Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2020.010130
Article

Tech Science Press

A PSO-XGBoost Model for Estimating Daily Reference Evapotranspiration in the Solar Greenhouse

Jingxin Yu[1,3], Wengang Zheng[1,*], Linlin Xu[3], Lili Zhangzhong[1], Geng Zhang[2] and Feifei Shan[1]

[1]National Engineering Research Center for Information Technology in Agriculture, Beijing, 100097, China
[2]National Agro-tech Extension and Service Center, Beijing, 100125, China
[3]School of Land Science and Technology, China University of Geosciences, Beijing, 100083, China
[*]Corresponding Author: Wengang Zheng. Email: zhengwg@nercita.org.cn

Abstract: Accurate estimation of reference evapotranspiration (ET_0) is a critical prerequisite for the development of agricultural water management strategies. It is challenging to estimate the ET_0 of a solar greenhouse because of its unique environmental variations. Based on the idea of ensemble learning, this paper proposed a novel ET_{0i} estimation model named PSO-XGBoost, which took eXtreme Gradient Boosting (XGBoost) as the main regression model and used Particle Swarm Optimization (PSO) algorithm to optimize the parameters of XGBoost. Using the meteorological and soil moisture data during the two-crop planting process as the experimental data, and taking ET_{0i} calculated based on the improved Penman–Monteith equation as the reference truth, the accuracy of model estimation was evaluated and the impact of less input variables on model estimation was tested. The results showed that PSO algorithm could optimize the parameters of XGBoost model stably, PSO-XGBoost model could accurately estimate ET_{0i} in various data modes, and the estimation accuracy of the model decreases with the decrease of the number of input variables. Compared with other integrated learning models, PSO-XGBoost model could obtain the best estimation performance of ET_{0i}.

Keywords: Reference evapotranspiration; XGBoost; particle swarm optimization; solar greenhouse

1 Introduction

Reference evapotranspiration (ET_0) is proposed by the United Nations Food and Agriculture Organization (FAO) in 1977, it is an important parameter for calculating Evapotranspiration (ET) [1]. At present, ET_0 has become a basic parameter in the field of water resources management and irrigation operation. FAO-56 Penman–Monteith (P–M) equation is usually used to estimate the ET_0 in farmland environment [2]. The air circulation and heat exchange in the solar greenhouse are quite different from that in outdoor farmland, some studies have proposed some improved P-M equation methods to calculate ET_0 in the solar greenhouse, but the calculation processes were still complicated and tedious [3–5]. Therefore, it is necessary to use Artificial Intelligence (AI) algorithms to simplify the calculation process and reduce the input parameters to accurately estimate ET_0 in the greenhouse environment, which can reduce the cost of data acquisition and help facility agriculture improve the efficiency of irrigation water use.

In recent years, machine learning algorithms, especially AI algorithms, have been applied to ET_0 estimation because of its high fitting accuracy and flexibility [6]. Antonopoulos et al. [7] used limited meteorological input parameters and Artificial Neural Network (ANN) algorithm to predict ET_0 in northern Greece, and proved that ANN algorithm could predict ET_0 with less input variables. Shiri et al. [8] proposed a Gene Expression Programming (GEP) approach to estimate ET_0 of four weather stations in northern Spain,

and found that this approach was superior to Adaptive Neuro-Fuzzy Inference System (ANFIS), Priestley-Taylor and Hargreaves-Samani models. Pour-Ali Baba et al. [9] estimated the ET_0 of 2 sites in South Korea by combining ANFIS and ANN algorithm, and found that this method could estimate ET_0 precisely. Furthermore, some researchers compared the effect of different machine learning algorithms in estimating ET_0. Sanikhani et al. [10] compared Multi-Layer Perceptron (MLP), Generalized Regression Neural Networks (GRNN), Radial Basis Neural Networks (RBNN), integrated ANFIS with grid partitioning and subtractive clustering (ANFIS-GP and ANFIS-SC), and GEP and other machine learning algorithms carried out model estimation for ET_0 in Mediterranean region of Turkey, he found that GEP and GRNN models can obtain better prediction accuracy. Wu et al. [11] compared eight algorithms of four types of models: Neuron-based (MLP, GRNN and ANFIS), kernel-based (SVM, KNEA), tree-based (M5Tree) and curve-based (MARS) models. It was found that the kernel-based SVM, KNEA and curve-based MARS models could achieve better estimation results. However, it is difficult for a single machine learning model to further improve the estimation ability of ET_0, and it shows obvious instability, especially when we build the model with less input variables, it is difficult to obtain better estimation performance.

At present, ensemble learning is the popular development trend of AI algorithms. It combines independent models into stronger learners, which can achieve better stability and prediction effect compared with individual models [12]. Some studies have applied ensemble learning model to estimate ET_0, Fan et al. compared four tree-based algorithms, including Random Forest (RF), M5 model tree (M5Tree), Gradient Boosting Decision Tree (GBDT) and XGBoost, to estimate the ET_0 of 8 sites in China, and found that XGBoost could obtain the best prediction accuracy and the modeling time cost was lower [13]. Huang et al. [14] evaluated the CatBoost method for prediction of ET_0 in humid regions and found that CatBoost performed better than RF and SVM models when the complete input data were available. Fan et al. [15] proposed Light Gradient Boosting Machine (LightGBM) for predicting daily ET_0, the LightGBM outperformed M5Tree, RF and four empirical models. However, the setting of AI model parameters has a great impact on the estimation ability of the model. The adjustment of model parameters usually takes a lot of time and requires solid professional knowledge, especially when the input items are relatively complex [16]. Ensemble learning algorithms have advantages over individual machine learning models. Among ensemble learning models, XGBoost model is a tree-based ensemble learning model proposed in 2016 [17], which has won many machine learning competitions and been widely used in industry and academia [18].

To improve the efficiency of parameter adjustment of AI model, there are many studies using optimization algorithm to optimize the parameters of ET_0 estimation model. For example, the Genetic Algorithm (GA) was used to optimize the SVM model to develop the GA-SVM model to estimate ET_0 of the semi-arid environment in northwest China. The results showed that the estimation ability of GA-SVM model was better than that of SVM and ANN model [19]. Han et al. [20]combined with a bat algorithm with XGBoost to estimate ET_0 in the arid and semiarid regions of China, and compared the different meteorological input variables, the results of the study found that bat algorithm optimized XGBoost model to get a better prediction effect. Liu et al. [21] proposed an Extreme Learning Machine (ELM) method optimized by Particle Swarm Optimization (PSO) algorithm (PSO-SWELM) to realize more accurate evapotranspiration estimation with limited environmental data, the results showed that PSO-SWELM estimation effect was better than other models (BP, PSO-BP, SVM, ELM and PSO-ELM) in the estimation of ET_0. In addition, whale optimization algorithm, flower pollination algorithm, cuckoo search algorithm and ant colony optimization are also used to optimize the ET_0 prediction model in recent studies [22–23]. Among the optimization algorithms, PSO algorithm is the widely used and stable optimization algorithm based on swarm intelligence [24].

This study was based on the meteorological and soil moisture data of the two-crop planting process from 2018 to 2019 in the solar greenhouse located in Beijing, China. The ET_0 calculated by the improved P-M equation for solar greenhouse was used as the reference truth to evaluate the accuracy of model estimation, and then the ET_0 estimation model based on AI algorithms was constructed. XGBoost model has strong data fitting ability and PSO optimization algorithm can effectively improve the performance of

the model, however, few studies have applied the combination of the two to the estimation of ET_0 in solar greenhouse. The main purpose of this study is: (1) To propose an intelligent model PSO-XGBoost which uses the PSO algorithm to optimize parameters; (2) To evaluate the effect of the PSO-XGBoost model in estimating ET_0 in the solar greenhouse; (3) To evaluate the performance of the PSO-XGBoost model with less input variables.

2 Methodology

2.1 XGBoost Model

XGBoost is a machine learning algorithm realized by gradient lifting technology, it is an enhanced GBDT algorithm. Its base classifier is the Classification and Regression Tree (CART). XGBoost is a tree integration model combines multiple CART [17]. The XGBoost model is built by adding trees iteratively. The predicted value of the i-th sample in the t-th iteration can be expressed as follows:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(X_i) \tag{1}$$

The tree is added iteratively to minimize the objective function, which can be expressed as:

$$obj^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)\right) + \Omega(f_t) \tag{2}$$

where obj is the loss function and $\Omega(f_t)$ represents the model complexity.

To optimize the objective quickly, the second-order Taylor expansion [25] is used for Eq. (2), as shown in Eq. (3).

$$obj^{(t)} \approx \sum_{i=1}^{n}\left(L\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_i(X_i) + \frac{1}{2} h_i f_i^2(X_i)\right) + \Omega(f_i) \tag{3}$$

where $g_i = \partial_{j_i^{(t-1)}} L\left(y_i, \hat{y}_i^{(t-1)}\right)andh_i = \partial^2_{\hat{y}_i^{(t-1)}} L\left(y_i, \hat{y}_i^{(t-1)}\right)$ are the first and second derivatives of loss function terms respectively.

When adding the t-th tree, the previous t-1 tree has completed the training, that is, $L\left(y_i, \hat{y}_i^{(t-1)}\right)$ is a constant term. Remove this term to obtain the simplified objective function of step t, which can be expressed as:

$$\widetilde{obj}^{(t)} = \sum_{i=1}^{n}\left(g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i)\right) + \Omega(f_t) \tag{4}$$

Define $I_j = \{i | q(X_i) = j\}$ as the sample set of leaf node j, by expanding the regular term Ω, the Eq. (4) can be transformed into:

$$\widetilde{obj}^{(t)} = \sum_{i=1}^{n}\left(g_i \omega_{q(x_j)} + \frac{1}{2} h_i \omega^2_{q(x_i)}\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{5}$$

where ω_j^* is the weight of leaf node j.

Finally, the objective function is optimized, and the optimal solution can be expressed as:

$$\omega_j^* = -\sum_{i \in I_j} g_i / \left(\sum_{i \in I_j} h_i + \lambda\right) \tag{6}$$

$$\widetilde{obj}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{7}$$

Leaf node split is based on the input variable of the model, through the input variables are applied to divide the number of leaf nodes to calculate the scores of the importance of input variables, the score reflects the correlation between input variables and the model output, so we can according to the relative importance score of input variables to determine XGBoost input variables.

2.2 Particle Swarm Optimization

PSO algorithm is a kind of evolutionary algorithm based on swarm intelligence activity, each potential solution in the PSO algorithm is considered a point or particle, all potential solutions group into particle swarm, each particle has the velocity and position of these two properties, velocity on behalf of the particle movement speed, position represents the direction of the particle movement. Each particle searches for the optimal solution separately in the n-dimensional search space and records it as the current individual extremum, then shares the individual extremum with other particles in the whole particle swarm. The optimal individual extremum is the current global optimal solution of the whole particle swarm.

All particles in the particle swarm adjust their velocity and position according to the current individual extremum and the current global optimal solution shared by the whole particle swarm. The PSO algorithm [24] randomly initializes the velocity and position of particles in the search space. Then define the fitness function, generate the global optimal solution by eradicating the individual optimal solution of each particle, and then compare the current global optimal with the historical global optimal to determine whether to update the global optimal. The update of the velocity and position of each particle can be expressed as:

$$V_{id} = \omega V_{id} + C_1 random(0,1)(P_{id} - X_{id}) + C_2 random(0,1)\left(P_{gd} - X_{id}\right) \tag{8}$$

$$X_{id} = X_{id} + V_{id} \tag{9}$$

In the formula, C_1 and C_2 are individual and global learning factors, P_{id} represents the d-th dimension of individual extreme value of the i-th particle, and P_{gd} represents the d-th dimension of global optimal solution. ω is the inertia factor, which is used to adjust the global optimization performance and local optimization performance. The linear decreasing strategy used in this study can be expressed as follows:

$$\omega = \frac{\omega_{max} + (iter - iter_i) \times (\omega_{max} - \omega_{min})}{iter} \tag{10}$$

where $iter$ is the maximum iteration, $iter_i$ is the current iteration, and ω_{max} and ω_{min} are the maximum and minimum values of ω respectively.

Table 1: Parameters setting of the PSO algorithm

Parameters	Value
Particle numbers	80
Maximum number of iterations	200
Local learning factor C1	2
Local learning factor C2	2
Decreasing range of inertia weight	(0.3,0.9)

2.3 Performance Evaluation Measures

Four evaluation measures were selected to indicate the performance of the ET_0 estimation models.

Mean Absolute Error (MAE) is:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|\left(y_i - \hat{y_i}\right)\right| \tag{11}$$

Mean Squared Error (MSE) is:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y_i}\right)^2 \tag{12}$$

Root Mean Squared Error (RMSE) is:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y_i}\right)^2} \tag{13}$$

R-Squared (R^2) is:

$$R^2 = 1 - \frac{\sum_i \left(\hat{y_i} - y_i\right)^2}{\sum_i \left(\overline{y_i} - y^i\right)^2} \tag{14}$$

In the above formula, $\hat{y_i}$ is the predicted value, y_i is the true value, and $\overline{y_i}$ is the average value. MAE can reflect the actual situation of the predicted value error. MSE is the expected value of the square of the difference between the modeled value and the observed value. It can evaluate the degree of the data change, and the smaller value of the MSE, the better accuracy of the prediction model. RMSE is the arithmetic square root of MSE. R^2 can eliminate the influence of dimension on the evaluation measure.

2.4 PSO-XGBoost Model

In this study, XGBoost model was used as the basic algorithm for ET_0 estimation model. Parameter optimization of XGBoost model is the key to model construction. Therefore, the idea of this study on PSO-XGBoost model is to optimize the parameters of XGBoost by using PSO algorithm, and then to use the optimized XGBoost model for model fitting.

Six important parameters for tree booster in the XGBoost model were selected for optimization, including: learning rate (eta), max_depth, min_child_weight, min_split_loss (gamma), subsample and colsample_bytree. Tab. 2 shows the details of each parameter.

Table 2: Information of main parameters of XGBoost model

Parameters	Default value	Range	Explanations
eta	0.3	[0, 1]	Step size shrinkage used in update to prevents overfitting.
max_depth	6	[0, ∞]	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
min_child_weight	1	[0, ∞]	Minimum sum of instance weight needed in a child.
gamma	0	[0, ∞]	Minimum loss reduction required to make a further partition on a leaf node of the tree.
subsample	1	(0, 1]	Subsample ratio of the training instances.
colsample_bytree	1	(0, 1]	Subsample ratio of columns when constructing each tree.

Fig. 1 shows the technical flow chart of the PSO-XGBoost model. The main process of building the PSO-XGBoost model is as follows:

Process and convert the data first. The multi-source data collected by solar greenhouse equipment are further processed to form input variables, which are then converted into tabular form.

The PSO algorithm optimizes the parameters of XGBoost. The optimization target is six parameters of XGBoost, so each particle of PSO is a six-dimensional vector, in which each dimension corresponds to the optimal solution of one XGBoost parameter.

Considering that XGBoost is used to solve the regression problem in this study, we set MSE as the objective optimization function of PSO, and the fitness of the i-th particle at time t can be expressed as:

$$F_{i(t)} = \left(P_{i(t)} \rightarrow XGBoost\big|_{\text{training set}}\right)_{[\text{metric=MSE}]} \tag{15}$$

The local optimal value of the i-th individual particle at time t can be expressed as:

$$Pbest_{i(t)} = max\left(F_{i(j)}\right), 0 \le j \le t \tag{16}$$

The global optimal value of the global m-individual particle at time t can be expressed as:

$$Gbest_{(t)} = max\left(Pbest_{k(t)}\right), 1 \le k \le m \tag{17}$$

Finally, according to the target error value or the maximum number of iterations, the model training is terminated to obtain the optimized parameter value of XGBoost.

Generalized performance evaluation in the estimating model. The test data set was substituted into PSO-XGBoost estimation model to get ET_0 value of model estimation. The accuracy of the model was evaluated by comparing with reference ET_{0i} value.

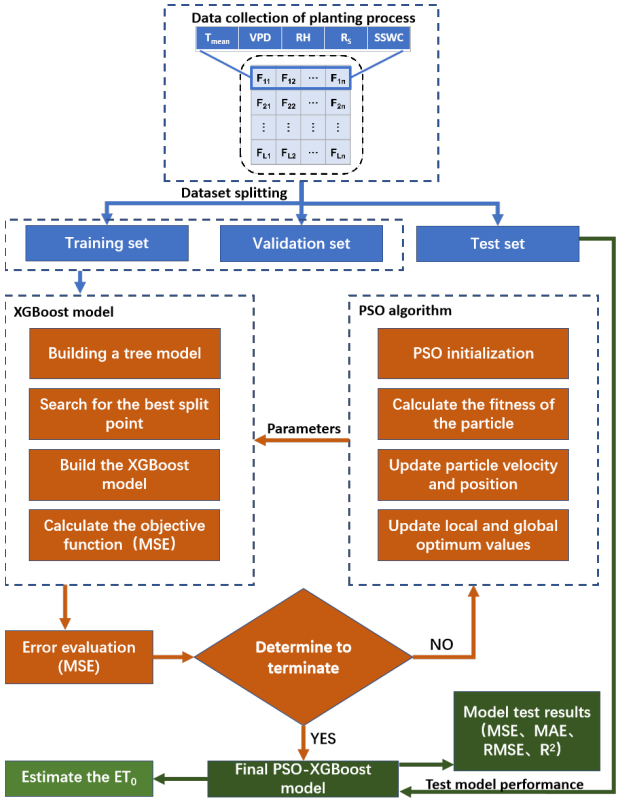

Figure 1: The construction process of the PSO-XGBoost model for estimating the ET_0 of the solar greenhouse

3 Data Source and Processing

3.1 Study Area and Data Acquisition

The experiment of this study was carried out from 2018 to 2019. The greenhouse is located in the precision agriculture demonstration base of national agricultural information technology research center in Xiaotangshan Town, Changping District, Beijing, China (116°34′–117°00′ E, 40°00′–40°21′ N). The length, span and height of the greenhouse are 30 m, 6.5 m and 3 m. The outdoor annual average temperature is 10–13°C, the daily average sunshine hours are 6.5–8.5 h, the total annual radiation is 5413 MJ/m^2, the annual sunshine hours are 2700.3 h, the frost-free period is about 186-200 d, the annual average rainfall is 602.2 mm, and the groundwater depth is 10 m. The terrain of the experimental area is flat, among which, the soil is tidal soil with medium fertility, which is typical in the North China Plain. The meteorological data of this experiment were collected from a micro weather station located in the middle of the greenhouse, 2 m above the ground, and collected hourly. The measured data included air temperature (°C), relative humidity (%) and net surface radiation (MJ/(m^2·d)). The surface soil moisture content (%) of 0–20 cm was collected by the soil moisture content sensor in the greenhouse and collected every hour. The planting crop of this experiment is tomato. The spring crop was planted from March 20, 2018 to July 2, 2018, and the autumn crop was planted from August 24, 2018 to January 3, 2019.

3.2 Data Processing and Analysis

We processed the data of meteorological stations, and obtained the daily average air temperature (T_{mean}, °C) and daily average relative humidity (RH, %) by averaging the hourly data, and obtained the daily total surface net radiation (RS, MJ/(m^2·d)) by adding the hourly net surface radiation. The average hourly surface soil moisture content collected by the soil moisture sensor was converted to the daily average surface soil water content (SSWC, %). In addition, T_{mean} and RH were used to further calculate the saturated vapor pressure difference (VPD) according to the Eq. (18).

$$VPD = 0.61078 \times e^{\frac{17.27 \times T_{mean}}{T_{mean}+237.3}} \times (1 - RH) \tag{18}$$

We defined the reference evapotranspiration of solar greenhouse as ET_{0i}, and calculated ET_{0i} according to the Penman-Monteith formula modified by [3], and the formula is:

$$ET_{0i} = \frac{0.408\Delta(R_n - G) + \gamma\frac{1713}{T+273}(e_s - e_a)}{\Delta + 1.64\gamma} \tag{19}$$

The relevant parameters in the formula were calculated according to [26]:

$$\Delta = \frac{2504 \cdot exp\left(\frac{17.27T}{T+237.3}\right)}{(T+237.3)^2} \tag{20}$$

$$e_s = \frac{e_s(T_{max}) + e_s(T_{\min})}{2} \tag{21}$$

$$e_s(T_{\max/min}) = 0.6108 \cdot \exp\left(\frac{17.27T_{\max/min}}{T_{\max/\min}+237.3}\right) \tag{22}$$

$$e_a = \frac{e_s(T_{min})\frac{RH_{max}}{100} + e_s(T_{max})\frac{RH_{min}}{100}}{2} \tag{23}$$

where ET_{0i} as the reference evapotranspiration of solar greenhouse, mm/d; Δ is the tangent slope of the temperature saturated vapor pressure curve at air temperature T, kPa/°C; R_n is net radiation, MJ/(m^2·d); G is soil heat flux, MJ/(m^2·d); T is the average daily temperature at the height of 2 m above the ground, °C; $T_{\max/min}$ is the daily maximum/minimum temperature at the height of 2 m above the ground; r is the constant of wet and dry meter; e_s is the average saturated vapor pressure, kPa; e_a is the actual water vapor pressure, kPa, RH_{max} and RH_{min} are the daily maximum and minimum relative humidity respectively, %.

Fig. 2 shows the changes of ET_{0i} during the spring and autumn crop planting process. Fig. 3 shows the linear correlation between ET_{0i} and T_{mean}, RH, VPD, RS and SSWC. The linear correlation between ET_{0i} and RS, VPD and RH is relatively high. These three parameters can be used to build a model to obtain better prediction ability, and combining with other indicators can further improve the prediction accuracy of the model.

Tab. 3 performs statistical analysis on the spring crop, autumn crop, and mixed data, showing the maximum (Max), minimum (Min), average (Mean), and standard deviation (SD) of T_{mean}, RH, VPD, RS, SSWC and ET_{0i}.

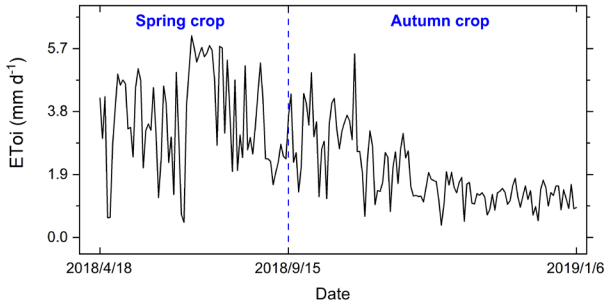

Figure 2: The variation of ET_{0i} in greenhouse during planting of spring and autumn crop

The data obtained in this study included the planting process data of spring and autumn crops, to better verify the model building ability of PSO-XGBoost model and the generalization performance of fitting with the data of a single planting process. All the data was divided into three dataset modes, in which the dataset mode 1 and the dataset mode 2 respectively using spring or autumn crop planting process data as model training data, using the part that did not participate in the modeling data to evaluate the prediction error, the training and validation data in modes 1 and 2 were selected randomly. Dataset Mode 3 merged two planting process data, a random sample 50% was used for modeling, used the other 50% of the data that did not participate in the modeling for testing. In this study, the training set was used for model training, the verification set was used for evaluation of each model iteration to obtain the best model parameters, and the test set was used for error evaluation of the trained model.

To improve the accuracy and fitting speed of model training, the data were normalized. In this study, the normalization method of Min-Max was used to process the characteristic values. The normalization formula of data is as follows:

$$x = \frac{x-min}{max-min} \tag{24}$$

The model training environment was a graphics workstation configured with CPU: Intel(R) Xeon(R) CPU E5-1620 v4@3.50 GHz, GPU: NVIDIA Quadro K2200 and RAM: 32 GB. Ananconda platform was used as the basic platform for model training, XGBoost 1.1.0 was used as the model framework, and the Python version was 3.7.

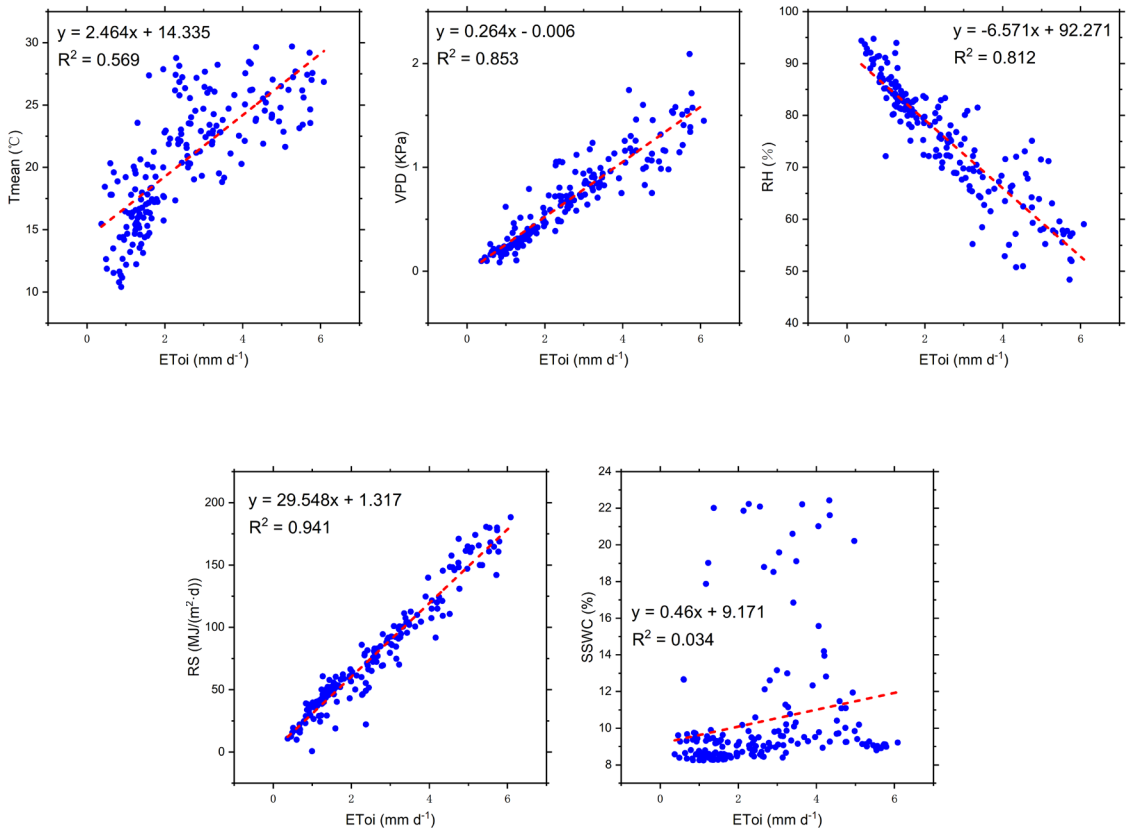

Figure 3: Correlation between ET_{0i} and main greenhouse meteorological and soil parameters

Table 3: Statistical analysis of data in different planting periods

		T_{mean} (℃)	RH (%)	VPD (kPa)	RS (MJ/(m²·d))	SSWC (%)	ET_{0i} (mm/d)
Spring crop	Max	29.69	93.68	2.09	188.33	13.95	6.08
	Min	17.81	48.40	0.13	9.91	8.72	0.45
	Mean	24.93	69.78	0.98	104.45	9.87	3.58
	SD	2.73	10.61	0.39	50.78	1.18	1.47
Autumn crop	Max	26.18	94.75	1.46	179.83	22.42	5.54
	Min	10.41	50.79	0.08	0.73	8.24	0.36
	Mean	17.88	79.05	0.47	59.44	10.65	1.90
	SD	3.74	9.26	0.30	29.87	4.26	1.04
Spring and autumn crops	Max	29.69	94.75	2.09	188.33	22.42	6.08
	Min	10.41	48.40	0.08	0.73	8.24	0.36
	Mean	20.65	75.40	0.67	77.15	10.35	2.56
	SD	4.82	10.78	0.42	45.08	3.41	1.48

Table 4: Three data set partition standards

Dataset mode	Training set	Validation set	Test set
1	80% data of spring crop	20% data of spring crop	100% data of autumn crop
2	80 data of autumn crop	20% data of autumn crop	100% data of spring crop
3	40% data of autumn and spring crops	10% data of autumn and spring crops	50% data of autumn and spring crops

4 Results

4.1 PSO-XGBoost Model Architecture with All Input Variables

We trained three times for each dataset mode, and determined the optimal value of the PSO-XGBoost model parameters by comparing the MSE value of the validation set. Tab. 5 shows the model optimization results. The MSE in all the training results was less than 0.1, and the PSO-XGBoost model was able to fit the data accurately. In addition, the MSE obtained from multiple training of the same dataset mode was relatively stable, with an average MSE of 0.079, 0.017 and 0.028, respectively. The MSE value of dataset mode 2 was better than that of the other two modes.

The appropriate parameter values of the PSO-XGBoost model were obtained through multiple training, and then the estimation error was evaluated. Tab. 6 shows the estimation errors of the verification set and the test set for different modes. It can be seen from the results that the R^2 of the verification set were all higher than 0.9, and the R^2 of the test set were all higher than 0.92. The rank of the goodness of fit of the verification set was: dataset mode 1 > dataset mode 2 > dataset mode 3, while the rank of the goodness of fit of the test set was: dataset mode 3 > dataset mode 1 > dataset mode 2. In addition, MAE and RMSE of dataset mode 3 were higher than the other two.

To further study the influence of single variable on the estimation of ET_{0i} by PSO-XGBoost model, we calculated the relative importance of five input items after the completion of XGBoost training [27]. Fig. 4 shows the relative importance scores of each input variable in the PSO-XGBoost model, it is obvious that the importance of each variable varies greatly in the degree of model construction. The importance scores of RS and VPD were relatively high, with an average relative importance score of 0.38 and 0.36. The average relative importance ranking of each input variable was: RS > VPD > RH > T_{mean} > SSWC.

Table 5: Results of main parameters of XGBoost model optimized by the PSO algorithm

Dataset mode	MSE	max_depth	eta	min_child_weight	gamma	subsample	colsample_bytree
1	0.097	9	0.128	0.705	0.009	0.273	0.831
	0.096	5	0.195	2.629	0.261	0.646	0.694
	0.044	3	0.154	2.928	0.034	0.708	0.685
2	0.015	6	0.062	1.145	0.171	0.725	0.631
	0.016	7	0.153	1.529	0.039	0.667	0.721
	0.021	7	0.146	1.872	0.185	0.861	0.591
3	0.026	6	0.081	0.911	0.159	0.533	0.892
	0.032	8	0.079	2.119	0.279	0.909	0.703
	0.028	3	0.082	1.504	0.019	0.644	0.685

Table 6: The estimation accuracy evaluation of validation set and test set in different dataset modes

Dataset mode	Validation set estimation			Test set estimation		
	MAE	RMSE	R^2	MAE	RMSE	R^2
1	0.163	0.211	0.982	0.199	0.258	0.938
2	0.094	0.124	0.978	0.312	0.411	0.922
3	0.121	0.162	0.973	0.138	0.192	0.984

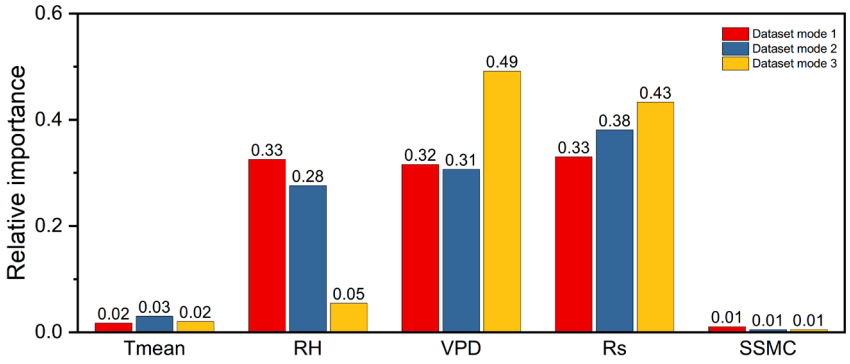

Figure 4: The relative importance of each input variable in different dataset modes

4.2 PSO-XGBoost Model Architecture with Less Input Variables

We conducted experiments on the PSO-XGBoost estimation model with less input variables. According to Fig. 3, we knew the linear correlation between ET_{0i} and other variables. In combination with the results of the relative importance of each input variable (Fig. 4), we designed six combinations of input variables to verify the modeling capability of PSO-XGBoost model with less input variables.

Tab. 7 shows the estimation accuracy evaluation results of models with different input variables. The MSE, MAE and RMSE increased with the less input variables. From the MSE, the difference between the 4 input variables (T_{mean}, RH, VPD, RS) and 3 input variables (RH, VPD, RS) was only 0.002, the average MSE of the 2 input variables was 0.085, and the MSE of the 1 input variable was significantly increased, when VPD was the only input variable, MSE increased by 0.367 compared with 4 input variables.

Fig. 5 shows the goodness of fit between the estimated value and the reference value of the PSO-XGBoost model constructed by different input variables. The R^2 decreased with the decrease of input variables, and when the input variables were only RS, the R^2 was still greater than 0.9.

Table 7: Estimation accuracy evaluation of models with less input variables

	4 Input variables T_{mean}, RH, VPD, RS	3 Input variables RH, VPD, RS	2 Input variables VPD, RS	2 Input variables RH, RS	1 Input variable RS	1 Input variable VPD
MSE	0.042	0.046	0.063	0.107	0.155	0.409
MAE	0.149	0.148	0.183	0.239	0.305	0.453
RMSE	0.206	0.215	0.251	0.328	0.394	0.639

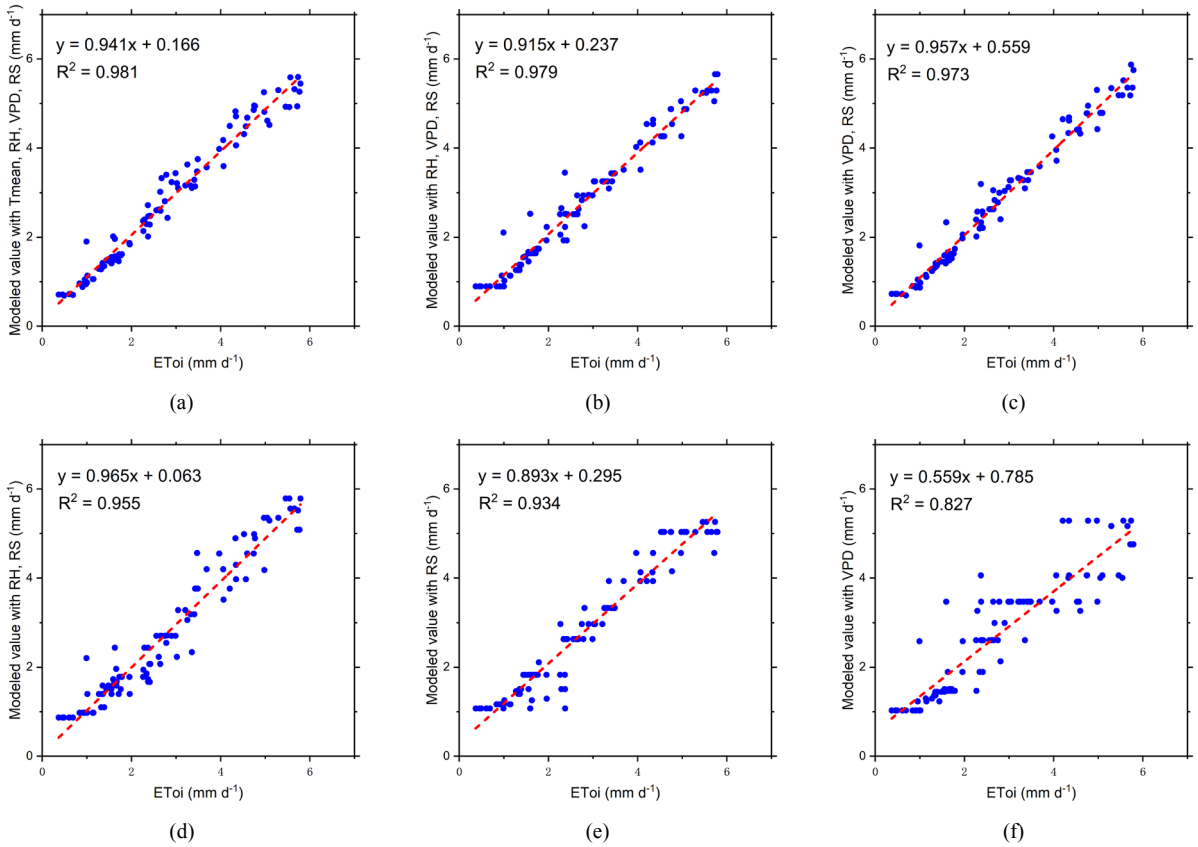

Figure 5: Linear correlation analysis between modeled value of model with different input variables and reference ET_{0i}. (a) 4 Input variables (T_{mean}, RH, VPD, RS); (b) 3 Input variables (RH, VPD, RS); (c) 2 Input variables (VPD, RS); (d) 2 Input variables (RH, RS); (e) 1 Input variable (RS); (f) 1 Input variable (VPD)

5 Discussion

We found that the MSE of multiple training in the same data set was relatively close (Tab. 5), while the error obtained from different training sets was significantly different, indicating that the ability of PSO-XGBoost model to fit ET_{0i} data was relatively stable. However, the model training accuracy of different training sets has obvious differences, which may be caused by the difference of statistical characteristics, Maier et al. [28] pointed out that the training set, test set and verification set used to train the AI model needed to maintain balanced statistical properties to obtain the best model building effect. From Tab. 6, the R^2 of test sets with different dataset modes were all higher than 0.92, indicating that good estimation results could be obtained by training the training model with limited planting period data, the MAE, RMSE and R^2 of test set estimation results of dataset mode 3 were better than the other two modes. This may be because that the mixed data sets could better obtain the data characteristics of different planting periods to achieve more accurate estimates, to obtain better model estimation effect, we should pay more attention to the division and selection of data sets.

According to the results, the linear correlation between each input variable and ET_0 (Fig. 3) was consistent with the relative importance of PSO-XGBoost model (Fig. 4), and the order was: RS > VPD > RH > T_{mean} > SSWC. This result showed that PSO-XGBoost could accurately identify the weight relationship between the input variables, and the research of constructing estimation model with XGBoost in other fields also selected the input items through the relative importance index [29]. In addition, we found that the input variable with the highest correlation with ET_{0i} was RS, while in the studies of [7] and [30], the ET_0 estimation model of farmland considered that temperature was the most critical indicator. This may be because the temperature in the greenhouse was higher than that outside and the temperature difference in the greenhouse was smaller. At the same time, the wind speed in the greenhouse was almost zero, which made the sunlight become the main source of transpiration. Tab. 7 compares the prediction accuracy of models with less input variables. We found that the best results could be obtained by taking all parameters as input. The overall error of test set increases with the decrease of input items, which was consistent with the research conclusion of [7].

To verify the effect of PSO-XGBoost model, we compared other ensemble learning methods based on dataset mode 3, such as Bagging [31], Random Forest [32], CatBoost [33] and AdaBoost [34]. In addition, traditional machine learning methods such as Artificial Neural Network (ANN) [7], Decision Trees (Tree) [35] and K-Nearest Neighbor (KNN) [36] were compared. Tab. 8 shows the comparison results between PSO-XGBoost model and other machine learning models. Through the error evaluation of the estimation results of the verification set and the test set, the estimation accuracy of the optimized PSO-XGBoost model was better than that of other models. The findings of other studies similarly supported that the estimation ability of XGBoost model optimized by PSO algorithm was improved and better estimation effect was obtained [37–38]. The PSO-XGBoost model had the advantages of high prediction accuracy and generalization capability, however, not incorporating the latest bionic optimization algorithms was a drawback of the model.

Table 8: Comparison of estimation accuracy between PSO-XGBoost model and other machine learning models

	Validation set estimation				Test set estimation			
	MSE	**MAE**	**RMSE**	**R^2**	**MSE**	**MAE**	**RMSE**	**R^2**
PSO-XGBoost	0.025	0.111	0.157	0.975	**0.031**	**0.125**	**0.177**	**0.987**
CatBoost	0.028	0.116	0.167	0.970	0.037	0.126	0.193	0.983
Bagging	0.029	0.125	0.172	0.969	0.503	0.152	0.224	0.978
XGBoost	0.031	0.124	0.175	0.969	0.037	0.144	0.193	0.982
ANN	0.034	0.145	0.185	0.964	0.076	0.219	0.276	0.968
AdaBoost	0.043	0.158	0.207	0.956	0.641	0.182	0.253	0.972
Random Forest	0.047	0.135	0.217	0.952	0.521	0.167	0.228	0.978
Tree	0.064	0.178	0.254	0.934	0.098	0.213	0.312	0.958
KNN	0.090	0.221	0.300	0.907	0.096	0.216	0.311	0.959

This study still has some limitations in terms of the selection of optimization algorithms and the richness of data. In the future research, we plan to collect data in different greenhouses to compare the impact of greenhouse and crop differences on model estimation, and adopt some new and powerful bio-inspired optimization algorithms.

6 Conclusion

In this study, a novel PSO-XGBoost model was proposed to estimate the ET_0 of solar greenhouse. PSO algorithm was used to optimize the XGBoost model parameters for optimal model performance. Meteorological and soil moisture data of the two-crop process in the solar greenhouse located in Beijing, China from 2018 to 2019 were selected as the basis. After data processing, five basic variables (T_{mean}, VPD, RH, RS and SSWC) were obtained. The accuracy of model estimation was evaluated by comparing ET_0 calculated by improved P-M equation based on solar greenhouse.

The experimental results showed that the training results of PSO-XGBoost model were relatively stable, and the correlation coefficients (R^2) between the estimated value and the reference value of the test set were higher than 0.92 under the different data set division types of dataset mode. The order of relative importance of each input variable in model construction was: RS > VPD > RH > T_{mean} > SSWC. Different input variables had great influence on the estimation ability of PSO-XGBoost mode. When all parameters were input, the model estimation could obtain the highest accuracy performance. As the number of input variables decreased, the model estimation error increased. Among them, the RMSE values estimated by the test sets of 4 input variables (T_{mean}, Rh, VPD, RS), 3 input variables (RH, VPD, RS) and 2 input variables (VPD, RS) were all less than 0.3, which could provide high estimation accuracy. Especially when only RS was used as the input variable, the RMSE of the model was 0.39, which was of reference significance for the situation that the estimation accuracy requirements were relatively wide. In addition, this study further compared PSO-XGBoost model with other integrated models (CatBoost, Bagging, XGBoost, AdaBoost and Random Forest) and classical machine learning models (Artificial Neural Network, Decision Trees and K-Nearest Neighbor), demonstrating that PSO-XGBoost model had the best model fitting ability and generalization performance. The findings of the study are of great significance to the development of ET_0 estimation model in greenhouse environment.

Funding Statement: This work was supported by Chongqing Yubei District Science and Technology Plan Project (2020-30), Ability Construction Project of Beijing Academy of Agriculture and Forestry Sciences (KJCX20200430, KJCX20180703), and Agricultural Sci-tech Extension of Beijing Academy of Agriculture and Forestry Sciences (20200401).

Conflicts of Interest: No potential conflict of interest was reported by the authors.

References

[1] J. Doorenbos and W. O. Pruitt, "Crop water requirements. FAO irrigation and drainage paper 24," *Land and Water Development Division*, vol. 144, 1977.

[2] R. G. Allen, L. S. Pereira, D. Raes and M. Smith, "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56," *Fao, Rome*, vol. 300, no. 9, D05109, 1998.

[3] J. Wang, H. Cai, H. Li and X. Chen, "Study and Evaluation of the Calculation Methods of Reference Crop Evapotranspiration in Solar-heated Greenhouse," *Journal of Irrigation and Drainage*, vol. 6, pp. 11–14, 2006.

[4] M. D. Fernández, S. Bonachela, F. Orgaz, R. Thompson, C. López *et al.,* "Measurement and estimation of plastic greenhouse reference evapotranspiration in a Mediterranean climate," *Irrigation Science*, vol. 28, no. 6, pp. 497–509, 2010.

[5] R. Qiu, J. Song, T. Du, S. Kang, L. Tong *et al.,* "Response of evapotranspiration and yield to planting density of solar greenhouse grown tomato in northwest China," *Agricultural Water Management*, vol. 130, pp. 44–51, 2013.

[6] L. Wu, Y. Peng, J. Fan and Y. Wang, "Machine learning models for the estimation of monthly mean daily reference evapotranspiration based on cross-station and synthetic data," *Hydrology Research*, vol. 50, no. 6, pp. 1730–1750, 2019. DOI 10.2166/nh.2019.060.

[7] V. Z. Antonopoulos and A. V. Antonopoulos, "Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables," *Computers and Electronics in Agriculture*, vol. 132, pp. 86–96, 2017.

[8] J. Shiri, Ö. Kişi, G. Landeras, J. J. López, A. H. Nazemi *et al.,* "Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain)," *Journal of Hydrology*, vol. 414–415, pp. 302–316, 2012. DOI 10.1016/j.jhydrol.2011.11.004.

[9] A. Pour-Ali Baba, J. Shiri, O. Kisi, A. F. Fard, S. Kim *et al.,* "Estimating daily reference evapotranspiration using available and estimated climatic data by adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN)," *Hydrology Research*, vol. 44, no. 1, pp. 131–146. 2013. DOI 10.2166/nh.2012.074.

[10] H. Sanikhani, O. Kisi, E. Maroufpoor and Z. M. Yaseen, "Temperature-based modeling of reference evapotranspiration using several artificial intelligence models: Application of different modeling scenarios," *Theoretical and Applied Climatology*, vol. 135, no. 1–2, pp. 449–462, 2019.

[11] L. Wu and J. Fan, "Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration," *PLoS One*, vol. 14, no. 5, 2019.

[12] R. Polikar, "Ensemble learning," in *Ensemble Machine Learning*, Springer, pp. 1–34, 2012.

[13] J. Fan, W. Yue, L. Wu, F. Zhang, H, Cai *et al.,* "Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China," *Agricultural and Forest Meteorology*, vol. 263, pp. 225–241, 2018.

[14] G. Huang, L. Wu, X. Ma, W. Zhang, J, Fan *et al.,* "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *Journal of Hydrology*, vol. 574, pp. 1029–1041, 2019.

[15] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu and W. Zeng, "Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data," *Agricultural Water Management*, vol. 225, 105758, 2019. DOI 10.1016/j.agwat.2019.105758.

[16] A. Anand and L. Suganthi, "Hybrid GA-PSO optimization of artificial neural network for forecasting electricity demand," *Energies*, vol. 11, no. 4, pp. 728, 2018.

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. of the 22nd ACM Sigkdd Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[18] D. Nielsen, "Tree boosting with xgboost-why does xgboost win "every" machine learning competition?," Master's Thesis, NTNU, 2016.

[19] Z. Yin, X. Wen, Q. Feng, Z. He, S. Zou *et al.,* "Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area," *Hydrology Research*, vol. 48, no. 5, pp. 1177–1191, 2017. DOI 10.2166/nh.2016.205.

[20] Y. Han, J. Wu, B. Zhai, Y. Pan, G. Huang *et al.,* "Coupling a bat algorithm with XGBoost to estimate reference evapotranspiration in the arid and semiarid regions of China," *Advances in Meteorology*, vol. 2019, pp. 1–16, 2019. DOI 10.1155/2019/9575782.

[21] T. Liu, Y. Ding, X. Cai, Y. Zhu and X. Zhang, "Extreme learning machine based on particle swarm optimization for estimation of reference evapotranspiration," in *2017 36th Chinese Control Conference (CCC)*, 2017, pp. 4567–4572. DOI 10.23919/ChiCC.2017.8028076.

[22] L. Wu, G. Huang, J. Fan, X. Ma, H. Zhou *et al.,* "Hybrid extreme learning machine with meta-heuristic algorithms for monthly pan evaporation prediction," *Computers and Electronics in Agriculture*, vol. 168, 105115, 2020. DOI 10.1016/j.compag.2019.105115.

[23] L. Wu, H. Zhou, X. Ma, J. Fan and F. Zhang, "Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio-inspired optimization algorithms: Application in contrasting climates of China," *Journal of Hydrology*, vol. 577, 123960, 2019. DOI 10.1016/j.jhydrol.2019.123960.

[24] F. Marini and B. Walczak, "Particle swarm optimization (PSO). A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 153–165, 2015.

[25] Q. Li, Z. Qiu and X. Zhang, "Eigenvalue analysis of structures with interval parameters using the second-order Taylor series expansion and the DCA for QB," *Applied Mathematical Modelling*, vol. 49, pp. 680–690, 2017. DOI 10.1016/j.apm.2017.02.041.

[26] R. L. Huffman, D. D. Fangmeier, W. J. Elliot, S. R. Workman and G. O. Schwab, *Soil and Water Conservation Engineering*. American Society of Agricultural and Biological Engineers St. Joseph, 2011.

[27] H. J. Lu, N. Zou, R. Jacobs, B. Afflerbach, X. G. Lu *et al.,* "Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion," *Computational Materials Science*, vol. 169, 109075, 2019. DOI 10.1016/j.commatsci.2019.06.010.

[28] H. R. Maier, A. Jain, G. C. Dandy and K. P. Sudheer, "Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions," *Environmental Modelling & Software*, vol. 25, no. 8, pp. 891–909, 2010.

[29] K. Song, F. Yan, T. Ding, L. Gao and S. Lu, "A steel property optimization model based on the XGBoost algorithm and improved PSO," *Computational Materials Science*, vol. 174, 109472, 2020.

[30] A. Laaboudi, B. Mouhouche and B. Draoui, "Neural network approach to reference evapotranspiration modeling from limited climatic data in arid regions," *International Journal of Biometeorology*, vol. 56, no. 5, pp. 831–841, 2012.

[31] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[32] A. Liaw and M. Wiener, "Classification and regression by random Forest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[33] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, pp. 6638–6648, 2018.

[34] T. Hastie, S. Rosset, J. Zhu and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[35] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[36] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 1883, 2009.

[37] L. T. Le, H. Nguyen, J. Zhou, J. Dou and H. Moayedi, "Estimating the heating load of buildings for smart city planning using a novel artificial intelligence technique PSO-XGBoost," *Applied Sciences*, vol. 9, no. 13, pp. 2714, 2019.

[38] H. Jiang, Z. He, G. Ye and H. Zhang, "Network intrusion detection based on PSO-xgboost model," *IEEE Access*, vol. 8, pp. 58392–58401, 2020. DOI 10.1109/ACCESS.2020.2982418.