



Research article

Construction of a photosynthetic rate prediction model for greenhouse strawberries with distributed regulation of light environment

Xinyan Chen, Zhaohui Jiang*, Qile Tai, Chunshan Shen, Yuan Rao and Wu Zhang

School of Information and Computer Science, Anhui Agricultural University, Hefei, 230036, China

* **Correspondence:** Email: jiangzh@ahau.edu.cn.

Abstract: In winter and spring, for greenhouses with larger areas and stereoscopic cultivation, distributed light environment regulation based on photosynthetic rate prediction model can better ensure good crop growth. In this paper, strawberries at flowering-fruit stage were used as the test crop, and the LI-6800 portable photosynthesis system was used to control the leaf chamber environment and obtain sample data by nested photosynthetic rate combination experiments under temperature, light and CO₂ concentration conditions to study the photosynthetic rate prediction model construction method. For a small-sample, nonlinear real experimental data set validated by grey relational analysis, a photosynthetic rate prediction model was developed based on Support vector regression (SVR), and the particle swarm algorithm (PSO) was used to search the influence of the empirical values of parameters, such as the penalty parameter C , accuracy ε and kernel constant g , on the model prediction performance. The modeling and prediction results show that the PSO-SVR method outperforms the commonly used algorithms such as MLR, BP, SVR and RF in terms of prediction performance and generalization on a small sample data set. The research in this paper achieves accurate prediction of photosynthetic rate of strawberry and lays the foundation for subsequent distributed regulation of greenhouse strawberry light environment.

Keywords: greenhouse; light environment; distributed regulation; photosynthetic rate model; PSO-SVR; strawberry

1. Introduction

Greenhouses create a good growing environment for crops and meet the growing demand of

society for yield, quality, variety, season, and origin of fruits and vegetables. Photosynthesis is the material basis of planting production, and it is affected by a variety of environmental factors, such as temperature, humidity, CO₂ concentration, and light intensity. The influence of the light environment is especially important for facility cultivation in the middle and high latitudes of the northern hemisphere [1], especially in winter and spring and in some regions during the rainy season, when low temperature, low light, and high humidity make it difficult to optimize the photosynthesis of crops, resulting in longer crop growth cycles, lower yields, and increased probability of pests, diseases, and other problems[2]. Y. Kong took peas as an example to prove that winter supplemental light has a great influence on the improvement of crop yield and quality [3]. Taking cherry tomatoes as an example, Y.C. Xu demonstrated that adding LED supplementary light under bright conditions can significantly increase the growth rate of crops [4]. Therefore, it is urgent to supplement light for protected crops to improve their photosynthetic effect.

Research on light environment regulation for facility cultivation is divided into two parts. The first is supplemental light equipment and control aspects, including supplemental light morphology improvement and the selection of fixed positions. Y.J. Zheng proved through experiments that different blue light intensities can regulate the growth of pakchoi in greenhouse [6]. L. Xu installed LED lights at the top, middle and bottom of plant canopy to provide additional photosynthetic effective radiation [7]. O.D. Palmitessa, taking tomato as an example, summarizes the present situation of LED application and development, focusing on the requirements related to latitude [8]. F.F. He put forward an expert system technology database to store the empirical values of real-time light intensity values, so as to identify the best position where LED lights need to be turned on and the best position where LED arrays lit by driving circuits need to be located, identify the number of LED supplementary lights and solve the problem of light intensity optimization [9]. C.Y. Li constructed the database of greenhouse light intensity optimization control model under limited light resources, and optimized the distribution of light resources [10]. As well as the optimization of greenhouse construction with the goal of reducing costs and maximizing economic benefits, which provide the physical basis for light environment regulation. M. Carlini analyzed the multi-parameters of the solar greenhouse to obtain the best development conditions for the growth of the sample crops [11]. N. Choab discussed the performance data of greenhouse covering materials and the comparison of several cladding materials. The selection process of greenhouse shapes and orientations under different climatic conditions was introduced, and several shapes and orientations were compared [12]. J.T. Chen analyzed and compared six typical greenhouse shapes (uniform span, uneven span, oval, arch, zigzag and vinery) common in southern China, in order to select the type that is most suitable for capturing solar radiation to the greatest extent [13]. R. Liu introduced a new transient greenhouse model, which can be used for the management and control of solar greenhouse conditions in China, and as a decision support tool for greenhouses [14]. Another more important aspect is the light environment regulation model or strategy. It is an important research topic to make the best use of resources while meeting the needs of crop growth. The existing methods are generally based on empirical light filling or adjustment of a given threshold range, so it is difficult to achieve high cost performance. Because the growth of crops can be evaluated by photosynthetic rate, the regulation method based on photosynthetic rate monitoring or prediction has gradually become the mainstream. The other more important aspect is the light environment regulation model or strategy, and meeting crop growth needs while maximizing resource utilization has become an important research topic. The existing methods are generally based on empirical light supplementation or regulation based on a given threshold range, which makes it difficult

to achieve high cost performance. Since a crop growth condition can be evaluated by its photosynthetic rate, regulation methods based on photosynthetic rate monitoring or prediction are gradually becoming mainstream. T. Liu et al. modelled and achieved the accurate prediction of the photosynthetic rate of tomatoes using the LSSVM for tomato seedlings [15]. P.P. Xin uses ACO-SVM model [16] and ANN model [17] to optimize and control the light environment of cucumber. J. Hu established the photosynthetic rate model of cucumber seedling based on BP neural network and SVR [18]. J.T. Ding used ABPM-BP model to predict and adjust the growth environment parameters of *Dendrobium candidum* to improve its yield [19]. Y.W. Liu et al. established a greenhouse environmental climate prediction model based on an LSTM model and validated it with tomatoes, cucumbers, and peppers [20]. X.Y. Zhang et al. measured multiple prediction methods based on poplar leaf data and determined the best performance of the prediction model established using the XGBoost algorithm [21]. K.P. Zheng et al. analysed the association between photorespiration and multiple factors in cucumber leaves and achieved the best fit with the XGBoost algorithm [22]. D.H. Jung predicted temperature, humidity, and CO₂ with RNN-LSTM [23].

Because of the large area of a greenhouse and the stereoscopic cultivation, distributed light environment regulation is needed. We propose a greenhouse light environment regulation scheme based on distributed photosynthetic monitoring, as shown in Figure 1. The greenhouse light environment regulation scheme in this study relies on a main control terminal and multiple monitoring terminals. Each monitoring terminal includes temperature sensors, CO₂ sensors, photosynthetic photon flux density (PPFD) sensors, and control circuits to automatically monitor real-time environmental parameters at multiple locations in the greenhouse and transmit data to the main control terminal via a wireless communication module. At the main control terminal, a corresponding model is invoked to determine the light saturation point, and the location where the light intensity needs to be changed is regulated through a fill light regulation circuit to complete the feedback according to the result of the search.

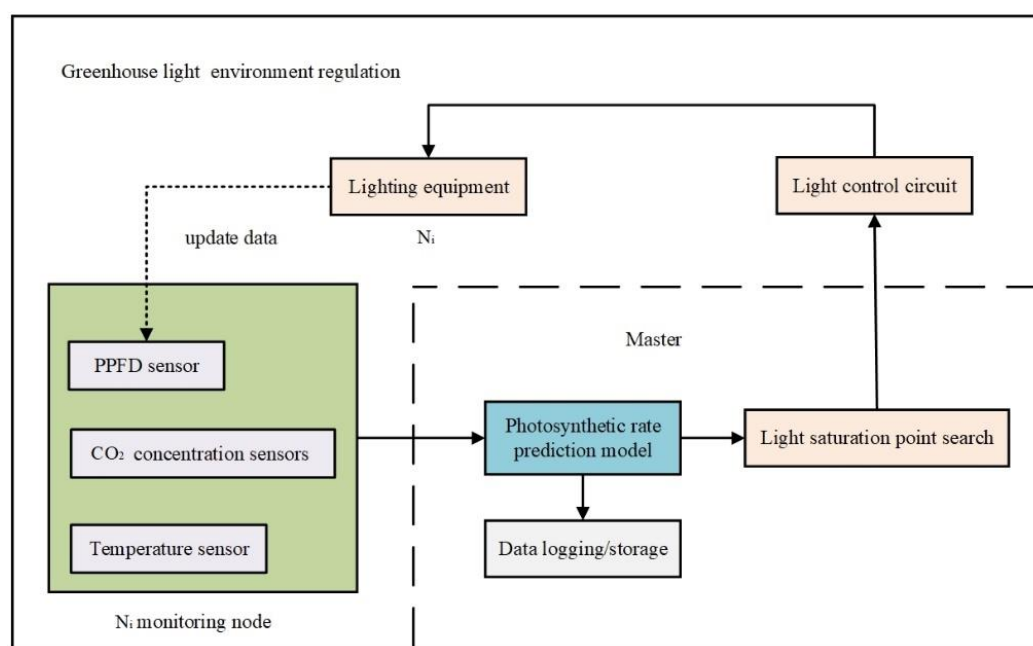


Figure 1. Greenhouse light regulation solution for distributed photosynthesis monitors.

Strawberries are one of the most popular fruits in the world, and the greenhouse strawberry cultivation industry is developing rapidly and on a large scale [24]. In this study, we take strawberries as an example to study the method of photosynthetic rate prediction model construction.

Choosing a modeling approach with a good fit is crucial to ensure model usability [25]. Common models include multiple linear regression (MLR), random forest (RF), back propagation neural network (BPNN), and support vector regression (SVR). In this paper, a support vector machine model based on particle swarm optimization (PSO-SVR) is used for photosynthetic rate prediction to improve the prediction performance and lay the foundation for subsequent distributed regulation of the greenhouse light environment.

2. Data collection and preprocessing

2.1. *Experimental data acquisition*

In July 2021, four good, uniformly growing strawberry plants ('Hong yan') at flowering-fruit stage were obtained from the plastic shed in Changfeng County, Hefei City, Anhui Province. We transplanted them to a $2 \times 4 \times 2.5 \text{ m}^3$ greenhouse in the Science Building of Anhui Agricultural University, Hefei, Anhui Province, where the soil and air conditions were essentially the same as their original growing environment. During the trial period, fertilisation, pest control, and environmental control measures were carried out according to the Anhui local strawberry production standards provided by the Anhui Standardisation Information Service Platform.

The data for this experiment was collected with the LI-6800 portable photosynthesis system. The LI-6800 has a CO_2 injector, a light control module, and a temperature control module, so it can be automatically programmed to simulate a greenhouse under various environmental conditions in a dedicated leaf chamber module [26]. This type of measurement avoids the need for extensive adjustment of the greenhouse environment and reduces the operational difficulty of the test. Therefore, we used the LI-6800 portable photosynthesis system as the core for building the basic test platform and conducted a multi-factor correlation photosynthesis rate combination test to quantitatively obtain photosynthesis rate related parameters under multi-factor nested conditions, verify the key environmental factors affecting photosynthesis rate in strawberry, and provide refined data for building a multi-factor correlation photosynthesis rate prediction model.

The leaf bin environment, including temperature, CO_2 volume concentration and photon flux density (PPFD), was controlled using several sub-modules attached to the LI-6800 photosynthesis system from 09:00 to 11:30 and 14:30 to 17:00 each day, while the net photosynthetic rate of the samples was measured. The data were collected strictly according to the guidelines of control variables. On-demand control was performed around fresh strawberry leaves using several sub-modules included in the LI-6800 Photosynthesis System. Data were collected strictly following the guidelines of the control variables. Among them, the temperature was controlled at 18, 20, 23, 26, 29, and 32 °C in six gradients; the CO_2 volume ratio was set at 200, 300, 400, 600, 800, 1000, 1200, and 1500 $\mu\text{mol/mol}$ in eight concentrations; and the PPFD was set at eleven gradients of 100, 200, 300, 400, 600, 800, 1000, 1200, 1500, 1800, and 2000 $\mu\text{mol/m}^2\text{s}$. The humidity module was set the relative humidity inside the leaf chamber to between 50% and 60%. After each adjustment of the environmental parameters, the leaves to be tested were acclimatised in the leaf chamber for 30 minutes before recording photosynthetic rates to obtain stable data.

To improve robustness, a leaf was randomly selected on each of the four strawberry plants. These leaves were used for the environmental control and photosynthetic rate tests described above, and 2112 sets of experiments were performed. After removing the few missing and abnormal values due to human oversight in the experiments, the overall error was found to be small, so the data from one randomly selected strawberry plant sample per set of environmental conditions resulted in a sample data set with 528 valid data sets.

2.2. Experimental parameter correlation verification

There are many environmental parameters that affect the photosynthetic rate, and the selection of predictors is necessary to construct a suitable prediction model. To exclude the arbitrariness of selecting subjective choices as much as possible and to combine relevant information, we used grey correlation analysis to verify the correlation between the predictors (environmental parameters) and the corresponding photosynthetic rates of the constituent sample data [27]. Grey relational analysis (GRA) does not require a large sample size, is relatively small in computational effort and can effectively analyze the correlation between predictors and photosynthetic rate. The specific steps are as follows:

- 1) The photosynthetic rate is noted as the reference sequence, denoted as

$$X_0 = [X_0(1), X_0(2), \dots, X_0(n)] \quad (1)$$

- 2) The predictors in the data set are set as a comparison series, denoted as

$$(X_1, X_2, \dots, X_m) = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_m(1) \\ x_1(2) & x_2(2) & \cdots & x_m(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(n) & x_2(n) & \cdots & x_m(n) \end{bmatrix} \quad (2)$$

where n is the length of the series and m is the numbers of predictors. $X_i = (x_i(1), x_i(2), \dots, x_i(n))^T, i = 1, 2, \dots, m$.

- 3) Data nondimensionalization. Commonly used methods for dimensionless processing are the extreme value method, the initial value method, and the mean method. Here, the extreme value method is used to standardize the statistical data, and the values are controlled to be between $[0, 1]$ to eliminate the influence of the magnitude and size of each variable. The extreme value method formula is

$$x'_p(k) = \frac{x_p(k) - \min_{1 \leq i \leq n} x_p}{\max_{1 \leq i \leq n} x_p - \min_{1 \leq i \leq n} x_p} \quad (3)$$

where $x'_i(j)$ is the normalized value of $x_i(j)$, p is the evaluation object, k is the evaluation parameter, $p = 1, 2, \dots, m$, $k = 1, 2, \dots, n$, and $\max_{1 \leq i \leq n} x_p$ and $\min_{1 \leq i \leq n} x_p$ denote the maximum and minimum values of the sequence p , respectively. The dimensionless data are as follows:

$$(X'_0, X'_1, \dots, X'_m) = \begin{bmatrix} x'_0(1) & x'_1(1) & \cdots & x'_m(1) \\ x'_0(2) & x'_1(2) & \cdots & x'_m(2) \\ \vdots & \vdots & \ddots & \vdots \\ x'_0(n) & x'_1(n) & \cdots & x'_m(n) \end{bmatrix} \quad (4)$$

4) The correlation coefficient is calculated. The number of grey correlation coefficients between the corresponding elements of each comparison sequence and the reference sequence are calculated as follows:

$$\xi_p(k) = \frac{\min_p \min_k |x'_0(k) - x'_p(k)| + \zeta \max_p \max_k |x'_0(k) - x'_p(k)|}{|x'_0(k) - x'_p(k)| + \zeta \max_p \max_k |x'_0(k) - x'_p(k)|} \quad (5)$$

where $\zeta (0 < \zeta < 1)$ is the resolution coefficient, and is usually set to 0.5; $\xi_p(k)$ is the number of grey correlation coefficients between the p -th evaluation object and the k -th parameter.

5) The degree of relevance is calculated. The data are substituted into Eq. (6) and the correlation value between each predictor and the photosynthetic rate is obtained.

$$\gamma_q = \frac{1}{n} \sum_{k=1}^n \xi_p(k) \quad (6)$$

As seen in Figure 2, when the T, CO₂ volumetric concentration and PPFD were used as predictors, the resulting γ_q were all greater than 0.6, which correlated strongly with the crop photosynthetic rate. Clearly, the three environmental parameters selected in this paper can have a large impact on the photosynthetic rate when they are adjusted.

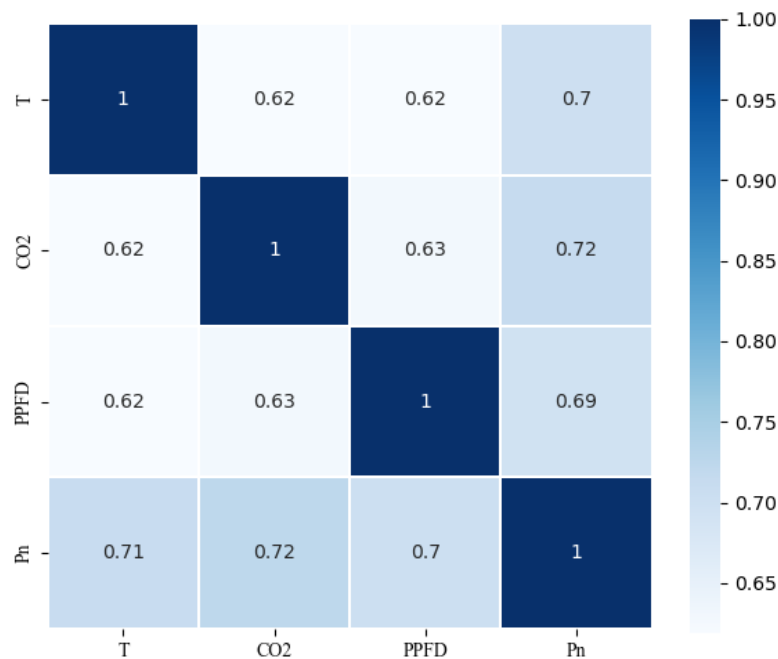


Figure 2. Correlation analysis diagram of selected environmental parameters and the photosynthetic rate.

3. Methods

3.1. PSO-SVR prediction model

3.1.1. Support vector regression (SVR)

Support vector machine (SVM) is a very convenient binary classification model originating from statistics that usually solves classification and regression problems by transforming the original space into a high-dimensional feature space. It may be called Support vector regression (SVR) in the treatment of regression problems. In this study, the photosynthetic rate is nonlinearly related to environmental parameters such as temperature, CO₂ volume concentration, and PPFD, which are applicable to the model [18].

Let the samples in the data set be $(x_i, y_i) \in R^n \times R$, where x_i is an input independent variable and y_i is the output dependent variable. A total of 528 items (ambient temperature, CO₂ concentration, photosynthetically active radiation, and photosynthetic rate under the corresponding conditions) were included in the data set. We took the ambient temperature, CO₂ concentration, and photosynthetically active radiation as inputs and the apparent photosynthetic rate under the corresponding conditions as the output, thus transforming the model prediction into a nonlinear regression problem ($i = 1, 2, 3; n = 528$). The goal of SVR modelling is to find the relationship between the independent and dependent variables and to provide a function that represents the relationship as close as possible. When a new set of x_i is input, the method can provide the corresponding predicted values. It is expressed by:

$$f(x) = \sum_{i=1}^n \omega * \varphi(x) + b \quad (7)$$

where $\varphi(x)$ is a mapping function, ω represents the weight, and b represents the deviation, and ω and b are the targets to be studied, as they determine the linear hyperplane suitable for the data set being used for training. To map x and y , which do not have a linear relationship per se, to a higher dimensional space, $\varphi(x)$ and y in the higher dimensional space have a linear relationship, and $\varphi(x)$ is a nonlinear mapping about x . The fitting accuracy must also comply with Eq. (8):

$$|f(x_i) - y_i| \leq \varepsilon \quad (8)$$

where ε is an insensitive loss function, which is introduced to ignore the error within a certain range between the obtained function value and the true value. Then, a penalty parameter $C (C > 0)$ and nonnegative relaxation factors ξ^i and ξ_i^* are introduced, and this nonlinear regression problem can be transformed into a convex quadratic optimization problem that can be used to represent the deviation of the value of the training set fitting function $f(x)$ from the actual value. The SVR minimizes the regression risk by considering Eq. (9).

$$\begin{aligned} \text{Min: } & \frac{1}{2} \omega^2 + C \sum_{i=1}^l (\xi + \xi^*) \\ \text{s. t. } & \begin{cases} f(x_i) - y_i \leq \varepsilon + \xi_i^*, i = 1, \dots, l \\ y_i - f(x_i) \leq \varepsilon + \xi_i, i = 1, \dots, l \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, l \end{cases} \end{aligned} \quad (9)$$

To reduce the difficulty of the calculation, the above equation can be transformed into its dual

form by introducing the Lagrange multiplier a_i^*, a_i and kernel function $K(x_i, x_j) = [\varphi(x_i), \varphi(x_j)]$, as in Eq. (10):

$$f(x) = \sum_{i=1}^n (a_i^* - a_i) K(x_i, x_j) + b$$

$$s. t. 0 \leq a_i^* \leq C, \quad 0 \leq a_i \leq C \quad (10)$$

The kernel function can greatly reduce the computational complexity of the SVR for nonlinear problems. Generally speaking, functions satisfying the Mercer theorem can be used as kernel functions. Nowadays, the commonly used kernel functions are linear kernel function, RBF kernel function, polynomial kernel function, etc. In order to avoid the complexity changing with parameters during the computation, we choose a Gaussian kernel function, which is also called radial basis kernel function (RBF, Radial basis function kernel).

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{2g^2}\right) \quad (11)$$

3.1.2. Particle swarm algorithm

After determining the use of the RBF, the SVR model must obtain the optimal values of the penalty coefficient C , the parameter g of the RBF, and the accuracy ε . The PSO algorithm is a commonly used population intelligence theory algorithm with high confidence and generalization ability because of the advantages of fast convergence, fewer adjustment parameters, and less likelihood of falling into local optimality, and it has shown superior performance in solving complex nonlinear optimization problems [28].

The solution principle is as follows. By tracking the individual and global extremes of a particle, the velocity and position of the particle are calculated and updated.

Specifically, suppose there exists a D -dimensional solution space and a population $X = (x_1, x_2, \dots, x_n)$ consisting of m particles; then, the position of the i -th particle in the space is expressed as $X_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T$, and the velocity is expressed as $V_i = [v_{i1}, v_{i2}, \dots, v_{in}]^T$, which is updated using Eqs. (12) and (13). Setting a reasonable particle swarm size can ensure a uniform initial population distribution of the particle swarm algorithm.

$$v_{id} = wv_{id}(k) + c_1r_1(k)(pbest_{id}(k) - x_{id}(k)) + c_2r_2(k)(gbest_{id}(k) - x_{id}(k)) \quad (12)$$

$$x_{id}(k+1) = x_{id}(k) + v_{id}(k+1) \quad (13)$$

Where c_1 and c_2 represent learning factors, w represents inertia weights, r_1 and r_2 are pseudorandom numbers, which are independent of each other, and $r_1, r_2 \sim U(0,1)$ and d represent particle dimensions. $pbest_{id}(k)$ represents the individual extreme value of the particle in the d -th dimension, while $gbest_{id}(k)$ represents the global extreme value in the k -th iteration.

3.1.3. Particle swarm algorithm based SVR parameter tuning process

The specific algorithm process is as follows.

Step 1: To prevent large data overwhelming small data and facilitate the calculation and prediction of data with differences in units, forms and attributes, we perform a standardization operation on the data input as independent variables. The standardized formula is shown in Eq. (14).

$$x_i(j) = \frac{x_i(j) - x_i}{z_i} \quad (14)$$

where $x_i(j)$ represents the j -th sample value of the i -th input variable, $i = 1, 2, \dots, m$ represents the input variable, and $j = 1, 2, \dots, n$ represents the sample number.

$$z_i = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_i(j) - x_i)^2} \quad (15)$$

$$x_i = \frac{1}{n} \sum_{j=1}^n x_i(j) \quad (16)$$

Since the output of the model is only the photosynthetic rate, there is no need to normalize the output and no need to reverse-normalize the predicted results, saving computational costs.

Step 2: Import the pre-processed data and divide it randomly into training and test sets.

Step 3: Take the penalty parameter C of the SVR, the loss function ε and parameter g of the RBF as the optimization objects and set the initialization information of the particle swarm. Set the initial parameters of the PSO algorithm to optimize the SVR model: the particle population size is set to 10, the maximum number of iterations is 10, w is set to 0.4, and the learning factors C_1 and C_2 are set to 1.5 and 1.6, respectively, with restrictions on the variation range of the velocity and position.

Step 4: Initialize the particle swarm, substitute the divided training set and test set data, substitute the initialized parameters from step 3 into the SVR model for training, and use the output model prediction accuracy as the fitness function of the PSO algorithm.

Step 5: Compare the fitness value of each particle with the saved optimal position, determine the optimal position of the particle, continuously update it, and calculate and collect the fitness value of the particle in each new round.

Step 6: When the set maximum number of iterations is completed or the particle fitness value no longer changes significantly with a new iteration, we obtain the penalty coefficient C , parameter g of the RBF and the loss function ε of the current SVR model.

Step 7: The values of each parameter obtained in step 6 are substituted into the SVR model for prediction.

The training process of the PSO-SVR model in this study is shown in Figure 3.

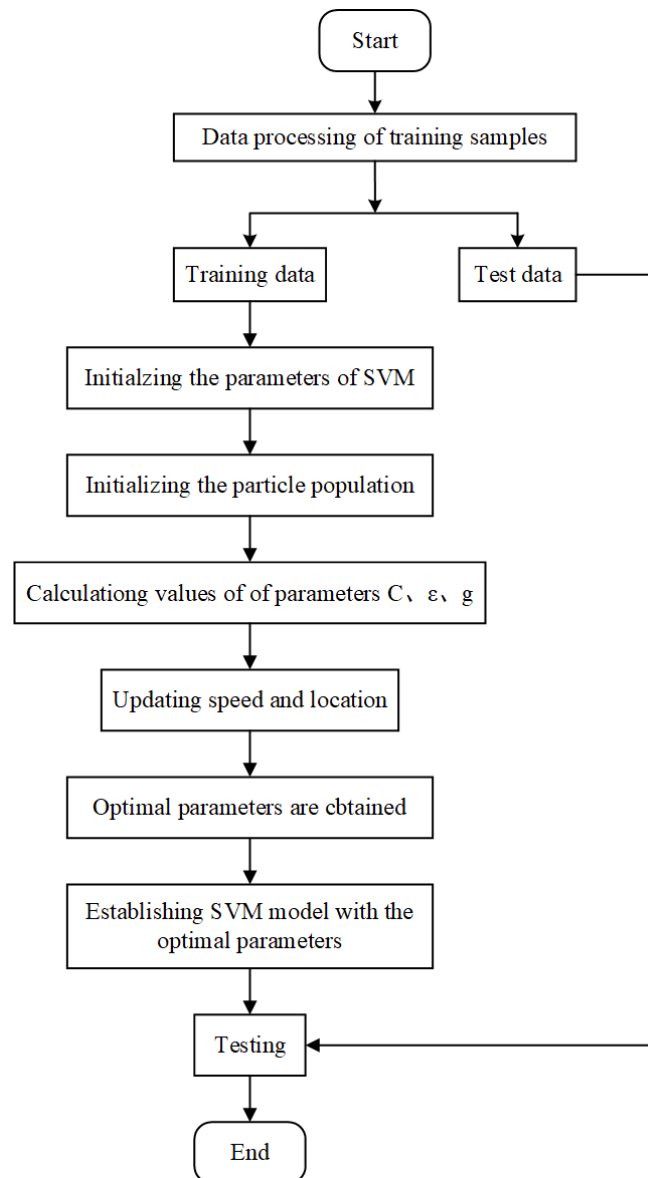


Figure 3. Flow chart of PSO-SVR model training.

3.2. Control methods

3.2.1. Multiple linear regression (MLR)

Multiple Linear Regression (MLR) is a common and effective statistical tool for describing more complex input-output relationships. The regression line in MLR can be expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_t x_t + \alpha \quad (17)$$

where y is the dependent variable, x_i is the i -th independent variable, β_i is the polynomial coefficients of x_i , t is the number of independent variables, and α is the possible variation form.

Then, the obtained MLR model can be adopted to predict the possible dependent variable related with the newly input vector.

3.2.2. Back-propagation neural network (BPNN)

Back Propagation Neural Network (BPNN), a supervised algorithm based on gradient descent, is composed of two processes, namely, the positive propagation of information and the reverse propagation of error [18].

BPNN consists of four steps: 1) Initialize the weights and thresholds in the network. 2) Perform forward propagation. 3) Calculate the error and perform backpropagation. 4) Adjust the weights and thresholds in the network.

3.2.3. Random Forest (RF)

Random Forest (RF) is a kind of ensemble algorithm (Ensemble Learning) that was developed as an extension of classification and regression trees (CART) to improve the estimation accuracy [29].

The training stage is to train multiple binary decision trees. Every classification regression tree can grow completely, and then get a low offset tree. Simultaneously, the method of random attribute selection and bagging makes lower correlation of each individual in random forest. The RF model obtains a better prediction result by modeling multiple trees instead of just one.

3.3. Evaluation Indicators

To verify the predictive performance of the PSO-SVR model for the photosynthetic rate, the determination coefficient (R^2), Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Error (MAE) were selected as model evaluation indexes. The calculation equations of each evaluation index are as follows.

$$R^2 = 1 - \frac{\sum_i^n (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2} \quad (18)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (19)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (20)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (21)$$

Where y_i is the true value to be predicted, $\hat{y}_i$ is the predicted value, and N is the total number of predictions. After we fit the data, the actual values were compared with the predicted values in the validation set.

4. Simulation results and analysis

4.1. Data set devision

We divided the data set in two ways and performed validation tests on their corresponding test sets using different validation criteria, such as MSE, MAE, RMSE, and R^2 . As shown in Table 1, we use D1 to represent one-tenth of the data for testing. The test set contains 53 sets of data and the remaining 475 sets of data are involved in training; D2 to represent the division into training and test sets by 7:3, resulting in 370 sets of data in the training set and 158 sets of data in the test set.

Table 1. Two ways of dividing the data set.

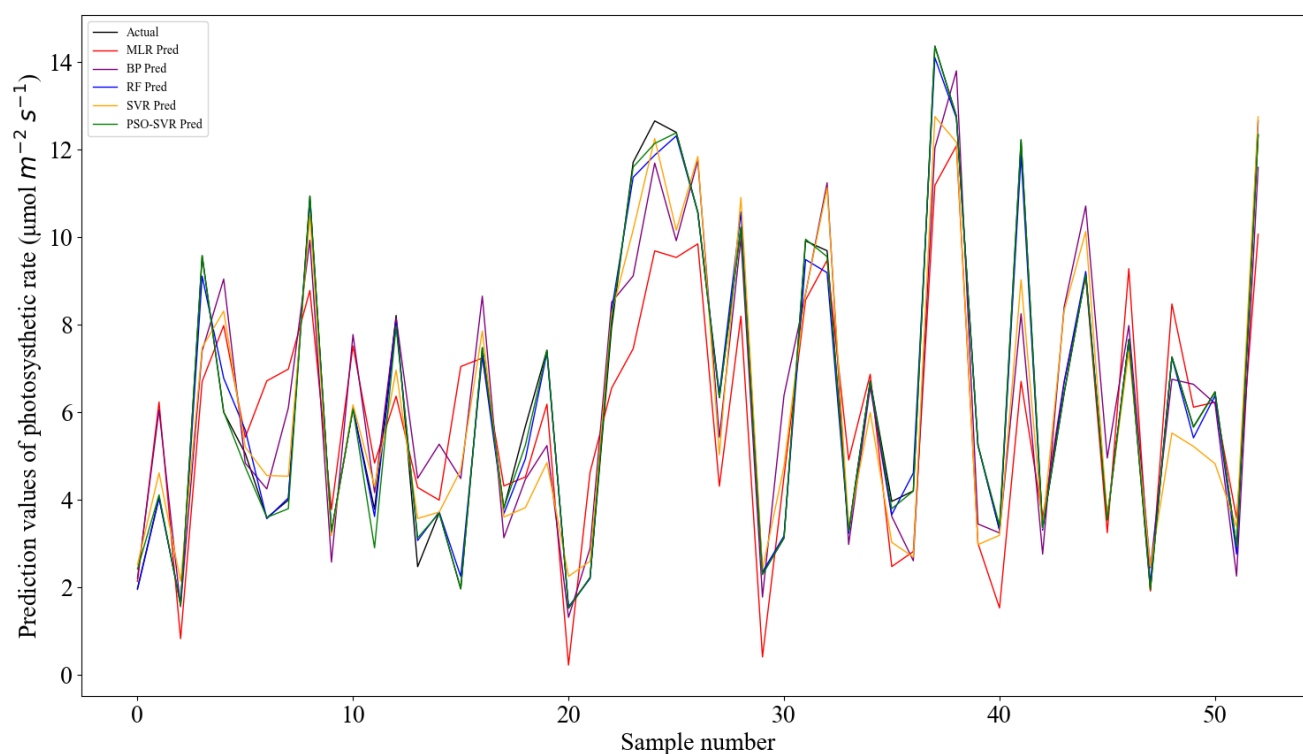
Data set	Training set	Validation set	Test set	Total
D1	422	53	53	528
D2	370	0	158	528

4.2. Parameter optimization results

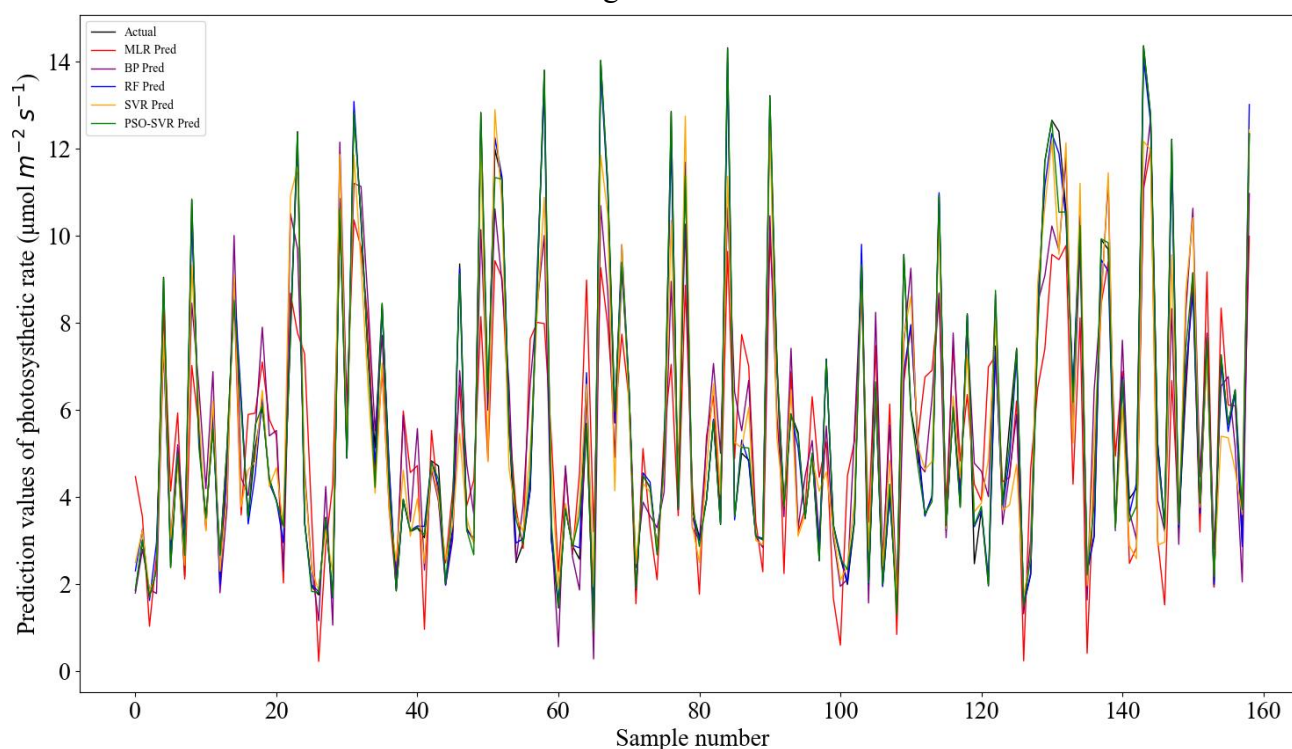
Different hyperparameters have different effects on the generalization ability of the SVR model. The best hyperparameters should be selected as much as possible to maximize the accuracy of the model. In this paper, the particle swarm (PSO) algorithm is chosen to optimize the parameters of the SVR model. The particle population size is set to 30, the maximum number of iterations is 30, the inertia weight is set to 0.4, and the learning factors c_1 and c_2 are set to 1.5 and 1.6, respectively, while the range of variation of velocity and position is limited. After parameter tuning, the optimal parameters for the PSO-SVR photosynthetic rate prediction model were determined to be $C=10$, $\varepsilon=0.01$, and $g=4.3282$ in the division case of D1, and $C=10$, $\varepsilon=0.02$, and $g=2.6653$ in the division case of D2.

4.3. Photosynthetic rate prediction analysis and error statistics

For a comprehensive assessment of the predictive performance of the photosynthetic rate prediction model based on the PSO-SVR algorithm, LR, RF, SVR, and BPNN were also used to model on the same data set. Figure 4 a, b depicts the upper limits of the tested effects of the five algorithms for both D1 and D2 division cases, respectively.



a. Prediction curves of the five algorithms on the test set in D1 condition.

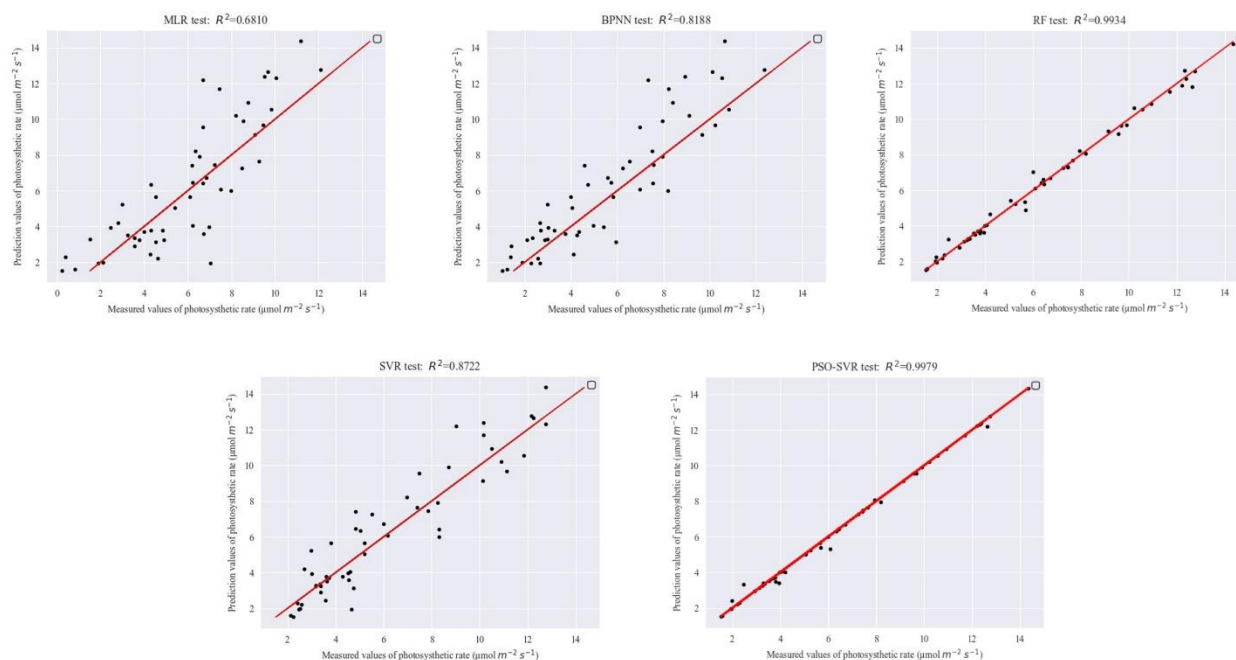


b. Prediction curves of the five algorithms on the test set in D2 condition.

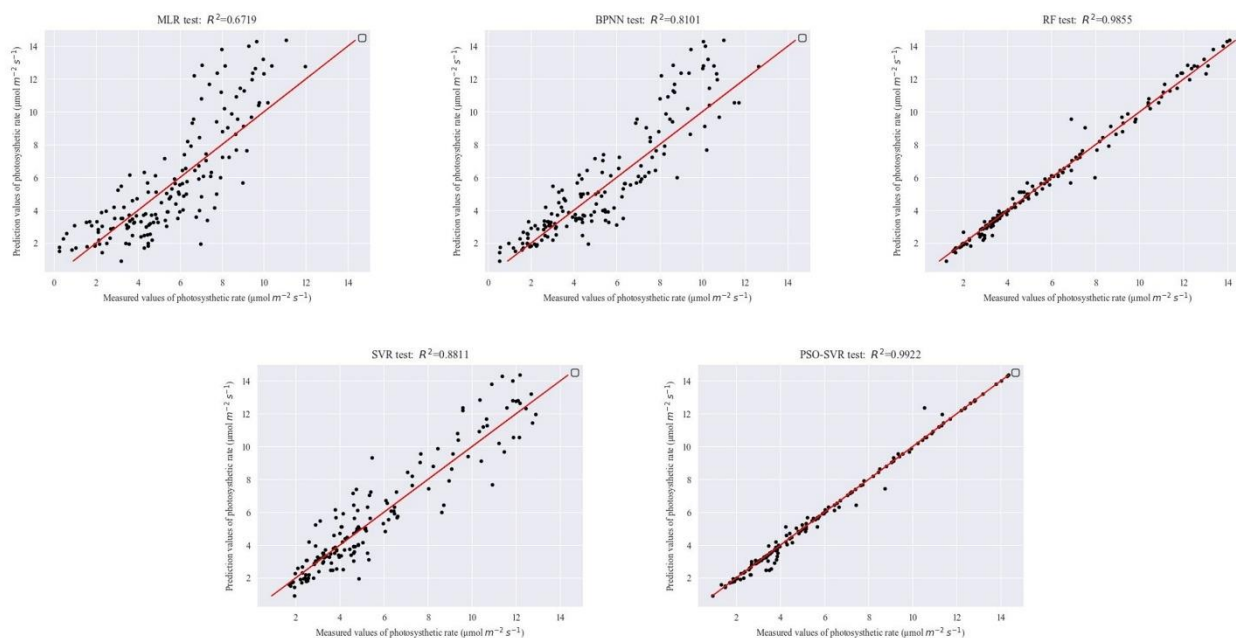
Figure 4. Prediction curves of the five models.

From Figure 5 a, b, we can analyze more specifically the correlation between the measured and predicted values of the photosynthetic rate models optimized in different ways on different test sets. It

is clear that almost all models with good performance in D1 condition show a decrease in performance when the data used for training are reduced. However, in both D1 and D2 conditions, the regression coefficient of PSO-SVR is closer to 1 compared to the rest of the models, which means that the PSO-SVR model has better performance on the photosynthetic rate data set.



a. Correlation between predicted and measured photosynthetic rate values under D1 conditions.



b. Correlation between predicted and measured photosynthetic rate values under D2 conditions.

Figure 5. Predictive correlation demonstration for five models.

To further analyze the prediction accuracy and generalization ability of the five models, we plotted Table 2. Table 2 shows the results of the error comparison of R^2 , MSE, MAE, and RMSE on different

test sets. Among the five methods, the R^2 of BPNN, RF, and PSO-SVR are all above 0.9000, indicating that these models have good fitting effects and the independent variables explain the dependent variable to a considerable extent.

It is generally accepted that random forest has slightly lower prediction accuracy than SVR on small samples [29]. In this paper, in the division case of D1, the R^2 of both RF and PSO-SVR are above 0.9900. the MSE of RF is 0.0843, the MAE is 0.1843, and the RMSE is 0.2903. the MSE of PSO-SVR is 0.0258, the MAE is 0.0739, and the RMSE is 0.1608. both models have small errors. However, in the case of D2 division, in the face of reduced training data, although the R^2 of both RF and PSO-SVR are above 0.9800, the prediction accuracy of RF decreases more than that of PSO-SVR. the MSE of RF is 0.1643, MAE is 0.2318, and RMSE is 0.4053. the MSE of PSO-SVR is 0.0926, MAE is 0.1461, and RMSE is 0.3043. the prediction error increases for both models, but the PSO-SVR model has a smaller variation in error. This still indicates that the PSO-SVR model established in this paper has high prediction accuracy and generalization ability on the photosynthetic rate data set. In addition, the running time of random forest is smaller than that of PSO-SVR, but there is a non-negligible difference between the prediction accuracy of the two algorithms on the small sample data set, and the PSO-SVR with better generalization ability should not be abandoned because of the slightly faster running speed of RF. under comprehensive consideration, the PSO-SVR model is the best model on the small sample data.

Table 2. Comparison of the experimental accuracy and error of the different prediction methods.

Method		R^2	MSE	MAE	RMSE
MLR	D1	0.6810	3.9071	1.5458	1.9766
	D2	0.6719	3.9203	1.5312	1.9799
BP	D1	0.8188	2.2559	1.2139	1.5020
	D2	0.8101	2.2698	1.1510	1.5066
RF	D1	0.9934	0.0843	0.1843	0.2903
	D2	0.9855	0.1643	0.2318	0.4053
SVR	D1	0.8722	1.5906	0.9758	1.2611
	D2	0.8811	1.4206	0.38840	1.1919
PSO-SVR	D1	0.9979	0.0258	0.0739	0.1608
	D2	0.9922	0.0926	0.1461	0.3043

5. Conclusion

To meet the need for distributed regulation of the greenhouse light environment, this paper attempts to find an optimal modeling approach for environmental factors associated with photosynthetic rate and compares the performance of several models for predicting photosynthetic rate in strawberry leaves. In the strawberry photosynthetic data set based on correlations verified by gray correlation analysis, temperature, CO_2 concentration, and PPFD measurements near the leaves were used as input variables to the model, and photosynthetic rate was used as an output variable. In addition,

the difference in prediction accuracy between different methods of dividing the data set was also tested. We observe that the method of optimizing the parameters in the support vector regression model using the particle swarm algorithm has the best generalization ability, which exhibits an R^2 greater than 0.9900 on the test sets obtained with different ways of dividing the data set. The predictive model will enable the greenhouse light environment distributed control solution to dynamically measure the current greenhouse ambient temperature, CO₂ volume concentration and light flux density and accurately predict the light saturation point based on the distributed monitoring terminal deployed in the greenhouse, and then calculate the light replenishment strategy to achieve fine control of light in the local area. The model can be used as a soft sensor. In addition, it has some ability to be extended to other plants in the same family as strawberry. In the next step, we will add humidity and other parameters to the model and carry out the deployment of distributed regulation of the greenhouse light environment.

Acknowledgments

This project is supported by Key Research and Development Project of Anhui Province (202104a06020012 and 202204c06020022).

Appendix A. Supplementary material

For completeness, the real experimental data used in this paper can be found on this page at <http://github.com/AHAU662/Photosynthesis-experiment>.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Q. Li, R. H. Zhang, Y. Wang, Interannual variation of the wintertime fog-haze days across central and eastern China and its relation with East Asian winter monsoon, *Int. J. Climatol.*, **36** (2016), 346–354. <https://doi.org/10.1002/joc.4350>
2. X. Li, W. Lu, G. Y. Hu, X. C. Wang, Y. Zhang, Effects of light-emitting diode supplementary lighting on the winter growth of greenhouse plants in the Yangtze River Delta of China, *Bot. Stud.*, **57** (2016). <https://doi.org/10.1186/s40529-015-0117-3>
3. Y. Kong, D. Llewellyn, Y. B. Zheng, Response of growth, yield, and quality of pea shoots to supplemental light-emitting diode lighting during winter greenhouse production, *Can. J. Agric. Sci.*, **98** (2018), 732–740. <https://doi.org/10.1139/cjps-2017-0276>
4. Y. C. Xu, Y. X. Chang, G. Y. Chen, H. Y. Lin, The research on LED supplementary lighting system for plants, *Optik*, **127** (2016), 7193–7201. <https://doi.org/10.1016/j.ijleo.2016.05.056>
5. T. Kmet, M. Kmetova, Adaptive critic design and Hopfield neural network based simulation of time delayed photosynthetic production and prey–predator model, *Inf. Sci.*, **294** (2015), 586–599. <https://doi.org/10.1016/j.ins.2014.08.020>
6. Y. J. Zheng, Y. T. Zhang, H. C. Liu, Y. M. Li, Y. L. Liu, Supplemental blue light increases growth

- and quality of greenhouse pak choi depending on cultivar and supplemental light intensity, *J. Integr. Agric.*, **17** (2018), 2245–2256. [https://doi.org/10.1016/S2095-3119\(18\)62064-7](https://doi.org/10.1016/S2095-3119(18)62064-7)
7. L. Xu, R. Wei, L. Xu, Optimal greenhouse lighting scheduling using canopy light distribution model: A simulation study on tomatoes, *Lighting Res. Technol.*, **52** (2020), 233–246. <https://doi.org/10.1177/1477153519825995>
 8. O. D. Palmitessa, M. A. Pantaleo, P. Santamaria, Applications and development of LEDs as supplementary lighting for tomato at different latitudes, *Agronomy*, **11** (2021), 835. <https://doi.org/10.3390/agronomy11050835>
 9. F. F. He, L. H. Zeng, D. M. Li, Z. H. Ren, Study of LED array fill light based on parallel particle swarm optimization in greenhouse planting, *Inf. Process. Agric.*, **6** (2019), 73–80. <https://doi.org/10.1016/j.inpa.2018.08.006>
 10. C. Y. Li, H. Y. Shan, C. H. Zhang, H. Q. Liu, X. C. Guo, M. Z. Xu, Optimal regulation model of Greenhouse light under limited light resources, *IOP Conf. Ser. Earth Environ. Sci.*, **792** (2021), 12025. <https://doi.org/10.1088/1755-1315/792/1/012025>
 11. M. Carlini, S. Castellucci, A. Mennuni, S. Morelli, Numerical modeling and simulation of pitched and curved-roof solar greenhouses provided with internal heating systems for different ambient conditions, *Energy Rep.*, **6** (2020), 146–154. <https://doi.org/10.1016/j.egyr.2019.10.033>
 12. N. Choab, A. Allouhi, A. E. Maakoul, T. Kousksou, S. Saadeddine, A. Jamil, Review on greenhouse microclimate and application: Design parameters, thermal modeling and simulation, climate controlling technologies, *Solar Energy*, **191** (2019), 109–137. <https://doi.org/10.1016/j.solener.2019.08.042>
 13. J. T. Chen, Y. W. Ma, Z. Z. Pang, A mathematical model of global solar radiation to select the optimal shape and orientation of the greenhouses in southern China, *Solar Energy*, **205** (2020), 380–389. <https://doi.org/10.1016/j.solener.2020.05.055>
 14. R. Liu, M. Li, J. L. Guzmán, F. Rodríguez, A fast and practical one-dimensional transient model for greenhouse temperature and humidity, *Comput. Electron. Agric.*, **186** (2021), 106186. <https://doi.org/10.1016/j.compag.2021.106186>
 15. T. Liu, Q. Y. Yuan, Y. G. Wang, Hierarchical optimization control based on crop growth model for greenhouse light environment, *Comput. Electron. Agric.*, **180** (2021), 105854. <https://doi.org/10.1016/j.compag.2020.105854>
 16. P. P. Xin, B. Li, H. H. Zhang, J. Hu, Optimization and control of the light environment for greenhouse crop production, *Sci. Rep.*, **9** (2019), 1–13. <https://doi.org/10.1038/s41598-019-44980-z>
 17. P. P. Xin, H. H. Zhang, J. Hu, Z. Y. Wang, Z. Zhang, An improved photosynthesis prediction model based on artificial neural networks intended for cucumber growth control, *Appl. Eng. Agric.*, **34** (2018), 769–787. <https://doi.org/10.13031/aea.12634>
 18. J. Hu, P. P. Xin, S. W. Zhang, H. H. Zhang, D. J. He, Model for tomato photosynthetic rate based on neural network with genetic algorithm, *Int. J. Agric. Biol. Eng.*, **12** (2019), 179–185. <https://doi.org/10.25165/j.ijabe.20191201.3127>
 19. J. T. Ding, H. Y. Tu, Z. L. Zang, M. Huang, S. J. Zhou, Precise control and prediction of the greenhouse growth environment of *Dendrobium candidum*, *Comput. Electron. Agric.*, **151** (2018), 453–459. <https://doi.org/10.1016/j.compag.2018.06.037>
 20. Y. W. Liu, D. J. Li, S. H. Wan, F. Wang, W. C. Dou, X. L. Xu, et al., A long short-term memory-based model for greenhouse climate prediction, *Int. J. Intell.*, **37** (2022), 135–151.

<https://doi.org/10.1002/int.22620>

21. X. Y. Zhang, Z. Y. Huang, X. H. Su, A. Siu, Y. P. Song, D. Q. Zhang, et al., Machine learning models for net photosynthetic rate prediction using poplar leaf phenotype data, *PLoS One*, **15** (2020), e228645. <https://doi.org/10.1371/journal.pone.0228645>
22. K. P. Zheng, Y. Bo, Y. D. Bao, X. L. Zhu, J. Wang, Y. Wang, A machine learning model for photorespiration response to multi-factors, *Horticulturae*, **7** (2021), 207. <https://doi.org/10.3390/horticulturae7080207>
23. D. H. Jung, H. S. Kim, C. Jhin, H. J. Kim, S. H. Park, Time-serial analysis of deep neural network models for prediction of climatic conditions inside a greenhouse, *Comput. Electron. Agric.*, **173** (2020), 105402. <https://doi.org/10.1016/j.compag.2020.105402>
24. H. G. Choi, B. Y. Moon, N. J. Kang, Effects of LED light on the production of strawberry during cultivation in a plastic greenhouse and in a growth chamber, *Sci. Hortic*, **189** (2015), 22–31. <https://doi.org/10.1016/j.scienta.2015.03.022>
25. S. V. Archontoulis, F. E. Miguez, Nonlinear regression models and applications in agricultural research, *Agron. J.*, **107** (2015), 786–798. <https://doi.org/10.2134/agronj2012.0506>
26. M. Riches, D. Lee, D. K. Farmer, Simultaneous leaf-level measurement of trace gas emissions and photosynthesis with a portable photosynthesis system, *Atmos. Meas. Tech.*, **13** (2020), 4123–4139, <https://doi.org/10.5194/amt-13-4123-2020>, 2020
27. X. X. Huang, M. Chen, W. J. Wang, Y. Ge, J. Xie, Shelf-life Prediction of chilled *Penaeus vannamei* using grey relational analysis and support vector regression, *J. Aquat. Food Prod. Technol.*, **29** (2020), 507–519. <https://doi.org/10.1080/10498850.2020.1766616>
28. Y. Zhang, H. Sun, Y. Guo, Wind power prediction based on PSO-SVR and grey combination model, *IEEE Access*, **7** (2019), 136254–136267. <https://doi.org/10.1109/ACCESS.2019.2942012>
29. J. H. Li, D. S. Zhu, C. X. Li, Comparative analysis of BPNN, SVR, LSTM, Random Forest, and LSTM-SVR for conditional simulation of non-Gaussian measured fluctuating wind pressures, *Mech. Syst. Signal. Process*, **178** (2022), 109285. <https://doi.org/10.1016/j.ymssp.2022.109285>



AIMS Press

©2022 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)