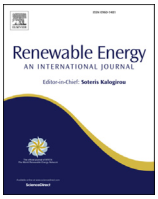
Renewable Energy 145 (2020) 2034–2045

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene



Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions

Junliang Fan [a], Lifeng Wu [a, *], Xin Ma [b], Hanmi Zhou [c], Fucang Zhang [a]

[a] Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas of Ministry of Education, Northwest A&F University, Yangling, 712100, China

[b] School of Science, Southwest University of Science and Technology, Mianyang, 621010, China

[c] College of Agricultural Engineering, Henan University of Science and Technology, Luoyang, 471003, China

ARTICLE INFO

Article history:
Received 26 April 2019
Received in revised form
20 July 2019
Accepted 23 July 2019
Available online 23 July 2019

Keywords:
Air pollution
Support vector machines
Extreme gradient boosting
Particle swarm optimization algorithm
Bat algorithm
Whale optimization algorithm

ABSTRACT

Increasing air pollutants significantly affect the proportion of diffuse (R_d) to global (R_s) solar radiation. This study proposed three new hybrid support vector machines (SVM) with particle swarm optimization algorithm (SVM-PSO), bat algorithm (SVM-BAT) and whale optimization algorithm (SVM-WOA) for predicting daily R_d in air-polluted regions. These models were further compared to standalone SVM, multivariate adaptive regression spline (MARS) and extreme gradient boosting (XGBoost) models. The results showed that models with suspended particulate matter with aerodynamic diameter smaller than 2.5 μm and 10 μm ($PM_{2.5}$ and PM_{10}) and ozone (O_3) produced more accurate daily R_d estimates than those without air pollution parameters, with average relative decreases in root mean square deviation (RMSD) of 11.1%, 10.0% and 10.4% for sunshine duration-based, R_s-based and combined models, respectively. SVM showed better accuracy than XGBoost and MARS. However, compared to SVM, SVM-BAT further enhanced the prediction accuracy and convergence rate in daily R_d modeling, followed by SVM-WOA and SVM-PSO, with relative decreases in RMSD of 2.9%–5.6%, 1.9%–4.9% and 1.1%–3.3%, respectively. The results highlighted the significance of incorporating air pollutants for more accurate estimation of daily R_d in air-polluted regions. Heuristic algorithms, especially BAT, are highly recommended for improving performance of standalone machine learning models.

1. Introduction

Reliable prediction of diffuse solar radiation (R_d), an important component of global solar radiation (R_s), is crucial for the optimum design and management of solar photovoltaic or thermal systems for renewable and clean energy production, which is also of significance for the evaluation of photosynthesis, canopy gas exchange and evapotranspiration processes in ecosystems [2–8]. The use of solar energy can significantly alleviate the adverse influence of traditional coal-fired electricity generation on our environment [9–13]. However, compared with other climatic parameters (e.g. sunshine hours, ambient temperature and even R_s), R_d data is much less readily accessible at most sites across the globe due to the costs of the instruments required to measure this parameter [14–16]. Taking China as an example, only 17 out of 726 long-term national meteorological stations record R_d [17]. Since measurements of R_d are often unavailable, various categories of empirical models have been proposed for prediction of R_d based on various readily available meteorological variables, such as sunshine duration-based models [18–21,71]), global solar radiation-based models [1,22–24,25,26] and combined models that predict R_d based on both R_s and sunshine hours as well as additional climatic parameters, e.g., air temperature and relative humidity [17,23,26–29]. However, it has been found that the increasing anthropogenic emission of air pollutants, e.g., fine particulate matter with an aerodynamic diameter smaller than 10 μm ($PM_{2.5}$), coarse particulate matter with an aerodynamic diameter smaller than 2.5 μm (PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO) and ozone (O_3), have scattered or absorbed the radiation from the Sun, and thus decreased the solar radiation that reaches the Earth's surface in heavily air-polluted regions [30–32]. The air pollution index (API), air quality index (AQI), suspended fine and coarse particulate matter ($PM_{2.5}$ and PM_{10}) have thus been

* Corresponding author. School of Hydraulic and Ecological Engineering, Nanchang Institute of Technology, Nanchang, 330099, China.
E-mail address: china.sw@163.com (L. Wu).

https://doi.org/10.1016/j.renene.2019.07.104
0960-1481/© 2019 Elsevier Ltd. All rights reserved.

incorporated in empirical models for more accurate prediction of R_d in air-polluted regions [33–37].

Empirical models have been largely utilized for R_d prediction globally, but these models are difficult to apply when involving complex relations between the input and output variables in noisy environments, e.g., under rainy, cloudy and air-polluted conditions [38,39]. Therefore, various machine learning models have been established for prediction of R_d, such as artificial neural networks (ANNs), e.g., radial basis function neural networks (RBNN) and generalized regression neural networks (GRNN) [40,41]; adaptive neuro fuzzy inference system (ANFIS) [42,43]; support vector machines (SVM) [44,45]; extreme learning machine (ELM) [72]; ktree-based soft computing models, e.g., boosted regression tree (BRT) and random forests (RF) models [46,47]. Ref. [48] developed an ANN model for prediction of hourly R_d in São Paulo of Brazil. They revealed that the proposed ANN models were more accurate and reliable compared with the studied empirical equations. Similar results have also been reported by Ref. [49] in Egypt, by Ref. [40] in China, by Ref. [41] in India and by Ref. [50] in Saudi Arabia, where the proposed ANN models remarkably outperformed the studied empirical models in predicting hourly, daily, monthly mean hourly or daily R_d based on R_s and sunshine duration [42]. adopted a neuro-fuzzy inference system (FIS) using relevance vector machines (RVM) for daily R_d estimation in Macedonia based on R_s and solar elevation angle. The results revealed the superiority of the FIS model for R_d estimation over the empirical models [46]. predicted daily R_d in Hong Kong of China using a BRT model. They revealed that the BRT model produced better R_d estimates than the regression model, and the cloud amount was the most important parameter for R_d estimation besides sunshine duration [43]. also explored the most influencing variables for daily R_d prediction in Kerman of Iran using the ANFIS model. They found that the combination of R_s and extra-terrestrial solar radiation (R_a) was the best two-parameter combination, while the combination of R_s, R_a and sunshine hours was the best three-parameter combination.

Ref. [44] explored the potential of the SVM model and single hidden layer feed-forward neural networks (SLFN) for predicting daily R_d on a tilted surface in Jeddah and Qassim of Saudi Arabia. The result showed that the SVM model gave much better daily R_d estimates, and it was more robust and faster relative to the SLFN model. Ref. [74] explored the potential of tree-type models for hourly and daily R_d estimation at five sites in Egypt. They revealed that the support vector regression (SVR) model exhibited both best estimation accuracy and model stability, while the tree-type models matched the performance of the SVR model, with significantly higher computational efficiency. Ref. [45] employed the SVM models for prediction of daily R_d in Beijing of China by considering the effects of air pollution. It was found that $PM_{2.5}$ and O_3 were the two major influencing air pollution-related variables for improving the prediction of daily R_d, while $PM_{2.5}$, PM_{10} and O_3 were the three optimum influencing air pollutants. The machine learning models have also started to be coupled with pre-treatment or heuristic algorithms for daily R_d prediction, such as wavelet transform (WT) and generic algorithm (GA), to optimize the training processes and further enhance the accuracy of standalone machine learning models. For instance, Ref. [51] coupled the SVM model and wavelet transform (SVM-WT) model to predict daily R_d in Kerman of Iran with sunshine duration. The results revealed that the hybrid SVM-WT showed much higher accuracy than the radial basis function SVM (SVM-RBF) and ANNs model, especially the third degree empirical equation. Ref. [72] proposed the hybrid backpropagation neural networks with genetic algorithm (GANN) for daily R_d estimation at two sites in North China with R_s and sunshine duration data, and compared its performance with ELM, RF, GRNN and empirical Iqbal models. They argued that the hybrid GANN model performed better than the ELM, followed by the RF, GRNN and Iqbal models.

Machine learning models have started to draw researchers' attention for prediction of R_d around the world, but studies have largely been limited to non-air-polluted climates. The increasing air pollutants can significantly affect the proportions of direct and diffuse solar radiation, thereby influencing the prediction accuracy of R_d using meteorological variables. Besides, a direct comparison of various types of machine learning models for R_d prediction at a specific site has been minimal, particularly some conventional and new models that have not been applied for R_d prediction yet, such as multivariate adaptive regression spline (MARS) and extreme gradient boosting (XGBoost). What's more important, machine learning models often have certain limitations in the appropriate selection or optimization of model parameters [52], and coupling the machine learning models with heuristic algorithms, especially the newly developed bat and whale optimization algorithms, can provide a better way to overcome these limitations. Therefore, the aims of this study are to: (1) to investigate the effects of various input combinations of variables, especially the air pollution-related variables, on the prediction accuracy of daily R_d in air-polluted Beijing of China as a case study, (2) to propose three types of machine learning models, including SVM, MARS and XGBoost, for prediction of daily R_d, and (3) to further evaluate the prediction accuracy of SVM models optimized by a classical and two new heuristic algorithms, including the particle swarm optimization algorithm (PSO), bat algorithm (BAT) and whale optimization algorithm (WOA), for more accurately estimating daily R_d. To the best of the authors' knowledge, this is the first implementation of SVM models with various heuristic algorithms for R_d prediction.

The rest of the paper is organized as follows. Section 2 describes the meteorological and air pollution data used in this study and the procedure for data quality control. It also provides a brief theoretical background of machine learning models and heuristic algorithms of evaluation. This section further introduces the developed models and the statistics for model evaluation. Main results are presented in Section 3 and the results are discussed is in Section 4, while the study conclusions along with future study are given in Section 5.

2. Materials and methods

2.1. Study site and data collection

As the capital and one of the severely polluted cities across China owing to the rapid urbanization and motorization over the past few decades, Beijing has started to draw the attention of researchers due to its serious air pollution (Fig. 1). Although the local air pollution has been alleviated in recent years by air pollution prevention and control actions, severe air pollution still occurs in adverse weather conditions. Daily observed meteorological variables, including global (R_s) and diffuse (R_d) solar radiation, sunshine hours (n), maximum and minimum temperatures (T_{max} and T_{min}), as well as three main air pollutants of $PM_{2.5}$, PM_{10} and O_3 were collected in Beijing over the period January 2014 to March 2017. The meteorological data were collected from the National Meteorological Information Center (NMIC) of China Meteorological Administration (http://www.nmic.cn/), while the air pollution dataset were collected from the China National Environmental Monitoring Center (http://www.cnemc.cn/). R_s and R_d were measured using the TBQ-2-B pyranometer (1 w m^{-2}, ±5%) [53]. Sunshine duration was measured by the Jordan sunshine recorder (0.1 h, ±0.2%) [54]. The WQG-11 dry- and wet-bulb thermometer (0.1 °C, ±2%) was used for maximum and minimum temperature measurements [55]. Continuous air quality monitoring systems

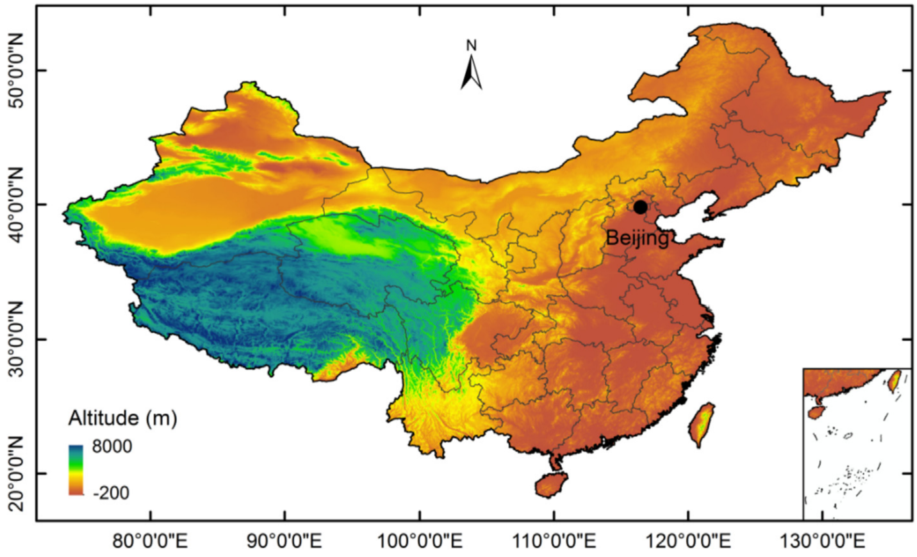


Fig. 1. Geographical location of Beijing in China.

were installed to measure $PM_{2.5}$, PM_{10} and O_3, where the β absorption method (1 nmol/mol, ±1%) was used to measure $PM_{2.5}$ and PM_{10}, and O_3 was determined using the UV-spectrophotometry method (1 nmol/mol, ±1%) [56]. The potential sunshine hours (N) and extra-terrestrial solar radiation (R_a) data were estimated on the basis of geographical, seasonal and solar information [39]. The data quality was further controlled by excluding dataset if any of the above meteorological and air quality data was missing or the ratio of R_s/R_a, R_d/R_s and n/N was greater than 1. The final dataset contained 1175 pairs of data points. Data from January 2014 to December 2015 (60% of total data points, 725 pairs) were used to train the models, while data during January 2016–March 2017) (40% of total data points, 450 pairs) were used for model testing (Fig. 2).

Table 1 presents the statistical values of the input and output variables during the training and testing stages. As seen from Table 1, T_{max} ranged from −10.9 °C in January to 41.1 °C in July, while T_{min} varied from −15.2 °C in January to 27.9 °C in July. The average sunshine duration was 6.7 h d^{-1} with the maximum value of 13.9 h d^{-1} in June and minimum value of 0 h d^{-1} in November. The highest concentrations of $PM_{2.5}$ and PM_{10} were observed in the cold and dry winter (476.0 μg m^{-3} and 501.0 μg m^{-3}, respectively) and the lowest values in the hot and wet summer (5.0 μg m^{-3} and 0 μg m^{-3}, respectively), while the concentration of O_3 had the

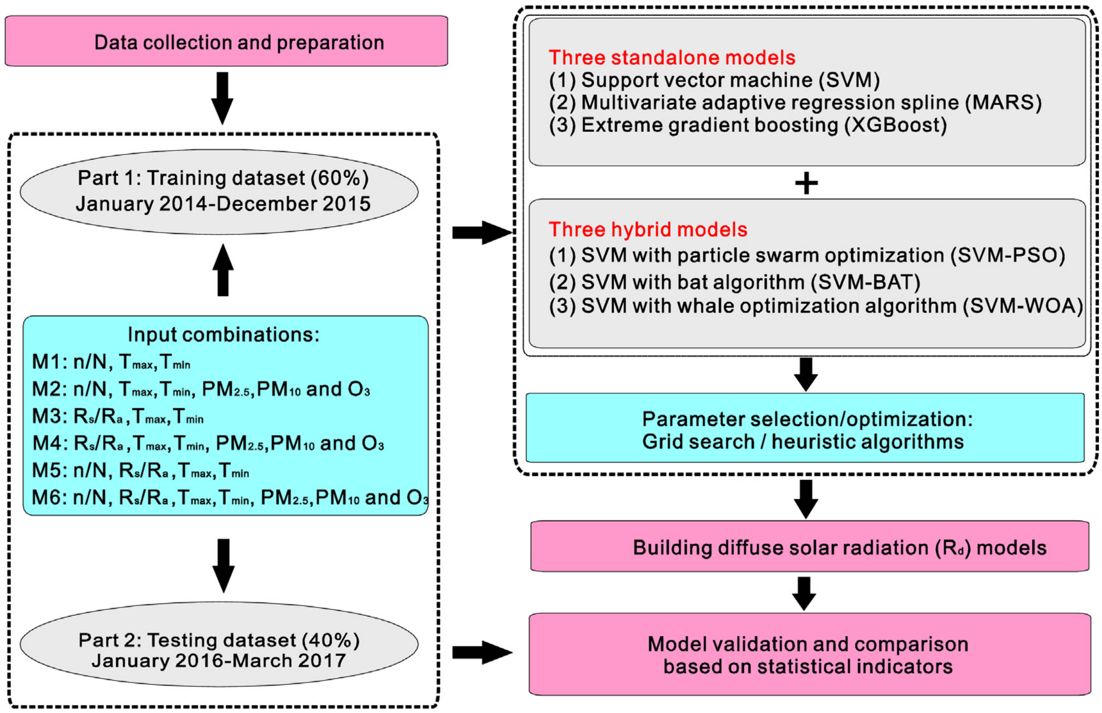

Data collection and preparation
Part 1: Training dataset (60%)
January 2014-December 2015
Input combinations:
M1: n/N, Tmax, Tmin
M2: n/N, Tmax, Tmin, PM2.5, PM10 and O3
M3: Rs/Ra, Tmax, Tmin
M4: Rs/Ra, Tmax, Tmin, PM2.5, PM10 and O3
M5: n/N, Rs/Ra, Tmax, Tmin
M6: n/N, Rs/Ra, Tmax, Tmin, PM2.5, PM10 and O3
Part 2: Testing dataset (40%)
January 2016-March 2017
Three standalone models
(1) Support vector machine (SVM)
(2) Multivariate adaptive regression spline (MARS)
(3) Extreme gradient boosting (XGBoost)
+
Three hybrid models
(1) SVM with particle swarm optimization (SVM-PSO)
(2) SVM with bat algorithm (SVM-BAT)
(3) SVM with whale optimization algorithm (SVM-WOA)
Parameter selection/optimization:
Grid search / heuristic algorithms
Building diffuse solar radiation (Rd) models
Model validation and comparison
based on statistical indicators

Fig. 2. Simple flowchart of the proposed methodology in the present study.

Table 1
Statistical parameters of the daily meteorological variables during training and testing. T_{max}: maximum temperature; T_{min}: minimum temperature; n: sunshine duration; $PM_{2.5}$ and PM_{10}: suspended particulate matter smaller than 2.5 μm and 10 μm in aerodynamic diameter; O_3: ozone; R_s: global solar radiation; R_d: diffuse solar radiation; SD: standard deviation; C_s: coefficient of skewness; K: kurtosis; the same below.

Parameter	T_{max} (°C)	T_{min} (°C)	n (h d^{-1})	$PM_{2.5}$ (μg m^{-3})	PM_{10} (μg m^{-3})	O_3 (μg m^{-3})	R_s (MJ m^{-2} d^{-1})	R_d (MJ m^{-2} d^{-1})
Training data (January 2014-December 2015)								
Maximum	41.1	27.9	13.6	476	464	169	30.9	17.5
Minimum	−1.8	−11.2	0.0	5	0	2	0.0	0.0
Mean	19.8	9.6	6.6	81	107	58	14.2	6.9
SD	11.1	10.4	4.1	70	76	39	7.5	3.9
C_s	−0.3	−0.2	−0.4	2.0	1.6	0.6	0.15	0.49
K	−1.3	−1.3	−1.2	5.4	3.7	−0.3	−0.98	−0.73
Testing data (January 2016-March 2017)								
Maximum	37.8	26.9	13.9	430	501	164	31.1	17.5
Minimum	−10.9	−15.2	0.0	6	0	2	0	0
Mean	17.0	6.9	6.8	73	96	54	14.4	6.4
SD	11.4	11.1	3.8	67	77	36	7.4	3.9
C_s	0	0.2	−0.5	1.9	1.8	0.7	0.17	0.61
K	−1.3	−1.3	−1.0	4.8	4.1	0	−0.87	−0.54

highest concentration in summer (169.0 μg m^{-3}) and lowest value in winter (2.0 μg m^{-3}). R_s and R_d were highest in June (31.1 MJ m^{-2} d^{-1} and 17.5 MJ m^{-2} d^{-1}, respectively), and lowest in November (both 0 MJ m^{-2} d^{-1}). It is also noted from Table 1 that the differences between the meteorological and air pollution-related variables during the training and testing stages did not vary substantially.

2.2. Machine learning models

2.2.1. Support vector machines (SVM)

The SVM model was established by Ref. [57]; adopting the theory of structural risk minimization instead of empirical risk minimization to reduce the over-fitting problem. The SVM model can be used for classification, pattern recognition and regression analysis problems. For a classification problem, the original problem is transformed into a convex quadratic programming problem. The SVM model can estimate the regressions on the basis of a set of kernel functions. These functions can implicitly transform the raw, low-dimensional input dataset to a high-dimensional feature space. The radial basis function-based kernel function was utilized in this study as a result of its outstanding performance for predicting R_d relative to other kernel functions [44], such as linear, polynomial and sigmoid functions. Further information about the SVM model is given by Ref. [57].

2.2.2. Multivariate adaptive regression spline (MARS)

The MARS model is a complicated non-parametric regression approach and was introduced by Ref. [58]. Using the so-called ‘divide-and-conquer’ rule, the training samples can be divided into different regions with their own regression lines. There is no need to make specific assumptions on the basic relations between input parameters and outputs. The endpoints of a line segment are referred to nodes. The knot can be any data point in the parameter range. The choice of knot includes searching all possible knot positions of single variables according to the adaptive regression algorithms. The knot is used to represent the end of a data region and the beginning of another. The resulting piecewise curve is called the basis function (BFs), which makes the model more flexible. The MARS model generates BFs through a step-by-step search. The adaptive regression algorithms are used to select node positions, which seek all the potential univariate node positions and interactions between all parameters. The model is divided into two parts: base function and pruning. Pruning is to reduce the degree of overfitting of the model. Further information about the MARS model is given in Ref. [58].

2.2.3. Extreme gradient boosting (XGBoost)

The XGBoost model was developed by Ref. [76] and originated from the idea of “boosting”. The XGBoost model integrates all the predicted values of a series of “weak” learners to develop “strong” learners via an additive training process. The XGBoost model is supposed to avoid the over-fitting problem and decrease the computational time, because the objective functions are simplified and the predictive and regularization terms are combined, while it maintains optimum computation efficiency at the same time. Parallel calculation is also automatically implemented during the training period to solve big-data science problems in a fast and accurate way. For more details about the XGBoost model refer to Ref. [59].

2.3. Heuristic algorithms

2.3.1. Particle swarm optimization algorithm (PSO)

The PSO algorithm was proposed by Ref. [60] based on the social psychological behavior of animals, e.g. bird flocking, insect swarming and fish schooling. It is composed by many individuals (denoted as particles with positions and velocities) to refine their positions in the specific search space. Every particle stands for a potential solution to the optimization issue. The particles alter their positions in the search space multidimensionally to find out the space positions of high fitness. The PSO algorithm first initializes the particles with random position and velocity. During optimizing, the fitness evaluation function is used to locate the positions of all particles and assign the fitness values to them. All the particles remember their best positions currently visited (the local best) by using a memory function. Meanwhile, the population memorizes the best position among all the individual best positions encountered so far (the global best) (Fig. 3). The inertia weight is implemented for balancing the global and local searching ability of the particles. Further details regarding the PSO algorithm can be found in Ref. [60].

2.3.2. Bat algorithm (BAT)

The BAT algorithm was first formulated by Ref. [61]; and is a bio-inspired optimization algorithm with fast convergence and simulates the behavior of echolocating bats in preying their foods. Bats utilize a kind of sonar named echolocation to locate the prey and avoid obstacles in the darkness. When preying their foods, the bats release ultrasonic pulses. The loudness at this time is the maximum, which helps lengthen the ultrasonic propagation distance. In the BAT algorithm, every virtual bat releases a series of loud ultrasound waves and detects the echoes returned from the

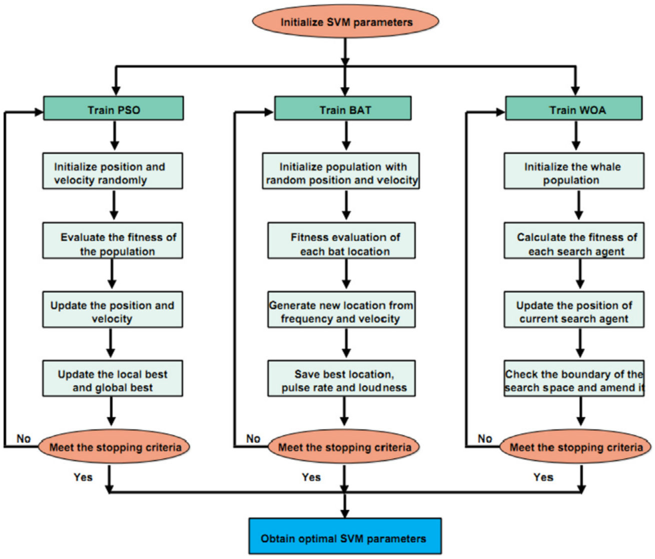

Initialize SVM parameters
Train PSO
Train BAT
Train WOA
Initialize position and velocity randomly
Evaluate the fitness of the population
Update the position and velocity
Update the local best and global best
Meet the stopping criteria
No
Yes
Initialize population with random position and velocity
Fitness evaluation of each bat location
Generate new location from frequency and velocity
Save best location, pulse rate and loudness
Meet the stopping criteria
No
Yes
Initialize the whale population
Calculate the fitness of each search agent
Update the position of current search agent
Check the boundary of the search space and amend it
Meet the stopping criteria
No
Yes
Obtain optimal SVM parameters

Fig. 3. Structure of the particle swarm optimization algorithm (PSO), bat algorithm (BAT) and whale optimization algorithm (WOA).

objects. The initial bat population is randomly generated from m real-valued vectors with dimension d in terms of Eq. (1):

$$x_{ij} = x_{\min\,d} + \mu(x_{\min\,d} - x_{\min\,d}) \tag{1}$$

where, i = 1, 2, …, m, and j = 1, 2, …, d. $x_{\max\ d}$ and $x_{\min\ d}$ are the upper and lower boundary of d-dimensional space, respectively. μ∈[0, 1] is a random value.

The position (X_i) and velocity (V_i) of each bat (i) in a d-dimensional search space are defined and updated during the iterations. The new positions and velocities of a virtual bat at time step (t) are updated by:

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \tag{2}$$

$$V_i^t = V_i^{t-1} + \left(V_i^{t-1} - V^*\right)f_i \tag{3}$$

$$V_i^t = V_i^{t-1} + V_i^t \tag{4}$$

where, f_i is the frequency value of the ith bat. $f_{\min}$ and $f_{\max}$ are the minimum and maximum frequency values, respectively. β is a random number uniformly drawn from [0, 1]. X* is the current global best solution, which is located after comparing all solutions within m bats.

This algorithm combines both the population-based and local search algorithms. The approach includes a series of iterations, in which a number of solutions alter by randomly modifying the signal bandwidths that increase with harmonics. For the local search phase, a new solution for each bat is locally generated using the random walk given by Eq. (5):

$$x_{new} = x_{old} + \varepsilon A^t \tag{5}$$

where, ε is a random number uniformly drawn from [0, 1], and $A^t = \langle A_i^t \rangle$is the average loudness of all the bats at this time step. The pulse rates and loudness are then renewed when the newer solution is satisfied. Further details regarding the BAT algorithm can be found in Ref. [61].

2.3.3. Whale optimization algorithm (WOA)

The WOA algorithm was established by Ref. [62] and simulates the behavior of humpback whales, which is a very new but powerful bio-inspired algorithm. The whales search their foods by a special behavior named bubble-net hunting, where bubbles are created by encircling or through '9'-shaped path. When hunting, the humpback whales go down in the water about 10–15 m deep and produce bubbles in a spiral shape that encircle the prey. The whales finally follow the produced bubbles and move towards the surface to catch the prey. In the WOA algorithm, the optimization process starts by initializing the population of whales randomly. The whales then seek the prey's (optimum) location, and attach (optimize) them by using the encircling approach or the bubble-net hunting method. To locate the food locations and attack them, two mechanisms are used by the whales. The preys are firstly encircled and the second consists of creating bubble nets. The search space is explored when the whales look for foods and the exploitation occurs when the whales attack the prey.

(1) Searching and encircling prey:

$$Z_i^{t+1} = Z_{rand} - A\left|C \cdot Z_{rand} - Z_i^t\right| \tag{6}$$

where, *A* and *C* are coefficients, computed as:

$$A = 2 \cdot a \cdot r - a \tag{7}$$

$$C = 2 \cdot r \tag{8}$$

where, '*a*' is linearly decreasing between 2 and 0, and '*r*' is the random number from 0 to 1.

$$Z_i^{t+1} = g_{best} - A \cdot \left|C \cdot g_{best} - Z_i^t\right| \tag{9}$$

Here, if*A* > 1, then will be used equation (13) represent searching prey behavior, otherwise, equation (16) will be used.

(2) Spirally updating position:

Position updating is expressed as Eq.(17):

$$Z_i^{t+1} = \begin{cases} g_{best} - A \cdot \left|C \cdot g_{best} - Z_i^t\right|, & p < 0.5 \\ \left|C \cdot g_{best} - Z_i^t\right| \cdot \exp(bl) \cdot \cos(2\pi l) + g_{bset}, & p \geq 0.5 \end{cases} \tag{10}$$

where, *p* is the random number ranging between 0 and 1, *l* varies from 0 to 1 and *b* is a constant to describe the spiral shape. In this study, the value was set as 1. More details about the WOA algorithm can be found in Ref. [62].

2.4. Input combinations and parameter optimization

Six input combinations of meteorological and environmental variables from three model groups were used in this study to explore the influence of various input parameters, especially the air pollution parameters, on daily R_d prediction. Three meteorological data-based models were first utilized for prediction of daily R_d, i.e., (1) n/N, T_{max} and T_{min}, (2) R_s/R_a, T_{max} and T_{minr}, (3) n/N, R_s/R_a, T_{max} and T_{min} (see Table 2). Three major air pollution-related variables, i.e. $PM_{2.5}$, PM_{10} and O_3, were further included in these baseline models to explore their impact on the accuracy enhancement for predicting daily R_d. For the SVM, MARS and XGBoost models, the main parameters of the models were selected with the grid search approach, while the parameters of the three hybrid SVM models

Table 2
The input combinations of meteorological variables for different machine learning models. N: potential sunshine hours; R_a: extra-terrestrial solar radiation; the same below.

Model group	Combination no.	Input combination	Output
Group 1: sunshine duration-based	M1	n/N, T_{max}, T_{min}	R_d
	M2	n/N, T_{max}, T_{min}, $PM_{2.5}$, PM_{10}, O_3	
Group 2: global solar radiation-based	M3	R_s/R_a, T_{max}, T_{min}	
	M4	R_s/R_a, T_{max}, T_{min}, $PM_{2.5}$, PM_{10}, O_3	
Group 3: combined	M5	n/N, R_s/R_a, T_{max}, T_{min}	
	M6	n/N, R_s/R_a, T_{max}, T_{min}, $PM_{2.5}$, PM_{10}, O_3	

were optimized by the three heuristic algorithms (Fig. 3).

2.5. Comparison statistics

The accuracies of the six machine learning models in daily R_d modeling were assessed with four commonly used statistical indices [63,64,75], i.e. coefficient of determination (R^2), root mean square deviation (RMSD), scatter index (SI) and mean absolute error (MAE), which can be expressed as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{m}\left(Y_{i,o}-\overline{Y}_{i,o}\right)\left(Y_{i,p}-\overline{Y}_{i,p}\right)\right]^2}{\sum_{i=1}^{m}\left(Y_{i,o}-\overline{Y}_{i,o}\right)^2\sum_{i=1}^{n}\left(Y_{i,p}-\overline{Y}_{i,p}\right)^2} \tag{11}$$

$$RMSD = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(Y_{i,o}-Y_{i,p}\right)^2} \tag{12}$$

$$SI = RMSD/\overline{Y}_{i,o} \tag{13}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|Y_{i,o}-Y_{i,p}\right| \tag{14}$$

where $Y_{i,o}$, $Y_{i,p}$, $\overline{Y}_{i,o}$ and m are the observed R_d, the predicted R_d, the mean of observed R_d and the number of observation data, respectively. Higher R^2 values indicate high prediction accuracy, whereas the smaller values of RMSD, SI and MAE suggest greater model performance. The model was considered to be excellent when $SI < 0.1$, good when $0.1 < SI < 0.2$, fair when $0.2 < SI < 0.3$ and poor when $SI > 0.3$ [65].

3. Results

The statistical results of the six machine learning models (MARS, XGBoost, SVM, SVM-PSO, SVM-BAT and SVM-WOA) for estimating daily R_d in air-polluted Beijing during training and testing under various input combinations are presented in Table 3. As shown in the table, the predicted daily R_d values were significantly different under various input combinations. Machine learning models in Group 3 (on average $R^2 = 0.917$, RMSD = 1.153 MJ m^{-2} d^{-1}, SI = 0.182 and MAE = 0.792 MJ m^{-2} d^{-1}) outperformed those in Group 1 (on overage $R^2 = 0.803$, RMSD = 1.719 MJ m^{-2} d^{-1}, SI = 0.267 and MAE = 0.906 MJ m^{-2} d^{-1}) and Group 2 (on overage $R^2 = 0.897$, RMSD = 1.318 MJ m^{-2} d^{-1}, SI = 0.204 and MAE = 0.792 MJ m^{-2} d^{-1}) during testing. It was obvious that the global solar radiation-based models gave much better daily R_d estimates than models based on sunshine hours, with an average relative decrease in RMSD of 23.0%. However, a combination of sunshine duration and R_s further enhanced the accuracy of daily R_d prediction, with average relative decreases in RMSD of 32.7% and 12.5% compared with the single sunshine duration and global solar radiation-based models, respectively. It can be seen from Table 3 that the R^2 values of most machine learning models in Groups 2 and 3 were greater than 0.9, indicating a strong relationship between the measured R_d values and those predicted by these models. The corresponding RMSD and MAE values were generally lower than 1.4 MJ m^{-2} d^{-1} and 1.0 MJ m^{-2} d^{-1} respectively, which indicated favorable prediction accuracy. The corresponding SI values of most models were less than 0.2, showing good performance for the prediction of daily R_d. However, all the machine learning models in Group 1 exhibited fair performance in predicting daily R_d, with $0.2 < SI < 0.3$.

It was also found that machine learning models with three main air pollutants of $PM_{2.5}$, PM_{10} and O_3 as inputs produced more accurate daily R_d estimates in air-polluted Beijing during testing (on average $R^2 = 0.826$, RMSD = 1.613 MJ m^{-2} d^{-1}, SI = 0.254 and MAE = 1.118 MJ m^{-2} d^{-1} in Group 1; $R^2 = 0.904$, RMSD = 1.249 MJ m^{-2} d^{-1}, SI = 0.197 and MAE = 0.887 MJ m^{-2} d^{-1} in Group 2; $R^2 = 0.926$, RMSD = 1.090 MJ m^{-2} d^{-1}, SI = 0.172 and MAE = 0.763 MJ m^{-2} d^{-1} in Group 3), compared with those without considering air pollution parameters (on average $R^2 = 0.781$, RMSD = 1.814 MJ m^{-2} d^{-1}, SI = 0.284 and MAE = 1.273 MJ m^{-2} d^{-1} in Group 1; $R^2 = 0.890$, RMSD = 1.249 MJ m^{-2} d^{-1}, SI = 0.197 and MAE = 0.887 MJ m^{-2} d^{-1} in Group 2; $R^2 = 0.907$, RMSD = 1.090 MJ m^{-2} d^{-1}, SI = 0.172 and MAE = 0.763 MJ m^{-2} d^{-1} in Group 3), with average relative decreases in RMSD of 11.1%, 10.0% and 10.4% in the three model groups, respectively.

As seen in Table 3, the predicted daily R_d values also differed among the six ELM models under different input combinations. In terms of statistical values averaged among the six input combinations, the SVM model ($R^2 = 0.872$, RMSD = 1.390 MJ m^{-2} d^{-1}, SI = 0.219 and MAE = 0.959 MJ m^{-2} d^{-1}) generally gave better accuracy in predicting daily R_d than the XGBoost model ($R^2 = 0.863$, RMSD = 1.455 MJ m^{-2} d^{-1}, SI = 0.227 and MAE = 0.994 MJ m^{-2} d^{-1}) and MARS model ($R^2 = 0.847$, RMSD = 1.536 MJ m^{-2} d^{-1}, SI = 0.243 and MAE = 1.079 MJ m^{-2} d^{-1}) during testing. However, relative to the standalone SVM model, the hybrid SVM-BAT model (on average $R^2 = 0.883$, RMSD = 1.336 MJ m^{-2} d^{-1}, SI = 0.208 and MAE = 0.920 MJ m^{-2} d^{-1}) further improved the prediction accuracy of daily R_d, followed by the SVM-WOA model (on average $R^2 = 0.881$, RMSD = 1.347 MJ m^{-2} d^{-1}, SI = 0.211 and MAE = 0.926 MJ m^{-2} d^{-1}) and SVM-PSO model (on average $R^2 = 0.879$, RMSD = 1.357 MJ m^{-2} d^{-1}, SI = 0.212 and MAE = 0.923 MJ m^{-2} d^{-1}). Specifically, the RMSD values of the hybrid SVM-BAT, SVM-WOA and SVM-PSO models were decreased by 2.9%–5.6%, 1.9%–4.9% and 1.1%–3.3% under different input combinations respectively, relative to the conventional SVM model. Obviously, the SVM-BAT model exhibited better prediction accuracy in daily R_d prediction and outperformed the other two hybrid SVM models.

The scatter plots of the estimated daily R_d values using the six soft computing models against the measured values during testing (January 2016–March 2017) under different input combinations are presented in Fig. 4. As shown in the figure, the plotted data points in Groups 2 and 3 generally correlated closer towards the 1:1 line compared those in Group 1. Nevertheless, the XGBoost and MARS models yielded more scattered R_d points relative to the other

Table 3
Statistical values of the six machine learning models for predicting daily R_d in Beijing during training and testing under different input combinations.

Combination/Model	Training				Testing			
	R^2	RMSD (MJ m^{-2} d^{-1})	SI	MAE (MJ m^{-2} d^{-1})	R^2	RMSD (MJ m^{-2} d^{-1})	SI	MAE (MJ m^{-2} d^{-1})
M1: n/N, T_{max}, T_{min}								
MARS	0.781	1.827	0.267	1.316	0.753	1.951	0.307	1.338
XGBoost	0.920	1.124	0.164	0.842	0.761	1.934	0.293	1.332
SVM	0.779	1.756	0.256	1.285	0.774	1.831	0.288	1.267
SVM-PSO	0.796	1.778	0.259	1.199	0.789	1.771	0.279	1.224
SVM-BAT	0.814	1.694	0.247	1.138	0.799	1.729	0.272	1.232
SVM-WOA	0.812	1.701	0.248	1.143	0.796	1.741	0.276	1.231
M2: n/N, T_{max}, T_{min}, $PM_{2.5}$, PM_{10}, O_3								
MARS	0.848	1.523	0.222	1.108	0.815	1.667	0.262	1.192
XGBoost	0.994	0.331	0.048	0.249	0.799	1.738	0.274	1.220
SVM	0.859	1.479	0.216	1.091	0.828	1.603	0.252	1.087
SVM-PSO	0.869	1.418	0.207	0.984	0.834	1.573	0.248	1.080
SVM-BAT	0.870	1.406	0.205	0.973	0.837	1.556	0.245	1.067
SVM-WOA	0.870	1.413	0.205	0.974	0.837	1.560	0.246	1.070
M3: R_s/R_a, T_{max}, T_{min}								
MARS	0.893	1.273	0.186	0.907	0.843	1.677	0.264	1.147
XGBoost	0.964	0.756	0.110	0.545	0.884	1.383	0.216	0.961
SVM	0.910	1.187	0.173	0.845	0.887	1.368	0.215	0.938
SVM-PSO	0.922	1.102	0.161	0.705	0.904	1.334	0.197	0.850
SVM-BAT	0.923	1.091	0.159	0.694	0.904	1.321	0.196	0.846
SVM-WOA	0.922	1.101	0.161	0.708	0.904	1.325	0.197	0.848
M4: R_s/R_a, T_{max}, T_{min}, $PM_{2.5}$, PM_{10}, O_3								
MARS	0.920	1.125	0.164		0.872	1.466	0.231	1.047
XGBoost	0.996	0.220	0.032	0.854	0.910	1.243	0.194	0.844
SVM	0.931	1.031	0.151	0.722	0.908	1.233	0.195	0.879
SVM-PSO	0.941	0.956	0.139	0.653	0.910	1.220	0.193	0.843
SVM-BAT	0.955	0.843	0.123	0.567	0.912	1.191	0.186	0.842
SVM-WOA	0.955	0.848	0.124	0.565	0.910	1.205	0.190	0.856
M5: n/N, R_s/R_a, T_{max}, T_{min}								
MARS	0.906	1.193	0.174	0.781	0.885	1.301	0.211	0.931
XGBoost	0.986	0.474	0.069	0.323	0.901	1.269	0.199	0.826
SVM	0.918	1.119	0.163	0.677	0.910	1.224	0.193	0.834
SVM-PSO	0.924	1.097	0.160	0.706	0.904	1.207	0.190	0.813
SVM-BAT	0.927	1.055	0.154	0.643	0.916	1.167	0.184	0.792
SVM-WOA	0.925	1.070	0.156	0.656	0.914	1.178	0.185	0.790
M6: n/N, R_s/R_a, T_{max}, T_{min}, $PM_{2.5}$, PM_{10}, O_3								
MARS	0.924	1.084	0.158	0.764	0.916	1.154	0.180	0.821
XGBoost	0.998	0.136	0.020	0.094	0.920	1.165	0.184	0.783
SVM	0.943	0.930	0.136	0.594	0.927	1.083	0.173	0.750
SVM-PSO	0.943	0.944	0.138	0.632	0.929	1.059	0.168	0.741
SVM-BAT	0.945	0.921	0.134	0.591	0.931	1.052	0.166	0.739
SVM-WOA	0.948	0.901	0.129	0.576	0.931	1.053	0.167	0.751

proposed machine learning models. The daily R_d values predicted from the hybrid SVM-BAT, SVM-WOA and SVM-PSO models were closer to the observed values under all input combinations. Fig. 5 presents the Taylor diagram visualizing the models' accuracy in daily R_d prediction under different input combinations. It was clear that the three hybrid SVM models, particularly the SVM-BAT model, were located closer to the reference point, indicating better performance compared to the other models. A further comparison of the convergence rate of the three bio-inspired algorithms showed that the SVM-BAT model generally converged faster and to a lower RMSD value than the SVM-WOA model under different input combinations, followed by the SVM-PSO model (Fig. 6).

4. Discussion

The results indicated that global solar radiation-based models outperformed models based on sunshine hours, while a combination of sunshine duration and R_s further improved the prediction accuracy of daily R_d. These generally agreed well with previous studies on R_d prediction using empirical and munchies learning models [17,66,67], where R_d was found to be more related to R_s and sunshine duration compared with other meteorological variables. It was also found that incorporation of air pollution parameters is important in readily available meteorological variable-based machine learning models for more accurate and reliable daily R_d estimates in air-polluted cities such as Beijing. Ref. [36] developed an empirical model for daily R_d estimation in four cities of China, and found that the accuracy of existing empirical models was improved by incorporating air quality index (AQI). Ref. [35] also developed an empirical model for R_d prediction in India based on the quadratic form of relative sunshine duration and exponential form of air pollution index (API). They found RMSD was decreased by about 10% compared with the traditional diffuse solar model. Ref. [68] explored the potential of SVM models for prediction of daily R_d in Beijing, China. They found that SVM models with AQI produced better prediction accuracy than traditional models based on only meteorological data.

However, AQI or API is calculated from the concentrations of six major air pollutants and are commonly used to evaluate the impact of air pollution on human health. The higher the pollutant concentration is, the greater the weight is. Taking China as an example, suspended particulate matter concentrations are higher in winter and ozone concentrations are higher in summer; therefore, the resulting AQI is high in both summer and winter. However, this air pollution parameter cannot explain which air pollutant influences the diffuse radiation most from the mechanism. Ref. [9] found that

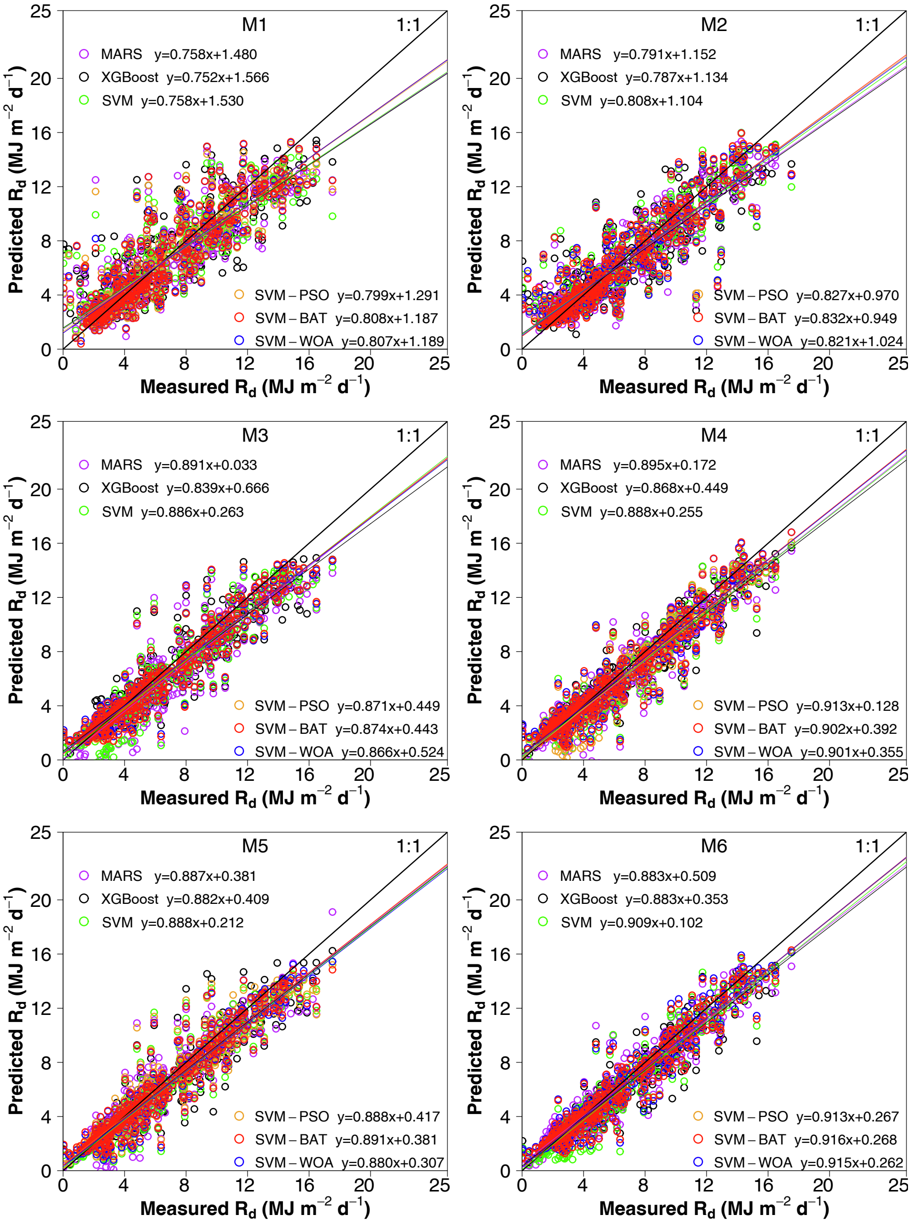


Fig. 4. Scatter plots of predicted daily R_d values by the six machine learning models against the observed R_d values during testing (January 2016–March 2017) under various input combinations in Beijing. Colored lines are the linear fits under various input combinations.

the incorporation of $PM_{2.5}$ and PM_{10} improved the prediction accuracy of ANN-based diffuse solar radiation models. In the present study, the specific concentrations of air pollutants instead of AQI or API were used as inputs to develop the models. This made the model more clear, and the changes of diffuse radiation in response to air pollutant concentrations can also be easily obtained through the developed model. This is of great significance for us to better understand the effect of air pollutant concentrations on radiation scattering. Suspended particulate matter mainly reflects the solar radiation to the outer atmosphere, while O_3 is the product of photochemical smog. Since NO_x substances absorb solar radiation and generate a large amount of ozone, higher ozone concentration indicates that more solar radiation would be consumed before reaching the Earth's surface.

This study also suggested that the hybrid SVM models showed much higher prediction accuracy in daily R_d modeling than the classical SVM model, which outperformed the XGBoost and MARS models. Particularly, the hybrid SVM-BAT model outperformed the SVM-WOA and SVM-PSO models in estimating daily R_d. Many studies have shown that the PSO algorithm often converges to the local optimal solution prematurely [69,73]. When the current optimal solution is determined after a series of iterations, the algorithm stops searching for alternative solutions, which is easy to cause premature convergence to the local optimal solution. However, the step sizes of BAT and WOA algorithms are gradually reduced. When the local optimal solution is obtained, the algorithms keep searching for potentially better solutions in the global space [70,77]. This ensured the better efficiency of the BAT and WOA algorithms in optimizing the SVM models for daily R_d prediction. Overall, the novel hybrid SVM-BAT model showed more accurate daily R_d estimates with faster convergence rate than the other models, and is thus highly recommended for accurately

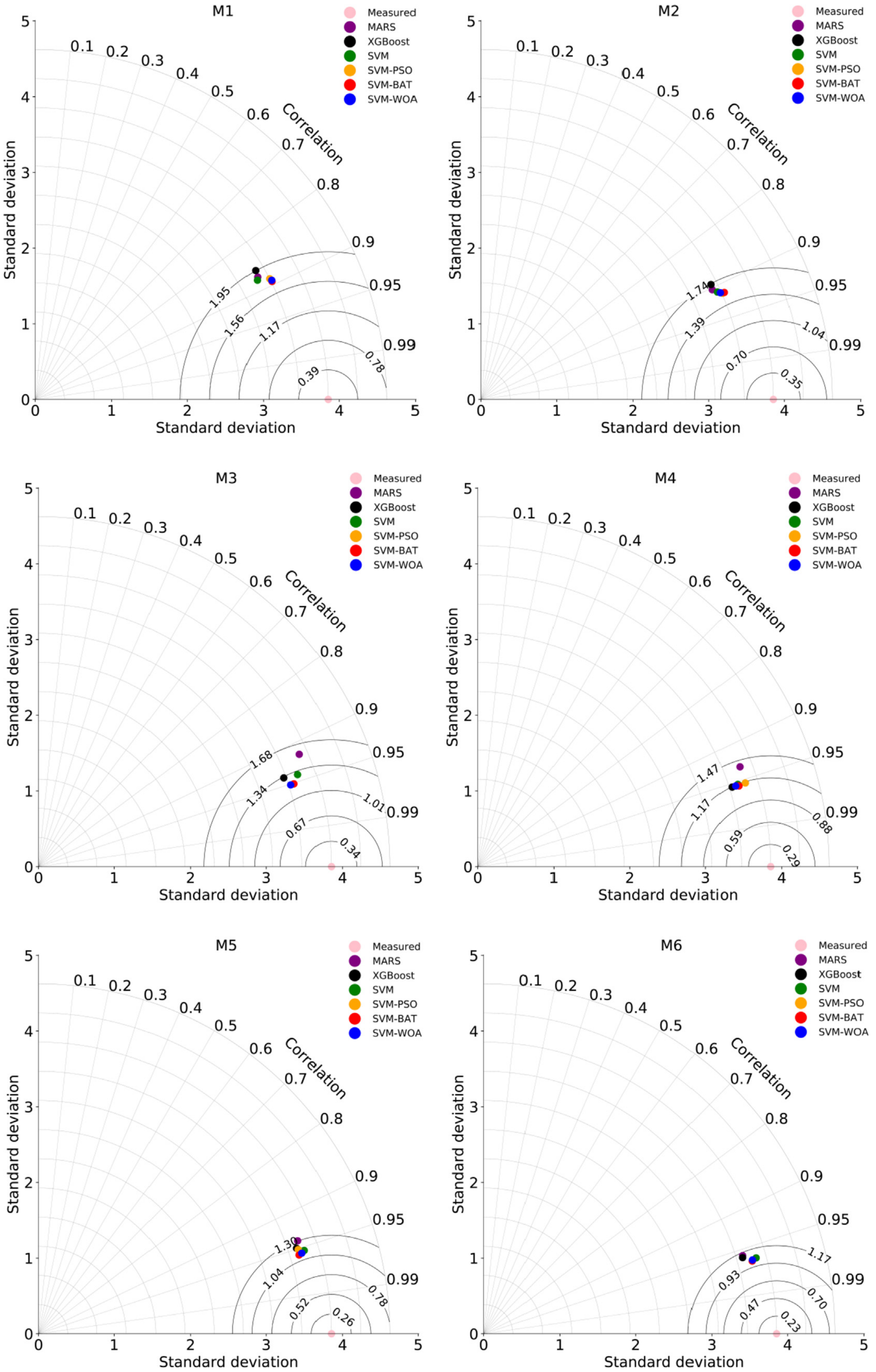


Fig. 5. Taylor diagram for the six machine learning models for daily R_d prediction during testing (January 2016–March 2017) under various input combinations in Beijing.

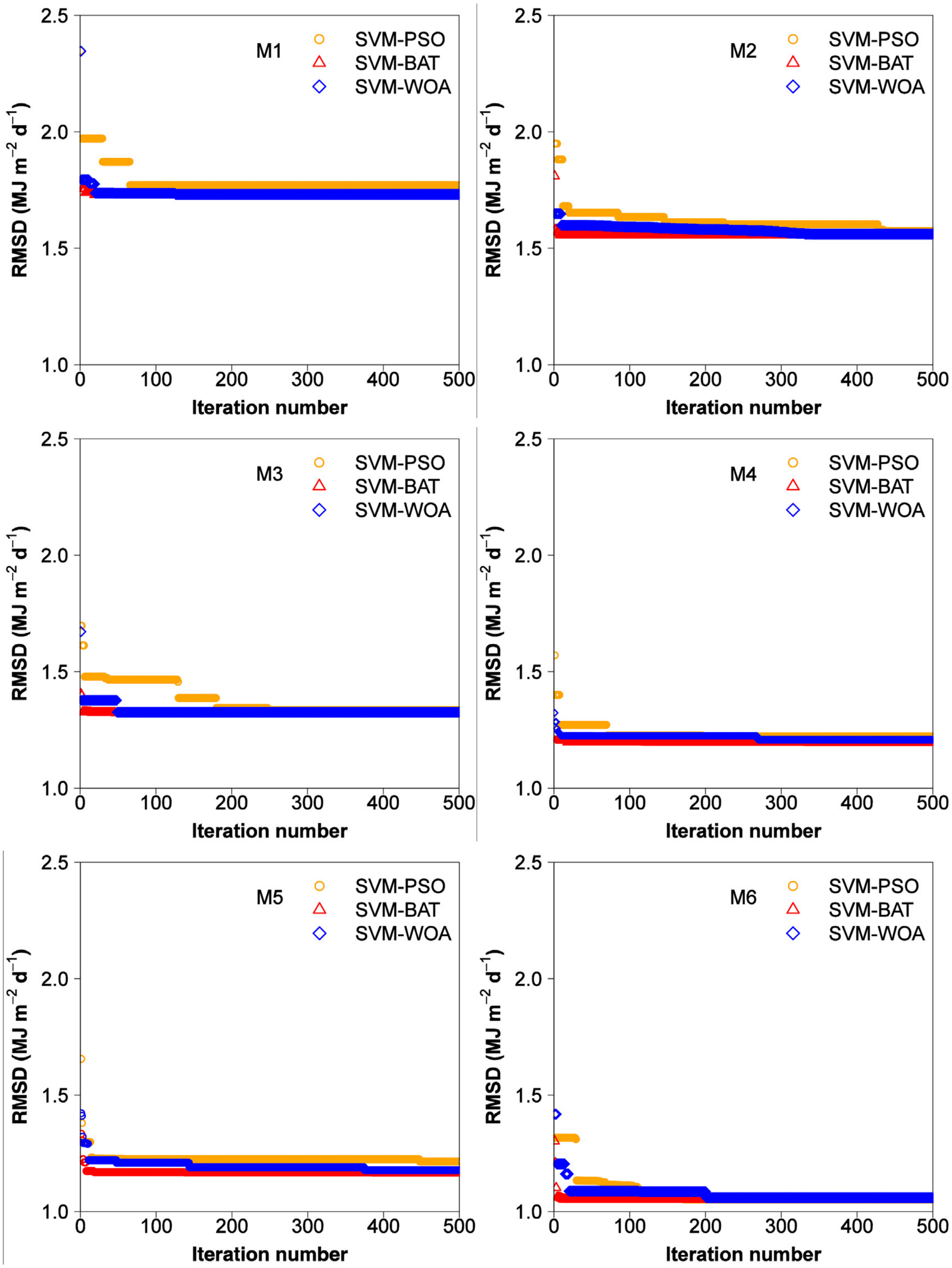


Fig. 6. Changes in the root mean square deviation (RMSD) of diffuse solar radiation obtained by the three hybrid SVM models.

predicting daily R_d in Beijing and maybe other air-polluted cities using various input variables.

5. Conclusions

Measuring or predicting the diffuse fraction of global solar radiation (R_d) is a precondition for the optimum design and evaluation of the solar photovoltaic or thermal system, which can generate renewable and clean electrical or thermal energy for industrial and agricultural production. However, direct measurement of R_d is not readily available at most global locations due to the high cost of the equipment; meanwhile, accurate prediction of R_d has also been difficult, especially in air-polluted regions due to increasing anthropogenic emissions of air pollutants. The increase of pollution load in the atmosphere can significantly affect the proportions of direct and diffuse solar radiation. The present study explored the potential of three hybrid support vector machines (SVM) with particle swarm optimization algorithm (SVM-PSO), bat algorithm (SVM-BAT) and whale optimization algorithm (SVM-WOA) for predicting daily R_d in air-polluted Beijing as a case study. The hybrid SVM models were further compared to the standalone SVM, multivariate adaptive regression spline (MARS) and extreme gradient boosting (XGBoost) models. Six input combinations of daily maximum/minimum temperature, sunshine duration, suspended particulate matter with an aerodynamic diameter smaller than 2.5 μm and 10 μm ($PM_{2.5}$ and PM_{10}), ozone (O_3) and global solar radiation (R_s) during January 2014–March 2017 were utilized to train and test the models.

The results showed that a combination of sunshine hours and R_s improved the prediction accuracy of daily R_d, with average relative decreases in root mean square deviation (RMSD) of 32.7% and 12.5% compared with the two single models, respectively. Models with $PM_{2.5}$, PM_{10} and O_3 produced more accurate daily R_d estimates than those without considering these air pollution parameters, with average relative decreases in RMSD of 11.1%, 10.0% and 10.4% for the sunshine duration-based, R_s-based and combined models, respectively. The standalone SVM model showed better accuracy for prediction of daily R_d than the XGBoost and MARS models. However, compared to the standalone SVM model, the SVM-BAT models further enhanced the accuracy of daily R_d prediction, followed by the hybrid SVM-WOA and SVM-PSO models, with relative decreases in RMSD of 2.9%–5.6%, 1.9%–4.9% and 1.1%–3.3%, respectively. The results advocated the possibility of incorporating air pollution-related variables and hybridizing the SVM model with the heuristic algorithms, especially BAT, for attaining more preciseness and reliability in predicting daily R_d in air-polluted regions. However, further study is needed to assess these models on various time scales, e.g. hourly or monthly. The hybridization of other machine learning models with these heuristic algorithms should also be tested for daily R_d prediction.

Acknowledgements

This study was jointly supported by the National Natural Science Foundation of China (No. 51879226, 51709143) and Natural Science Foundation of Jiangxi Province, China (No. 20181BAB206045). Thanks to the National Meteorological Information Center of China Meteorological Administration and the China National Environmental Monitoring Center for offering the data.

References

[1] T.E. Boukelia, M.-S. Mecibah, I.E. Meriche, General models for estimation of the monthly mean daily diffuse solar radiation (Case study: Algeria), Energy Convers. Manag. 81 (2014) 211–219.
[2] B. Jamil, A.T. Siddiqui, Generalized models for estimation of diffuse solar radiation based on clearness index and sunshine duration in India: applicability under different climatic zones, J. Atmos. Sol. Terr. Phys. 157 (2017) 16–34.
[3] S.C.S. Costa, A.S.A.C. Diniz, L.L. Kazmerski, Solar energy dust and soiling R&D progress: literature review update for 2016, Renew. Sustain. Energy Rev. 82 (2018) 2504–2536.
[4] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: a review, Renew. Energy 105 (2017) 569–582.
[5] D. Li, S. Lou, Review of solar irradiance and daylight illuminance modeling and sky classification, Renew. Energy 126 (2018) 445–453.
[6] A. Khosravi, R.N.N. Koury, L. Machado, J.J.G. Pabon, Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms, J. Clean. Prod. 176 (2018a) 63–75.
[7] J. Fan, L. Wu, F. Zhang, H. Cai, W. Zeng, X. Wang, H. Zou, 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 (2019a) 186–212.
[8] S. Yang, B. Fu, Y. Hou, S. Chen, Z. Li, S. Wang, Transient cooling and operational performance of the cryogenic part in reverse Brayton air refrigerator, Energy 167 (2019) 921–938.
[9] M. Vakili, S.R. Sabbagh-Yazdi, S. Khosrojerdi, K. Kalhor, Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data, J. Clean. Prod. 141 (2017) 1275–1285.
[10] A. Khosravi, R.O. Nunes, M.E.H. Assad, L. Machado, Comparison of artificial intelligence methods in estimation of daily global solar radiation, J. Clean. Prod. 194 (2018b) 342–358.
[11] J. Fan, X. Wang, L. Wu, F. Zhang, H. Bai, X. Lu, Y. Xiang, New combined models for estimating daily global solar radiation based on sunshine duration in humid regions: a case study in South China, Energy Convers. Manag. 156 (2018a) 618–625.
[12] Y. Feng, N. Cui, Y. Chen, D. Gong, X. Hu, Development of data-driven models for prediction of daily global horizontal irradiance in Northwest China, J. Clean. Prod. 223 (2019) 136–146.
[13] L. Wu, G. Huang, J. Fan, F. Zhang, X. Wang, W. Zeng, Potential of kernel-based nonlinear extension of Arps decline model and gradient boosting with categorical features support for predicting daily global solar radiation in humid regions, Energy Convers. Manag. 183 (2019a) 280–295.
[14] B. Jahani, Y. Dinpashoh, A.R. Nafchi, Evaluation and development of empirical models for estimating daily solar radiation, Renew. Sustain. Energy Rev. 73 (2017) 878–891.
[15] L. Zou, L. Wang, L. Xia, A. Lin, B. Hu, H. Zhu, Prediction and comparison of solar radiation using improved empirical models and Adaptive Neuro-Fuzzy Inference Systems, Renew. Energy 106 (2017) 343–353.
[16] M.A. Hassan, A. Khalil, S. Kaseb, M.A. Kassem, Potential of four different machine-learning algorithms in modeling daily global solar radiation, Renew. Energy 111 (2017a) 52–62.
[17] Y. Jiang, Estimation of monthly mean daily diffuse radiation in China, Appl. Energy 86 (2009) 1458–1464.
[18] M. Iqbal, Correlation of average diffuse and beam radiation with hours of bright sunshine, Sol. Energy 23 (1979) 169–173.
[19] A.A. El-Sebaii, A.A. Trabea, Estimation of horizontal diffuse solar radiation in Egypt, Energy Convers. Manag. 44 (2003) 2471–2482.
[20] C.K. Pandey, A.K. Katiyar, A comparative study to estimate daily diffuse solar radiation over India, Energy 34 (2009) 1792–1796.
[21] B. Jamil, N. Akhtar, Comparison of empirical models to estimate monthly mean diffuse solar radiation from measured data: case study for humid-subtropical climatic region of India, Renew. Sustain. Energy Rev. 77 (2017) 1326–1342.
[22] A. Miguel, J. Bilbao, R. Aguiar, H.D. Kambezidis, E. Negro, Diffuse solar irradiation model evaluation in the North Mediterranean belt area, Sol. Energy 70 (2) (2001) 143–153.
[23] I. Karakoti, B. Pande, K. Pandey, Evaluation of different diffuse radiation models for Indian stations and predicting the best fit model, Renew. Sustain. Energy Rev. 15 (2011) 2378–2384.
[24] T. Lealea, R. Tchinda, Estimation of diffuse solar radiation in the north and far north of Cameroon, Eur. Sci. Journal, ESJ 9 (2013).
[25] T.E. Boukelia, M.-S. Mecibah, I.E. Meriche, General models for estimation of the monthly mean daily diffuse solar radiation (Case study: Algeria), Energy Convers. Manag. 81 (2014) 211–219.
[26] N. Bailek, K. Bouchouicha, Z. Al-Mostafa, M. El-Shimy, N. Aoun, A. Slimani, S. Al-Shehri, A new empirical model for forecasting the diffuse solar radiation over Sahara in the Algerian Big South, Renew. Energy 117 (2018) 530–537.
[27] K. Bakirci, Models for the estimation of diffuse solar radiation for typical cities in Turkey, Energy 82 (2015) 827–838.
[28] H.D. Kambezidis, B.E. Psiloglou, D. Karagiannis, U.C. Dumka, D.G. Kaskaoutis, Recent improvements of the Meteorological Radiation Model for solar irradiance estimates under all-sky conditions, Renew. Energy 93 (2016) 142–158.
[29] H.D. Kambezidis, B.E. Psiloglou, D. Karagiannis, U.C. Dumka, D.G. Kaskaoutis, Meteorological Radiation Model (MRM v6.1): improvements in diffuse radiation estimates and a new approach for implementation of cloud products, Renew. Sustain. Energy Rev. 74 (2017) 616–637.
[30] Y. Wang, Y. Yang, N. Zhao, C. Liu, Q. Wang, The magnitude of the effect of air pollution on sunshine hours in China, J. Geophys. Res. Atmos. 117 (2012), D00V14.
[31] X. Yang, C. Zhao, L. Zhou, Y. Wang, X. Liu, Distinct impact of different types of aerosols on surface solar radiation in China, J. Geophys. Res. Atmos. 121 (2016) 6459–6471.
[32] J. Khodakarami, P. Ghobadi, Urban pollution and solar radiation impacts, Renew. Sustain. Energy Rev. 57 (2016) 965–976.
[33] C. Furlan, A.P. De Oliveira, J. Soares, G. Codato, J.F. Escobedo, The role of clouds in improving the regression model for hourly values of diffuse solar radiation, Appl. Energy 92 (2012) 240–254.
[34] N. Zhao, X. Zeng, S. Han, Solar radiation estimation using sunshine hour and air pollution index in China, Energy Convers. Manag. 76 (2013) 846–851.
[35] M. Suthar, G.K. Singh, R.P. Saini, Effects of air pollution for estimating global solar radiation in India, Int. J. Sustain. Energy 36 (2017) 20–27.
[36] W. Yao, C. Zhang, X. Wang, J. Sheng, Y. Zhu, S. Zhang, The research of new daily diffuse solar radiation models modified by air quality index (AQI) in the region with heavy fog and haze, Energy Convers. Manag. 139 (2017) 140–150.
[37] Q. Zhao, W. Yao, C. Zhang, X. Wang, Y. Wang, Study on the influence of fog and haze on solar radiation based on scattering-weakening effect, Renew. Energy 134 (2019) 178–185.
[38] O. Kisi, K.S. Parmar, Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution, J. Hydrol. 534 (2016) 104–112.
[39] V.H. Quej, J. Almorox, M. Ibrakhimov, L. Saito, Estimating daily global solar radiation by day of the year in six cities located in the Yucat{á}n Peninsula,

Mexico, J. Clean. Prod. 141 (2017) 75–82.

[40] Y. Jiang, Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models, Energy Policy 36 (2008) 3833–3837.

[41] S. Alam, S.C. Kaushik, S.N. Garg, Assessment of diffuse solar energy under general sky condition using artificial neural network, Appl. Energy 86 (2009) 554–564.

[42] E. Lazarevska, J. Trpovski, A neuro-fuzzy model of the solar diffuse radiation with relevance vector machine, in: 11th International Conference on Electrical Power Quality and Utilisation, 2011, pp. 1–6.

[43] K. Mohammadi, S. Shamshirband, D. Petković, H. Khorasanizadeh, Determining the most important variables for diffuse solar radiation prediction using adaptive neuro-fuzzy methodology; case study: city of Kerman, Iran, Renew. Sustain. Energy Rev. 53 (2016) 1570–1579.

[44] M.A.M. Ramli, S. Twaha, Y.A. Al-Turki, Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study, Energy Convers. Manag. 105 (2015) 442–452.

[45] J. Fan, L. Wu, F. Zhang, H. Cai, X. Wang, X. Lu, Y. Xiang, Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature, Renew. Sustain. Energy Rev. 94 (2018b) 732–747.

[46] S. Lou, D.H.W. Li, J.C. Lam, W.W.H. Chan, Prediction of diffuse solar irradiance using machine learning and multivariable regression, Appl. Energy 181 (2016) 367–374.

[47] L. Benali, G. Notton, A. Fouilloy, C. Voyant, R. Dizene, Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components, Renew. Energy 132 (2019) 871–884.

[48] J. Soares, A.P. Oliveira, M.Z. Božnar, P. Mlakar, J.F. Escobedo, A.J. Machado, Modeling hourly diffuse solar-radiation in the city of São Paulo using a neural-network technique, Appl. Energy 79 (2004) 201–214.

[49] H.K. Elminir, Y.A. Azzam, F.I. Younes, Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models, Energy 32 (2007) 1513–1523.

[50] S. Rehman, M. Mohandes, Splitting global solar radiation into diffuse and direct normal fractions using artificial neural networks. Energy sources, Part A recover, Util. Environ. Eff. 34 (2012) 1326–1336.

[51] S. Shamshirband, K. Mohammadi, H. Khorasanizadeh, L. Yee, M. Lee, D. Petković, E. Zalnezhad, Estimating the diffuse solar radiation using a coupled support vector machine–wavelet transform model, Renew. Sustain. Energy Rev. 56 (2016) 428–435.

[52] G.-B. Huang, An insight into extreme learning machines: random neurons, random features and kernels, Cognit. Comput. 6 (2014) 376–390.

[53] G.Y. Shi, T. Hayasaka, A. Ohmura, Z.H. Chen, B. Wang, J.Q. Zhao, L. Xu, Data quality assessment and the long-term trend of ground solar radiation in China, Journal of applied meteorology and climatology 47 (4) (2008) 1006–1016.

[54] X. Liu, X. Mei, Y. Li, J.R. Porter, Q. Wang, Y. Zhang, Choice of the Ångström–Prescott coefficients: are time-dependent ones better than fixed ones in modeling global solar irradiance? Energy Convers. Manag. 51 (12) (2010) 2565–2574.

[55] X. Ye, F. Chen, Z. Hou, The effect of temperature on thermal sensation: a case study in Wuhan city, China, Procedia Engineering 121 (2015) 2149–2156.

[56] Y. Wang, Q. Ying, J. Hu, H. Zhang, Spatial and temporal variations of six criteria air pollutants in 31 provincial capital cities in China during 2013–2014, Environ. Int. 73 (2014) 413–422.

[57] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 2013.

[58] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. (1991) 1–67.

[59] T. Chen, T. He, M. Benesty, others, Xgboost: extreme gradient boosting, 2015. R Packag. version 0.4-2 1–4.

[60] R. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, in: MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995, pp. 39–43.

[61] X.-S. Yang, A new metaheuristic bat-inspired algorithm BT - nature inspired cooperative strategies for optimization (NICSO 2010), in: J.R. González, D.A. Pelta, C. Cruz, G. Terrazas, N. Krasnogor (Eds.), Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 65–74.

[62] S. Mirjalili, A. Lewis, The whale optimization algorithm, Adv. Eng. Software 95 (2016) 51–67.

[63] J. Fan, W. Yue, L. Wu, F. Zhang, H. Cai, X. Wang, X. Lu, Y. Xiang, Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China, Agric. For. Meteorol. 263 (2018c) 225–241.

[64] X. Lu, Y. Ju, L. Wu, J. Fan, F. Zhang, Z. Li, Daily pan evaporation modeling from local and cross-station data using three tree-based machine learning models, J. Hydrol 566 (2018) 668–684.

[65] K. Mohammadi, S. Shamshirband, M.H. Anisi, K.A. Alam, D. Petković, Support vector regression based prediction of global solar radiation on a horizontal surface, Energy Convers. Manag. 91 (2015) 433–441.

[66] J. Zhang, L. Zhao, S. Deng, W. Xu, Y. Zhang, A critical review of the models used to estimate solar radiation, Renew. Sustain. Energy Rev. 70 (2017) 314–329.

[67] N.S. Chukwujindu, A comprehensive review of empirical models for estimating global solar radiation in Africa, Renew. Sustain. Energy Rev. 78 (2017) 955–995.

[68] W. Yao, C. Zhang, H. Hao, X. Wang, X. Li, A support vector machine approach to estimate global solar radiation with the influence of fog and haze, Renew. Energy 128 (2018) 155–162.

[69] Z. Dong, D. Yang, T. Reindl, W.M. Walsh, A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance, Energy 82 (2015) 570–577.

[70] L. Wu, H. Zhou, X. Ma, J. Fan, F. Zhang, Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio-inspired optimization algorithms: application in contrasting climates of China, J. Hydrol. 577 (2019b) 123960.

[71] J. Fan, L. Wu, F. Zhang, H. Cai, X. Ma, H. Bai, Evaluation and development of empirical models for estimating daily and monthly mean daily diffuse horizontal solar radiation for different climatic regions of China, Renew. Sustain. Energy Rev. 105 (2019b) 168–186.

[72] Y. Feng, N. Cui, Q. Zhang, L. Zhao, D. Gong, Comparison of artificial intelligence and empirical models for estimation of daily diffuse solar radiation in North China Plain, Int. J. Hydrogen Energy 42 (2017) 14418–14428.

[73] L.M. Halabi, S. Mekhilef, L. Olatomiwa, J. Hazelton, Performance analysis of hybrid PV/diesel/battery system using HOMER: a case study Sabah, Malaysia, Energy Convers. Manag. 144 (2017) 322–339.

[74] M.A. Hassan, A. Khalil, S. Kaseb, M.A. Kassem, Exploring the potential of tree-based ensemble methods in solar radiation modeling, Appl. Energy 203 (2017b) 897–916.

[75] X. Ma, M. Xie, W. Wu, X. Wu, B. Zeng, A novel fractional time delayed grey model with Grey Wolf Optimizer and its applications in forecasting the natural gas and coal consumption in Chongqing China, Energy 178 (2019) 487–507.

[76] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, 2016 (arXiv Prepr).

[77] X. Ma, A brief introduction to the Grey Machine Learning, J. Grey Syst 31 (1) (2019) 1–12.