and temperature does not correspond exactly. It is why the accuracy of the temperature-based models is lower than the sunshine-duration-based models.

The PSO-GEM can be recommended to estimate R_s in Northeast China. The proposed model can provide scientific support for evapotranspiration estimation, agricultural irrigation management and solar energy development. In this study, we considered a simple data set assignment for training machine learning models. K-fold cross-validation is an efficient training method recommended for training models (Shiri et al. 2015). In future research, we can combine PSO-GEM and K-fold cross-validation to further improve the accuracy of R_s estimation.

Conclusions

Five machine models with five groups of input parameters and two empirical models were evaluated for R_s prediction using meteorological data from four stations in Northeast China. The PSO-GEM with full climatic data as inputs showed the highest accuracy with RMSE, RRMSE, MAE, and E_{ns} values of 1.416 MJ m^{-2} d^{-1}, 10.27%, 1.018 MJ m^{-2} d^{-1}, and 0.962, respectively. The PSO-GEM showed the highest accuracy under other input scenarios. n is the most influential factor affecting R_s estimation by machine learning models.

Overall, the PSO-GEM5 is recommended for estimating R_s in Northeast China when all the meteorological variables are available. The PSO-GEM3 is recommended when only n and air temperature data are accessible. The PSO-GEM4 and GEM4 are recommended only when sunshine data and P data are available.

Acknowledgements We wish to thank the China Meteorological Administration for providing the data for this study.

Author contribution YJ analyzed and calculated the data and wrote the manuscript. HW got the data of the manuscript. PL, YS, and FW calculated the data and made the tables of the manuscript. SH made the figures for the manuscript. All authors read and approved the final manuscript.

Funding The research was funded by the Scientific Research Program of Higher Education in Hebei Province (QN2021227), the Water Conservancy Research and Extension Project of Hebei Province (2020–64), the Hebei Province Innovation Ability Promotion Plan, Soft Science Research (20557682D), the Colleges and Universities in Hebei Province Science and Technology Research (ZD2020348), and the Doctoral Scientific Research Foundation of Hebei University of Water Resources and Electric Engineering (SYBJ1902).

Data availability The data that support the findings of this study are available from National Meteorological Science Data Center (https://data.CMA.cn/) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of National Meteorological Science Data Center (https://data.CMAcma.cn/).

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Annandale J, Jovanovic N, Benade N, Allen R (2002) Sofware for missing data error analysis of Penman-Monteith reference evapotranspiration. Irrig Sci 21(2):57–67

Antonopoulos VZ, Papamichail DM, Aschonitis VG, Antonopoulos AV (2019) Solar radiation estimation methods using ANN and empirical models. Comput Electron Agric 160:160–167

Bailek N, Bouchouicha K, Al-Mostafa Z, El-Shimy M, Aoun N, Slimani A, Al-Shehri S (2018) A new empirical model for forecasting the diffuse solar radiation over Sahara in the Algerian Big South. Renew Energy 117:530–537

Belaid A, Mellit A (2016) Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate. Energy Convers Manag 118:105–118

Besharat F, Dehghan AA, Faghih AR (2013) Empirical models for estimating global solar radiation: a review and case study. Renew Sustain Energy Rev 21:798–821

Breiman L (2001) Random forests. Mach Learn 45:5–32

Bristow KL, Campbell GS (1984) On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric for Meteorol 31(2):159–166

Bueno CL, Mateo CC, Justo JS, Sanz SS (2019) Machine learning regressors for solar radiation estimation from satellite data. Sol Energy 183:768–775

Buja A, Swayne DF, Littman ML, Dean N, Hofmann H, Chen L (2008) Data visualization with multidimensional scaling. J Comput Graph Stat 17(2):444–472

Chen JL, Liu HB, Wu W, Xie DT (2011) Estimation of monthly solar radiation from measured temperatures using support vector machines – a case study. Renew Energy 36:413–420

Chukwujindu NS (2017) A comprehensive review of empirical models for estimating global solar radiation in Africa. Renew Sustain Energy Rev 78:955–995

Citakoglu H (2015) Comparison of artificial intelligence techniques via empirical equations for prediction of solar radiation. Comput Electron Agric 118:28–37

Demircan C, Bayrakçı HC, Keçebaş A (2020) Machine learning-based improvement of empiric models for an accurate estimating process of global solar radiation. Sustain Energy Technol Assess 37:100574

Desideri U, Zepparelli F, Morettini V, Garroni E (2013) Comparative analysis of concentrating solar power and photovoltaic technologies: technical and environmental evaluations. Appl Energy 102:765–784

Elias CL, Calapez AR, Almeida SFP, Chessman B, Simoes N, Feio MJ (2016) Predicting reference conditions for river bioassessment by incorporating boosted trees in the environmental filters method. Ecol Ind 69:239–251

Emamgolizadeh S, Bateni SM, Shahsavani D, Ashrafi T, Gorbani H (2015) Estimation of soil cation exchange capacity using genetic expression programming (GEP) and multivariate adaptive regression splines (MARS). J Hydrol 529(3):1590–1600

Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, Lu X, Xiang Y (2018) Comparison of support vector machine and extreme gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Convers Manag 164:102–111

Fan JL, Wu LF, Zhang FC, Cai HJ, Zeng WZ, Wang XK, Zou HY (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

Feng Y, Cui NB, Gong DZ, Zhang QW, Zhao L (2017) Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modeling. Agric Water Manag 193:163–173

Feng Y, Jia Y, Cui N, Zhao L, Li C, Gong D (2017) Calibration of Hargreaves model for reference evapotranspiration estimation in Sichuan basin of southwest China. Agriculture Water Manag 181:1–9

Feng Y, Jia Y, Zhang Q, Gong D, Cui N (2018) National-scale assessment of pan evaporation models across different climatic zones of China. J Hydrol 564:314–328

Feng Y, Cui N, Chen Y, Gong D, Hu X (2019) Development of data-driven models for prediction of daily global horizontal irradiance in northwest China. J Clean Prod 223:136–146

Feng Y, Cui N, Hao W, Gao L, Gong D (2019) Estimation of soil temperature from meteorological data using different machine learning models. Geoderma 338:67–77

Feng Y, Gong D, Zhang Q, Jiang S, Zhao L, Cui N (2019) Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. Energy Convers Manage 198:111780

Feng Y, Hao W, Li H, Cui N, Gong D, Gao L (2020) Machine learning models to quantify and map daily global solar radiation and photovoltaic power. Renew Sustain Energy Rev 118:109393

Feng Y, Zhang X, Jia Y, Cui N, Hao W, Li H, Gong D (2021) High-resolution assessment of solar radiation and energy potential in China. Energy Convers Manage 240:114265

Feng Y, Ziegler AD, Elsen PR, Liu Y, He X, Spracklen DV, Holden J, Jiang X, Zheng C, Zeng Z (2021) Upward expansion and acceleration of forest clearance in the mountains of Southeast Asia. Nature Sustain 4(10):892–899

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232

Gueymard CA (2001) Parameterized transmittance model for direct beam and circumsolar spectral irradiance. Sol Energy 71:325–346

Hargreaves GH, Samani ZA (1982) Estimating potential evapotranspiration. J Irrig Drain Div 108(3):225–230

Hassan GE, Youssef ME, Mohamed ZE, Ali MA, Hanafy AA (2016) New temperature-based models for predicting global solar radiation. Appl Energy 179:437–450

Hossain M, Mekhilef S, Olatomiwa L, Danesh M, Shamshirband S (2017) Application of extreme learning machine for short term output power forecasting of three grid-connected PV systems. J Clean Prod 167:395–405

Jahani B, Dinpashoh Y, Nafchi AR (2017) Evaluation and development of empirical models for estimating daily solar radiation. Renew Sustain Energy Rev 73:878–891

Jamil B, Akhtar N (2017) Estimation of diffuse solar radiation in the humid-subtropical climatic region of India: comparison of diffuse fraction and diffusion coefficient models. Energy 131:149–164

Jamil B, Siddiqui AT (2018) Estimation of monthly mean diffuse solar radiation over India: performance of two variable models under different climatic zones. Sustain Energy Technol Assess 25:161–180

Jia Y, Wang FC, Li PC, Huo SY, Yang T (2021) Simulating reference crop evapotranspiration with different climate data inputs using Gaussian exponential model. Environ Sci Pollut Res 28:41317–41336

Jiang S, Liang C, Cui N, Zhao L, Liu C, Feng Y, Hu XT, Gong DZ, Zou Q (2020) Water use efficiency and its drivers in four typical agroecosystems based on flux tower measurements. Agric for Meteorol 295:108200

Jin Z, Ye ZW, Gang Y (2005) General formula for estimation of monthly average daily global solar radiation in China. Energy Convers Manage 46(2):257–268

Kaba K, Sarıgül S, Avcı M, Kandırmaz M (2018) Estimation of daily global solar radiation using deep learning model. Energy 162:126–135

Katiyar AK, Pandey CK (2010) Simple correlation for estimating the global solar radiation on horizontal surfaces in India. Energy 35(12):5043–5048

Khatib T, Mohamed A, Sopian K (2012) A review of solar energy modeling techniques. Renew Sustain Energy Rev 16:2864–2869

Kisi O (2016) Modeling reference evapotranspiration using three different heuristic regression approaches. Agric Water Manage 169:162–172

Kisi O, Sanikhani H, Zounemat-Kermani M, Niazi F (2015) Long-term monthly evapotranspiration modeling by several data-driven methods without climatic data. Comput Electron Agric 115:66–77

Lesser B, Mucke M, Gansterer WW (2011) Effects of reduced precision on floating-point SVM classification accuracy. Procedia Comput Sci 4:508–517

Liu X, Mei X, Li Y, Wang Q, Jensen JR, Zhang Y, Porter JR (2009) Evaluation of temperature-based global solar radiation models in China. Agric Meteorol 149:1433–1446

Liu C, Zheng D, Zhao L, Liu C (2014) Gaussian fitting for carotid and radial artery pressure waveforms: comparison between normal subjects and heart failure patients. Bio-Med Mater Eng 24:271–277

Liu Y, Zhou Y, Wang D, Wan Y, Li Y, Zhu Y (2017) Classification of solar radiation zones and general models for estimating the daily global solar radiation on horizontal surfaces in China. Energy Convers Manag 154:167–179

Mecibah SM, Boukelia ET, Tahtah R, Gairaa K (2014) Introducing the best model for estimation the monthly mean daily global solar radiation on a horizontal surface (Case study:Algeria). Renew Sustain Energy Rev 36:194–202

Oates MJ, Ruiz-Canales A, Ferrández-Villena M, Fernández López A (2017) A low cost sunlight analyser and data logger measuring radiation. Comput Electron Agric 143:38–48

Pan T, Wu SH, Dai EF, Liu YJ (2013) Estimating the daily global solar radiation spatial distribution from diurnal temperature ranges over the Tibetan Plateau in China. Appl Energy 107:384–393

Persson G, Bacher P, Shiga T, Madsen H (2017) Multi-site solar power forecasting using gradient boosted regression trees. Sol Energy 120:423–436

Qiu RJ, Wang YK, Wang D, Qiu WJ, Wu JC, Tao WY (2020) Water temperature forecasting based on modified artificial neural network methods: two cases of the Yangtze River. Sci Total Environ 737:1–12

Quej VH, Almorox J, Arnaldo JA, Saito L (2017) ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. J Atmos Solar Terr Phys 155:62–70

Quinlan JR (1992) Learning with continuous classes. 5th Australian Joint Conference on Artificial Intelligence 92:343–348

Sattari MT, Pal M, Apaydin H et al (2013) M5 model tree application in Daily River flow forecasting in Sohu Stream Turkey. Water Resour 40(3):233–242

Shamshirband S, Mohammadi K, Tong CW, Zamani M, Motamedi S, Ch S (2016) A hybrid SVM-FFA method for prediction of monthly mean global solar radiation. Theoret Appl Climatol 125:53–65

Shiri J, Nazemi AH, Sadraddini AA, Landeras G, Kisi O, Fard AF, Marti P (2014) Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput Electron Agric 108:230–241

Shiri J, Sadraddini AA, Nazemi AH, Martí P, Fard AF, Kisi O, Landeras G (2015) Independent testing for assessing the calibration of the Hargreaves-Samani equation: New heuristic alternatives for Iran. Comput Electron Agric 117:70–80

Tabari H, Kisi O, Ezani A, Talaee PH (2012) SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. J Hydrol 44:78–89

Tian H, Zhao YQ, Luo M, He QQ, Han Y, Zeng ZL (2021) Estimating PM2.5 from multisource data: a comparison of different machine learning models in the Pearl River Delta of China. Urban Clim 35:100740

Vapink V (1999) The nature of statistical learning theory[M]. Springer-Verlag, New York

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 Sust Energy Rev 61:384–397

Wang Y, Witten IH (1997) Inducing model trees for continuous classes, In Proceedings of the ninth European conference on machine learning, pp 128–137

Wild M, Gilgen H, Roesch A, Ohmura A, Long CN, Dutton EG, Forgan B, Kallis A, Russak V, Tsvetkov A (2005) From dimming to brightening: decadal changes in solar radiation at Earth's surface. Sci 308(5723):847–850

Wu J, Lakshmi V, Wang D, Lin P, Pan M, Cai X, Wood EF, Zeng Z (2020) The reliability of global remote sensing evapotranspiration products over Amazon. Remote Sensing 12(14):2211

Wu ZJ, Cui NB, Hu XT, Gong DZ, Wang XS, Feng Y, Jiang SZ, Lu M, Han L, Xing LW, Zhu SD, Zhu N, Zhang YX, Zou QY, He ZL (2021) Optimization of extreme learning machine model with biological heuristic algorithms to estimate daily reference crop evapotranspiration in different climatic regions of China. J Hydrol 603:127028

Wu J, Feng Y, Liang L, He X, Zeng Z (2022) Assessing evapotranspiration observed from ECOSTRESS using flux measurements in agroecosystems. Agric Water Manag 269:107706

Wu J, Wang D, Li LZ, Zeng Z (2022) Hydrological feedback from projected Earth greening in the 21st century. Sustainable Horizons 1:100007

Yu HH, Chen YG, Hassan SG, Li DL (2016) Prediction of the temperature in a Chinese solar greenhouse based on LSSVM optimized by improved PSO. Comput Electron Agric 155:257–282

Zhang QW, Cui NB, Feng Y, Jia Y, Li Z, Gong DZ (2018) Comparative analysis of global solar radiation models in different regions of China. Advances in Meteorology 2018:1–21

Zhang Y, Cui N, Feng Y, Gong D, Hu X (2019) Comparison of BP, PSO-BP and statistical models for predicting daily global solar radiation in arid Northwest China. Comput Electron Agric 164:104905

Zheng MG, Hu SY, Liu XW, Wang W, Yin XC, Zheng L, Wang L, Lou YH (2019) Levels and distribution of synthetic musks in farmland soils from the Three Northeast Provinces of China. Ecotoxicol Environ Saf 172:303–307

Zhu B, Feng Y, Gong DZ, Jiang SZ, Zhao L, Cui NB (2020) Hybrid particle swarm optimization with extreme learning machine for daily reference evapotranspiration prediction from limited climatic data. Comput Electron Agric 173:105430

Zou L, Wang L, Xia L, Lin A, Hu B, Zhu H (2017) Prediction and comparison of solar radiation using improved empirical models and adaptive neuro-fuzzy inference systems. Renew Energy 106:343–353

Despotovic M, Nedic V, Despotovic D, Cvetanovic S (2015) Review and statistical analysis of different global solar radiation sunshine models. Renew Sustain Energy Rev 52:1869–1880

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.