Prediction of Soil Moisture in Grassland Based on Information Gain and Model Modification

Lizhu Yue, Kaidi Sun*

School of Business Administration, Liaoning Technical University, Huludao 125105, Liaoning, China

**Corresponding Author.*

Abstract:

Accurate prediction of grassland soil moisture is the key to understanding and responding to the destruction of grassland ecosystems. In order to improve the prediction effect of grassland soil moisture, this study proposed a grassland soil moisture prediction model based on information gain and model modification based on multi-dimensional and small sample size data from a testing station in Xilingol Grassland. First, we conduct correlation analysis on the 22-dimensional feature data of Xilin Gol grassland based on the Pearson correlation coefficient, and screen out the 15-dimensional feature data that can represent the whole; then calculate the information gain of the features on soil moisture, and screen out the 6-dimensional high information gain features; Secondly, by introducing the evolutionary boundary constraint processing scheme, Levy flight strategy and group fitness variance strategy to jointly improve the pathfinder algorithm (CLPPFA), it improves its global optimization capability, thereby optimizing extreme learning machine (ELM) related parameters and constructing a grassland soil moisture prediction model, and use 6-dimensional high information gain feature data to predict the preliminary results of grassland soil moisture; Finally, by establishing an ARIMA error correction model, the error prediction value and the preliminary prediction value are superimposed to obtain the final prediction result. The results show that the improved extreme learning machine's fitting degree R2 for grassland soil moisture prediction is 0.937, which is better than the PFA_ELM model and the ELM model; the ARIMA model is introduced to analyze the error sequence of the preliminary prediction results of CLPPFA_ELM, and ARIMA (2,0,2) The model is error corrected. The prediction fit R2 of the CLPPFA_ELM-ARIMA model is 0.988, which is significantly improved compared to the prediction effects of the CLPPFA_ELM, SVR, RF, BP and ridge regression models. In summary, it is shown that the model has good fitting effect and generalization ability in grassland soil moisture prediction. This model provides model reference and technical support for formulating effective grassland management and protection strategies.

Keywords: soil moisture, prediction model, information gain, pathfinder algorithm, extreme learning machine, error correction model

INTRODUCTION

Grassland ecosystems, as an important ecological component of the earth, cover about a quarter of the global land area, and their health and stability are directly related to many aspects such as biodiversity, soil and water conservation, and climate regulation [1]. Soil moisture, as a key environmental factor in grassland ecosystems, directly affects the carbon cycle, plant growth and biodiversity of grassland ecosystems, and plays a crucial role in maintaining the ecological balance of grasslands [2]. However, with global climate change and intensified human activities, grassland ecosystems are facing many challenges [3], such as drought, land degradation and biodiversity loss. Therefore, accurate prediction of grassland soil moisture is important for understanding and addressing these challenges and for developing effective grassland management and conservation strategies.

Traditional prediction methods often rely on complex physical models, which are computationally expensive and difficult to cope with complex environmental variables. Machine learning has obvious advantages in dealing with large-scale data and complex non-linear problems, and has become a key technology for solving scientific problems and industrial application problems. Research on the prediction of soil moisture using data-driven models based on past time series data has been widely studied and has been shown to be effective in various situations [4, 5], among them support vector machine [6, 7] (support vector machine, SVM), random forest [8] (random forest, RF), BP neural network [9], Extreme learning machine [10] (extreme learning machine, ELM), LSTM [11, 12] and other methods are widely used for soil moisture prediction.

In order to compensate for the shortcomings of a single model, Bates and Granger [13] proposed the method of combining prediction models, which aims to give full play to the advantages of a single prediction model by combining two or more different prediction methods for modelling, and at the same time reduce the prediction errors caused by parameter errors or model errors [14]. The application of this method to soil prediction research has received attention from many scholars. Yu et al. [15] designed a model for soil moisture at different depths in agricultural fields by integrating the spatio-temporal feature extraction advantages of ResNet and BiLSTM models; ElSaadani et al. [16] designed a deep learning algorithm that combines CNN and LSTM, and the experimental results showed that the algorithm can better predict the soil moisture changes in the study area after one day; Sun et al. [17] proposed to use genetic algorithm to improve the support vector machine as a way to make inverse prediction of soil moisture; Li et al. [18] used particle swarm optimisation algorithm to obtain the best parameters of the limit learning machine, and used GNSS-IR technology to obtain soil moisture data to verify the fitting ability of the model, and the test proved the reliability and superiority of the model.

In summary, the prediction of grassland soil moisture can be effectively improved by relying on the machine learning combination model. Combined with the high latitude and small sample of 10cm soil moisture data of Xilingol grassland, this paper proposes a grassland soil moisture prediction model of CLPPFA_ELM-ARIMA considering the information gain. The Pearson correlation coefficient is used to screen the features with high relevance in the Xilingol grassland; the information gain of the features on soil moisture is considered, and the features with high information gain are used as the input variables of the CLPPFA_ELM model; the evolutionary boundary constraint processing mechanism is introduced to solve the individual boundary-crossing problem of the Pathfinder Algorithm (PFA), and the The Levy flight and group fitness variance strategy are used to improve the global optimisation ability of the algorithm, which is used to construct the CLPPFA_ELM grassland humidity prediction model to obtain the preliminary prediction results; the error sequence characteristics are analysed, and an ARIMA error correction model is established to superimpose the error prediction value with the preliminary prediction value to obtain the final prediction results.

OVERVIEW OF THE STUDY AREA AND DATA SOURCES

Overview of the Study Area

The Xilingol Grassland in Inner Mongolia is representative and typical of temperate grasslands. It is one of the four major grasslands in China and is located on the Inner Mongolia Plateau, with geographic coordinates ranging between 110°50′~119°58′E longitude and 41°30′~46°45′N latitude, and an average annual precipitation of 340 mm. It is not only a nationally important base for the production of animal husbandry, but also an important. It national livestock production base, green ecological barrier, the of sandstorms and severe weather, for response mechanisms to human disturbances and global climate change, an of the International Geosphere-Biosphere Programme (IGBP) Terrestrial Sample China Terrestrial Ecosystem Sample Belt (NECT). It also represents a typical domain for investigating the response mechanisms of ecosystems to human-induced disturbances and global climatic variations, constituting a crucial component of the Northeast China Terrestrial Ecosystem Sample Zone (NECT) within the International Geosphere-Biosphere Programme (IGBP).

Data Acquisition and Analysis

This experiment relies on the Inner Mongolia Xilingol Grassland Ecosystem Observatory (Site No. 54102099999, 115°22′30′′E, 44°7′30′′N, 1004 m above sea level), which monitors and calculates the resulting data factors, including meteorological factor data (temperature, barometric pressure, precipitation, visibility, wind speed, etc.), soil evapotranspiration, runoff volume, vegetation index, Leaf Area Index, and soil moisture, with a time span of January 2012 to March 2022 for monthly statistics. In order to test whether the influence indicators of the soil moisture prediction samples are independent of each other, Pearson correlation coefficient was used as the evaluation index for the correlation analysis of the samples, and the heat map of correlation between the features is depicted in Figure 1.

According to the range of values of Pearson correlation coefficient, it is evident from Figure 1 that there is a linear correlation between different characteristic indicators, i.e., there is an overlap of information between soil moisture characteristic indicators. This article selects factors with low correlation coefficients and factors that represent the overall monthly coefficients from indicators with high correlation coefficients as predictors, i.e., average air temperature, precipitation, maximum single-day precipitation, number of days of precipitation, average sea level barometric pressure, average station barometric pressure, average visibility, minimum visibility, maximum visibility, average wind speed, maximum single-day average wind speed, soil evapotranspiration, leaf area index LA low-level vegetation, runoff volume, and vegetation index. In order to enhance the decision-making process efficiency and increase the robustness of the model, the correlation between the data needs to be further explored and the data dimensionality needs to be reduced.

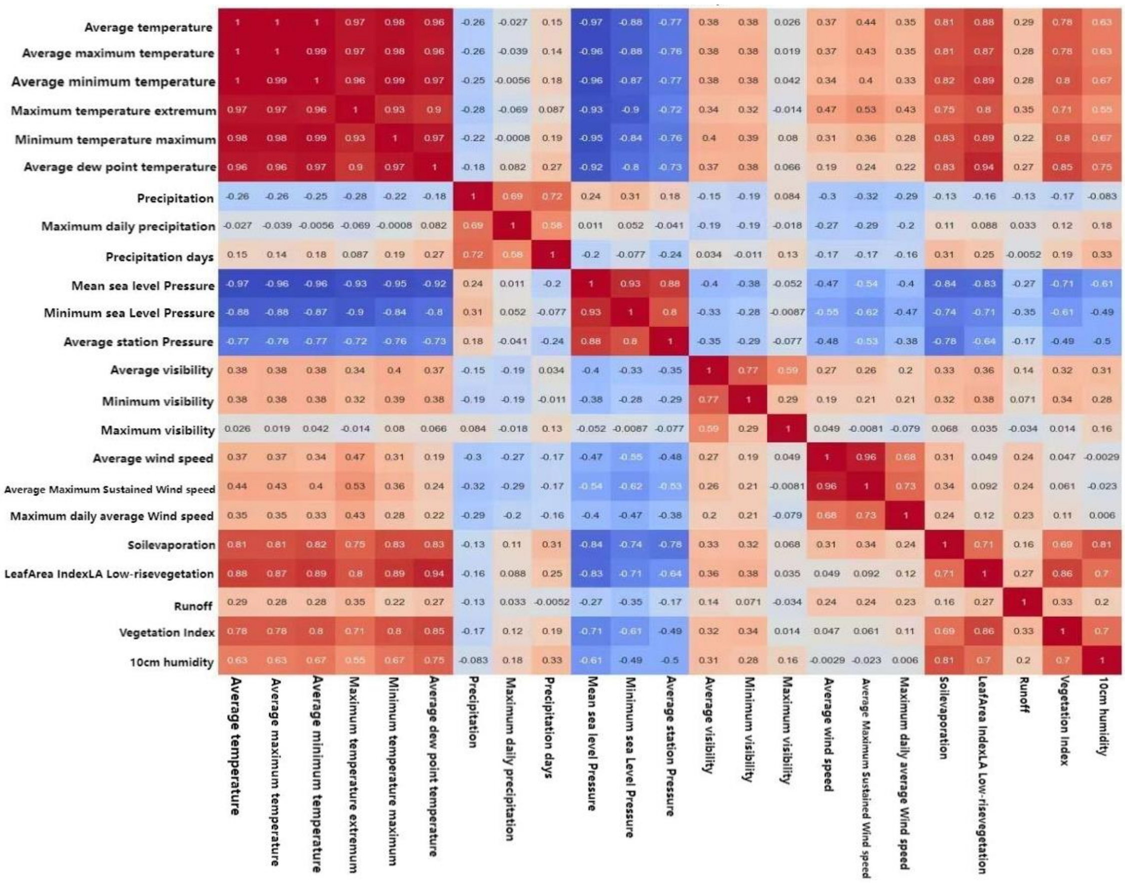


Figure 1. Feature indicator correlation heat map

Information Gain

Information gain is a measure of the predictive value of a feature for a target variable. In soil moisture prediction it indicates the degree of uncertainty reduction of information of a feature X in soil moisture$Y$, i.e., the disparity in entropy between of Y and the conditional information entropy of the attribute$X$, and its information gain is calculated by the following formula:

$$Gain(Y,X) = Ent(Y) - Ent(Y|X) \tag{1}$$

A high information gain signifies that the feature conveys substantial information regarding the target variable, which allows us to more accurately predict the value of the target variable. A low information gain, on the other hand, means that the feature does not help much in predicting the target variable, i.e. it hardly affects the uncertainty of the target variable. The information gain was calculated to obtain a graph of the percentage information gain for each type of indicator, as shown in Figure 2.

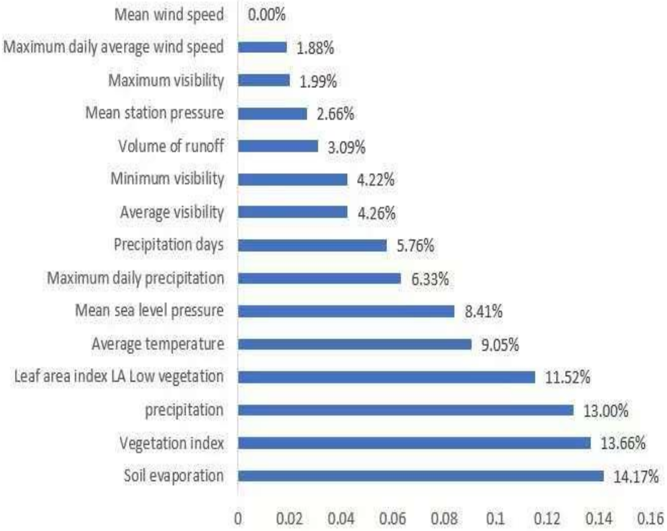


Figure 2. Information gain percentage chart of each factor indicator

As can be seen from Figure 2, soil evaporation, vegetation index, precipitation, leaf area index, mean air temperature and mean sea level pressure have the largest information gain, indicating that they are most closely related to soil moisture at 10 cm depth; the remaining factor indicators generally have small information gain, indicating that they are not closely related to soil moisture at 10 cm depth. Therefore, in this paper, the features with high information gain are used as input features of the prediction model to reduce data latitude and improve the model's prediction accuracy.

GRASSLAND SOIL MOISTURE PREDICTION MODEL CONSTRUCTION

Grassland soil moisture is subject to the joint action of multidimensional nonlinear influencing factors, and in the existing research on the problem of applying machine learning for moisture prediction, the learning algorithm of network structure has the best fitting effect on nonlinear data. ELM possesses the benefits of a straightforward structure and rapid learning speed. [19], and this paper proposes to use ELM model for grassland soil moisture prediction. By introducing the improved pathfinder algorithm to solve the problems of weak model generalisation ability and unsTable prediction effect caused by the random initialisation of $\omega$and b of ELM, and then improve its prediction accuracy in soil moisture. Analysing the characteristics of the error series, it is found that the change of the error series presents a certain pattern, and the ARIMA error correction model is established by treating the error series as a time series and superimposing the error prediction value with the preliminary prediction value to obtain the final prediction result.

CLPPFA Optimisation Algorithm Design

PFA

The PFA algorithm is a novel heuristic algorithm deduced from the community behaviour of group animals and their leadership system [20], which performs well in terms of global and local search capability by collaboratively searching for the global optimal solution through the communication between two roles, the pathfinder and the follower. The algorithm possesses the merit of ease of understanding, high performance and easy to operate and implement in parameter optimisation applications. However, PFA deals with the problem by pulling back the transgressing individuals, which makes the transgressing individuals gather at the boundary and affects the algorithm's population diversity and convergence speed; the follower in PFA tends to learn from the pathfinder, and is prone to fall into the problem of local optimum.

Introduction of an evolutionary boundary constraint treatment scheme

In order to solve the individual boundary crossing problem existing in PFA, this paper introduces the evolutionary boundary constraint processing mechanism proposed in the literature [21] to process the boundary crossing individuals to enhance the algorithm's performance. The processing method is as follows:

$$X_i^{`} = \begin{cases} a_1 \times lb_i + (1 - a_1) \times X_p, X_i < lb_i \\ a_2 \times ub_i + (1 - a_2) \times X_p, X_i > ub_i \end{cases} \tag{2}$$

Where$X_i^{`}$ is the position of the individual after the out-of-bounds treatment,X_i is the current position of the out-of-bounds individual,a_1 and a_2 are random numbers taken from the interval [0,1], ub and lb are the upper and lower limits of the population's individuals, respectively; andX_p is the present location of the pathfinder.

Introduction of the levy flight strategy

In the PFA algorithm, the followers will be misguided by the pathfinder and gradually gather around the pathfinder, falling into the local optimal solution. To mitigate the risk of the algorithm converging to a local optimal solution, this paper introduces the Levy flight strategy proposed in the literature [22], which conforms to the Levy distribution and randomly perturbs the position of the pathfinder. Its perturbation formula is as follows:

$$Levy(\beta) = 0.05 \times \frac{\mu}{|\nu|^{\frac{1}{\beta}}} \tag{3}$$

Where, $\beta = 1.5$, μ and ν obey the normal distribution as expressed in equation (4).

$$\mu \sim N(0, \sigma_x), \nu \sim N(0, \sigma_y)$$

$$\sigma_\mu = \left[\frac{\Gamma(1+\beta)\, sin\left(\frac{\pi\beta}{2}\right)}{\Gamma(\frac{1+\beta}{2}) \times \beta \times 2^{\frac{\beta-1}{2}}}\right]^{\frac{1}{\alpha}}, \sigma_\nu = 1 \tag{4}$$

Introducing a population fitness equation strategy

To address the issue that the follower in the PFA algorithm is prone to becoming trapped in the local optimum, this paper introduces the group fitness variance metric proposed in the literature [23] σ^2, as a kind of metric to measure the search state of the PFA, and σ^2 is calculated as follows:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} [f_i - f_\alpha]^2 \tag{5}$$

Where: N denotes the size of the Pathfinder group; f_i is the classification accuracy of individuals i, %; f_α is the average classification accuracy of the group, %.

σ^2 used to assess the degree of fluctuation of individual positions in the PFA algorithm, with larger fluctuations indicating global search and smaller degrees of fluctuation indicating local or global convergence. Comparing the indicator value σ^2 with the threshold value θ determines the search status of the PFA. If $\sigma^2 > \theta$, the global search is continued; if $\sigma^2 \leq \theta$, the two-point crossover operation of the genetic algorithm is used to locally update the position of each individual to avoid "premature" convergence.

CLPPFA_ELM Model Construction

Inputs: soil moisture dataset, maximum number of iterations T, population size N, population-adapted variance thresholdθ, number of hidden layer neurons$l$ and activation function$g(x)$.

Output: regression prediction model for example soil moisture data using CLPPFA optimized w and b training ELM.

(1) Randomly initialise the position of each individual according to the ELM optimisation object, the position is aK dimensional vector about$w$ and$b$ with values in the range [-1, 1].

$$K = s \times l + l \tag{6}$$

Where: s is the quantity of nodes present in the input layer of the ELM.

(2) The RMSE metric was chosen as the model fitness function as in Eq. (7), and the location of the individual possessing the minimal fitness value was recorded and set as the pathfinder.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \tag{7}$$

Where: y_i denotes the real value of soil moisture at , and $\hat{y}_i$ denotes the predicted value of soil moisture at i.

(3) Iterative updating of the pathfinder position, the pathfinder position is updated according to Eq. (8) and Eq. (9), and Eq. (2) is used to transgress the updated pathfinder position.

$$X_p^{t+1} = X_p^t + 2r_1(X_p^t - X_p^{t-1}) + A + Levy(\beta) \tag{8}$$

Where: t denotes the current iteration number;X_p^{t+1}, X_p^t,X_p^{t-1} denote the updated position of the pathfinder, the current position, and the position of the previous generation of the pathfinder, respectively; r_1 is a uniformly distributed random variable within the interval [0, 1], which denotes the step factor of the pathfinder;A is a set of perturbation vectors, which denotes the stochasticity of the pathfinder's updated position.

$$A = u_1 \cdot e^{\frac{-2t}{T}} \tag{9}$$

Where: T denotes the maximum number of iterations;u_1 is a set of random vectors in the range [-1, 1].

(4) Update the follower position according to Eq. (10) to Eq. (11), and use Eq. (2) to transgress the updated follower position.

$$X_i^{t+1} = X_i^t + \alpha \cdot r_2 \cdot (X_j^t - X_i^t) + \beta \cdot r_3 \cdot (X_p^t - X_i^t) + \varepsilon \tag{10}$$

Where:X_i^{t+1}, X_p^t, X_i^t, X_j^t denote the revised location of the follower.i, the current position of the pathfinder, the current position of the follower i , the current position of the follower j , respectively; r_2 and r_3 are uniformly generated random variables in the range of [0, 1], which are represented as the step factors of the movement with other followers and pathfinders; α, β denote the interaction coefficients of the followers, and the attraction coefficients of the pathfinders to the followers, which are both random numbers in the interval of [1, 2]. Both are random numbers in the interval [1, 2]; ε is the perturbation vector, which provides movement randomness for all the followers.

$$\varepsilon = (1 - \tfrac{t}{T}) \cdot u_2 \cdot D_{ij} \tag{11}$$

Where: u_2 is a set of random vectors in the interval [-1, 1];D_{ij} is the distance between follower i and follower j.

$$D_{ij} = \left\| X_i - X_j \right\| \tag{12}$$

Where:X_i denotes the location of the follower i;X_j denotes the location of the follower j.

(5) Calculate the population fitness variance σ^2, and if$\sigma^2 \leq \theta$, locally update the position of each individual using the two-point crossover operator of the genetic algorithm. Instead, continue the global search. Use the updated position of each individual as the initial position of each individual for the next iteration. Repeat steps 3)-5).

(6) Output the CLPPFA optimized w, b and train the ELM to fit the soil moisture data.

ARIMA Prediction Error Correction

Due to the limitations of the prediction model itself, the error sequence is constructed from the preliminary prediction results and the actual values, and the analysis reveals that its changes present a certain pattern. Considering the error series as a random time series, the error correction is carried out using the ARIMA model, the mathematical expression of which is shown in equation (13):

$$\left(1 - \sum_{i=1}^{p} \varphi_i \, L^i\right)(1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i \, L^i\right)\varepsilon_t \tag{13}$$

Where L is the lag operator, d is the difference order,φ_i is the adaptive coefficient,θ_i is the moving average coefficient,p is the autoregressive order,q is the moving average order, and ε_t is the residual series.

The steps to build the ARIMA error correction model are as follows:

(1) Identify the smoothness of the series. The use of ARIMA model needs to satisfy that the data series is a smooth series, so the initial prediction error value of the soil was tested for smoothness.

(2) Series smoothing. If the test yields a non-stationary series for the initial soil prediction error value, the error series is differenced using d times.

(3) Pattern recognition. Based on the ACF and PACF of the error sequence, the p, q values are determined.

(4) Model ordering. Determination of the parameters ,q often produces multiple combinations of eligible models, in order to select the most accurate prediction model, the AIC, BIC, HQ minimum criterion can be used for model comparison.

(5) AIRMA error correction model was developed and predicted.

Grassland Soil Moisture Prediction Process

The CLPPFA_ELM-ARIMA soil prediction method based on information gain is proposed in this paper to predict the 10cm thickness soil moisture in Xilingol grassland. The prediction process is shown in Figure 3.

The steps of the grassland soil moisture prediction model are as follows:

(1) Obtain the actual data of Xilinguole monitoring station, use Pearson correlation coefficient to eliminate the features with strong correlation, and select the features that can represent the whole.

(2) The screened features were individually calculated for their information gain on soil moisture, and the high information gain features were used as input values for model prediction.

(3) The selected features are dimensionless processed and divided into training and test sets to input into the CLPPFA-ELM model to obtain the predicted values $\hat{y}$.

(4) The error time seriesε is obtained from Eq. (14):

$$\varepsilon = y - \hat{y} \tag{14}$$

(5) Analyse the error characteristics and build an ARIMA error correction model, input the error sequence , train and predict, the error prediction result is ε'.

(6) Superimpose the preliminary and error prediction results to get the final soil moisture prediction .

$$Y = \hat{y} + \varepsilon' \tag{15}$$

(7) To assess the predictive performance of the model, Explained Variance Score (EVS), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R2 are chosen as evaluation metrics, and their calculation formulae are shown in Eq. (16)-(19).

$$EVS = 1 - \frac{Var\{y_i - \hat{y}_i\}}{Var\{y_i\}} \tag{16}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|(y_i - \hat{y}_i)| \tag{17}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{18}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{19}$$

Where, y_i denotes the real value of soil moisture at i, $\hat{y}_i$ denotes the predicted value of soil moisture at i, $\bar{y}$ denotes the mean value of soil moisture,m denotes the number of samples, and Var denotes the variance operation. EVS denotes the degree of explanation of the model on the fluctuation of the dataset, and the larger its value denotes the better the effect of the model; MAE and RMSE denote the error of the prediction results, and the smaller its value denotes the better prediction results; R2 denotes the degree of fit of the regression equation, and the larger its value denotes the better prediction performance. R2 indicates the goodness-of-fit of the regression equation, and the larger its value indicates the better prediction performance.

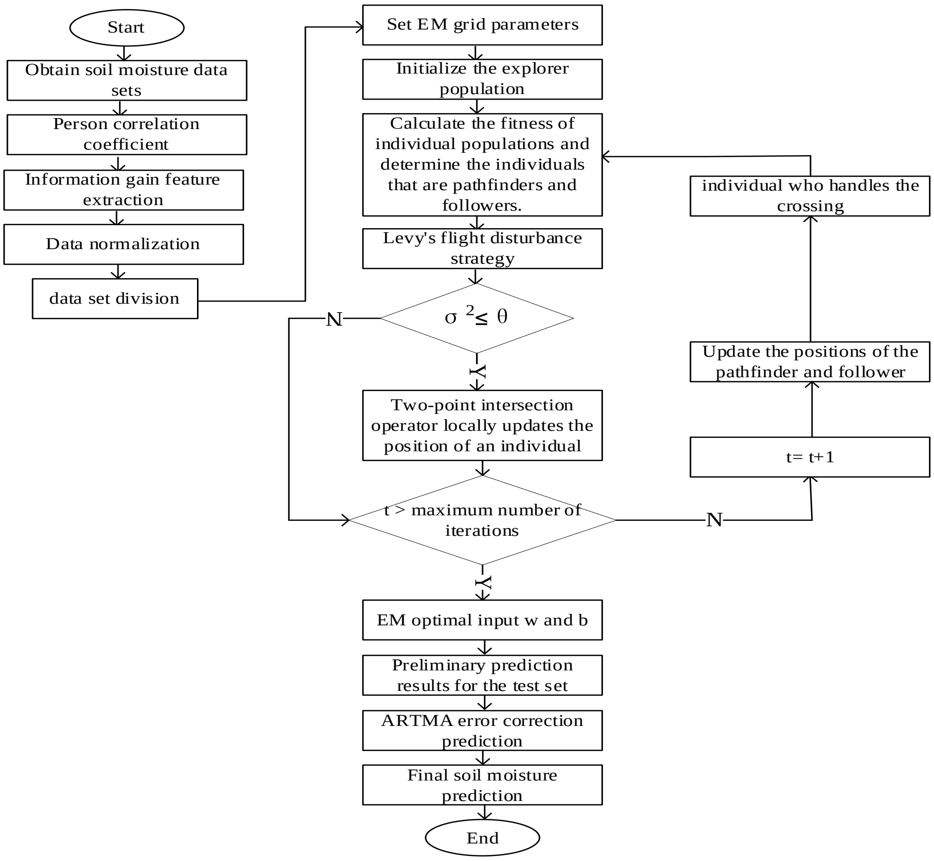

Start
Obtain soil moisture data sets
Person correlation coefficient
Information gain feature extraction
Data normalization
data set division
Set EM grid parameters
Initialize the explorer population
Calculate the fitness of individual populations and determine the individuals that are pathfinders and followers.
Levy's flight disturbance strategy
σ ²≤ θ
N
Y
Two-point intersection operator locally updates the position of an individual
t > maximum number of iterations
N
Y
EM optimal input w and b
Preliminary prediction results for the test set
ARTMA error correction prediction
Final soil moisture prediction
End
individual who handles the crossing
Update the positions of the pathfinder and follower
t= t+1

Figure 3. Soil moisture prediction process

RESULTS AND ANALYSES

In this paper, the experimental research is based on Python programming platform, the sample data before January 2020 is utilized as the training data set, and the rest of the data is used as the test set for the model prediction performance test. CLPPFA_ELM model adopts as the parameter of 105, as the parameter of 45, as the parameter of 100, and as the parameter of 10-4, and use the sigmoid function as the activation function. This experiment refers to the ideas of the “trial and error method” and the “grid search method”. The number of neurons in the hidden layer of the ELM model is selected between the intervals [1, 50]. In order to reduce the randomness of the optimization, the author compares the fitness values of each parameter setting through five repeated trials and determines that the optimal number of neurons is 15. The global optimisation capability and convergence speed of the CLPPFA algorithm is evaluated by comparing the changes in the fitness of the CLPPFA and PFA optimised ELM. The variation of fitness is shown in Figure 4.

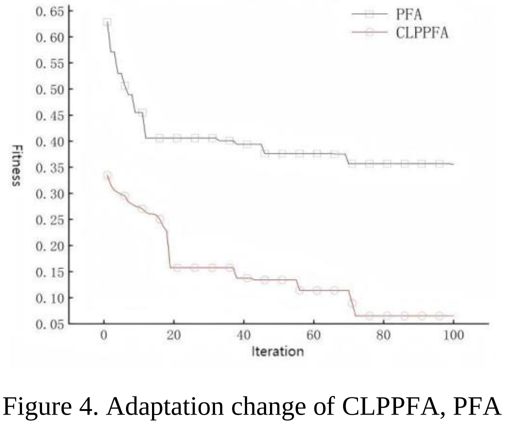
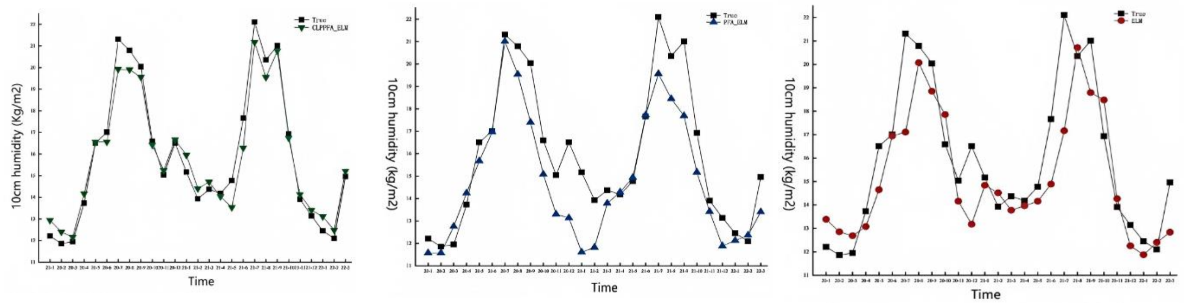


Figure 4. Adaptation change of CLPPFA, PFA

As can be seen from Figure 4, under the same experimental sample conditions, the CLPPFA algorithm has a smoother change in fitness value and a faster decrease in fitness value throughout the iteration process, and finally reaches the lowest fitness value: 0.0651, which shows better convergence speed and global optimisation capability. The results show that the CLPPFA algorithm performs best in optimising ELM with faster convergence speed and better global optimisation ability.

On the basis of information gain processing, this experiment evaluates the superiority of CLPPFA_ELM model in grassland soil moisture prediction by comparing the differences between the prediction results and the real values of CLPPFA_ELM, PFA_ELM and ELM models. The experimental results are shown in Figure 5.



(a) CLPPFA_ELM (b) PFA_ELM (c) ELM

Figure 5. Comparison chart of model prediction results of CLPPFA_ELM, PFA_ELM, ELM

According to the experimental results, Figure 5 presents the results of predicting the 10 cm soil moisture from January 2020 to March 2022. The results show that the fitting effect of the CLPPFA_ELM model has been significantly improved compared to the PFA_ELM and ELM models, and the CLPPFA_ELM model has improved the accuracy of soil moisture. After analysing and calculating, the errors between the measured values and the predicted results obtained using the above kinds of methods are shown in Table 1.

Table 1. Compare model error results

	EVS	MAE	RMSE	R^2
PFA_ELM	0.847	1.256	1.649	0.719
ELM	0.731	1.315	1.79	0.669
CLPPFA_ELM	0.938	0.517	0.639	0.937

As can be seen from Table 1, all the indicators of CLPPFA_ELM model are better than the unoptimised ELM model and the unimproved PFA_ELM model, in which the R2 reaches 0.937, the optimised algorithm improves the model effect and the fitting error has a significant improvement.

The ARIMA model is used to analyze the error sequence between the actual soil moisture value and the preliminary prediction result of CLPPFA_ELM. The ADF test method is used to test the stationarity of the error sequence. The specific test results can be found in Table 2. It can be found from the Table that when the order of the difference is 0, the significance p-value is 0.044**, which is significant at this level, the null hypothesis is rejected, and this sequence is a stationary time series.

Table 2. ADF test results

difference in order	t	P	AIC	threshold value		
				1%	5%	10%
0	-2.911	0.044^{**}	15.823	-3.809	-3.022	-2.651
1	-3.816	0.003^{***}	21.577	-3.724	-2.986	-2.633
2	-7.557	0.000^{***}	24.396	-3.738	-2.992	-2.636

The ARIMA (2,0,2) model was established for the grassland soil moisture series based on the minimization information criterion (AIC) by selecting the 0th order difference series, and the model test results are shown in Table 3. The predicted values of the error correction of the ARIMA (2,0,2) model were superimposed on the preliminary results of the CLPPFA_ELM model in order to correct the preliminary prediction results to obtain the final prediction results. As can be seen from Figure 6, the error correction prediction can be effectively carried out by ARIMA model to improve the prediction accuracy of the model.

Table 3. Model parameter results

	ratio	(statistics) standard deviation	t	P>|t|
a constant (math.)	0.054	0.021	2.571	0.010**
ar.L1	1.656	0.101	16.46	0.000 ***
ar.L2	-0.931	0.105	-8.826	0.000 ***
ma.L1	-1.259	0.465	-2.71	0.007 ***
ma.L2	0.293	0.365	0.803	0.422

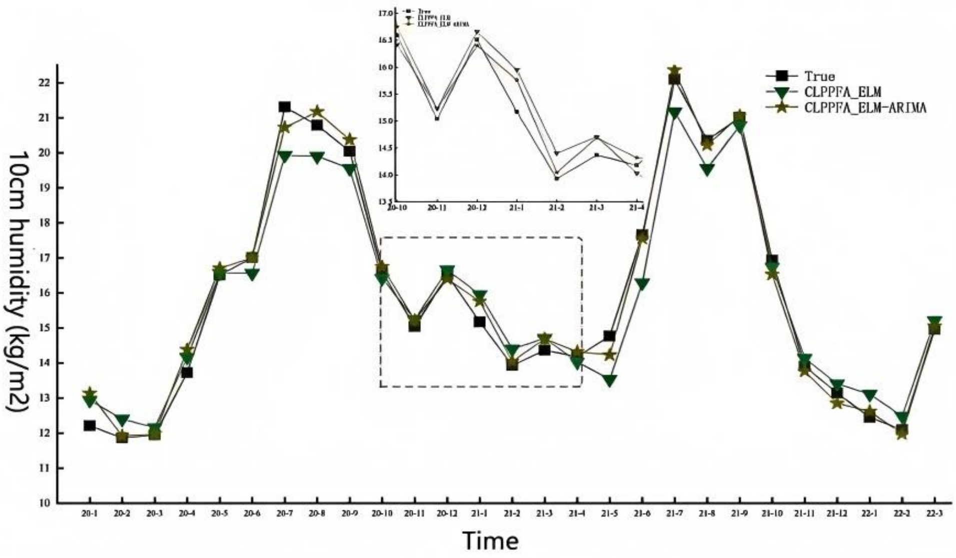


Figure 6. Error correction results

In order to further verify the advantages of CLPPFA_ELM-ARIMA model in grassland soil moisture prediction, SVR, RF, BP, and ridge regression were used as comparative tests for differential comparison of prediction effects in this experiment. The experimental comparison effect is shown in Figure 7.

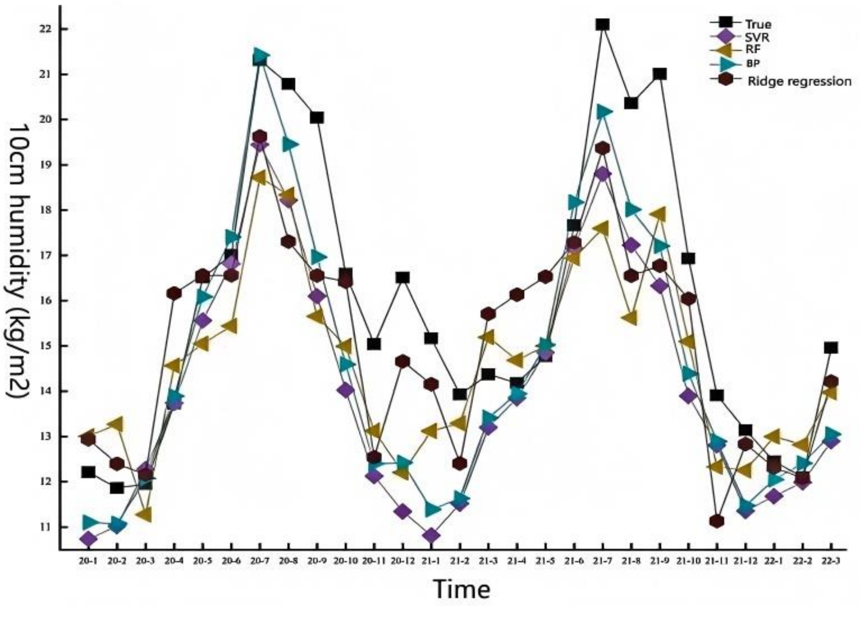


Figure 7. Comparative results of different models

From the waveform of Figure 7, the CLPPFA_ELM model and CLPPFA_ELM_ARIMA model fit the real value better, in which the prediction result of CLPPFA_ELM_ARIMA model is smoother and closer to the real value, especially in the moment when the data fluctuates greatly. The ridge regression and RF models can predict the overall trend of soil moisture to a certain extent, but the prediction effect deviates more from the actual value; SVR and BP neural network models are even unable to predict the overall trend of soil moisture when the data fluctuates greatly. It can be seen that the CLPPFA_ELM model proposed in this paper and the model modification of ARIMA can predict grassland soil moisture more accurately.

To further illustrate the advantages of this paper's model in grassland soil moisture prediction, the errors of the models on the test set were analysed comparatively, and the explained variance scores (EVS), mean absolute error (MAE), root mean square error (RMSE), and coefficients of determination (R2) of the models are presented in Table 4.

Table 4. Model error comparison

	EVS	MAE	RMSE	R^2
CLPPFA_ELM	0.938	0.517	0.639	0.937
SVR	0.758	1.857	2.343	0.411
RF	0.681	1.735	2.145	0.506
BP	0.807	1.386	1.783	0.668
mountain ridge return	0.675	1.914	1.974	0.598
CLPPFA_ELM-ARIMA	0.988	0.262	0.343	0.988

As can be seen from Table 4, the explained variance (EVS) of the CLPPFA_ELM_ARIMA model was 0. 988, which was much higher than that of the other models, indicating that the model developed in this paper has a high degree of explanation for the fluctuations of grassland soil data. The coefficient of determination of the coefficient of determination (R2) of the CLPPFA_ELM_ARIMA model was 0. 988, which was higher than that of the CLPPFA_ELM, SVR, RF, BP and ridge regression models by 0.051, 0.577, 0.482, 0.32, 0.39, respectively; the mean absolute error (MAE) decreased by 0.255, 1.595, 1.473, 1.124, 1.652, respectively; and the root mean square error (RMSE) decreased by 0.296, 2, 1.802, 1.44, respectively, 1.631, further validating the validity and superiority of the CLPPFA_ELM_ARIMA model proposed in this paper.

RESULTS AND ANALYSES

In order to achieve accurate prediction of grassland soil moisture, this paper proposes CLPPFA_ELM_ARIMA model for grassland soil moisture prediction considering the information gain of features and model error, and draws the following conclusions through the experimental verification analysis:

(1) Compared with traditional dimensionality reduction means such as PCA, the introduction of information gain can effectively retain the original data features. Using high information gain features as input features for the prediction model can reduce data latitude and improve the prediction accuracy of the model.

(2) The transgressing individuals in the population are effectively dealt with in PFA by evolving the boundary constraint processing scheme, and the Levy flight strategy and population fitness variance strategy are introduced to prevent the PFA algorithm from falling into the convergence problem prematurely. The CLPPFA_ELM model based on the information gain processing has significant improvement in all comparative indicators, and has better prediction effect compared with PFA_ELM and ELM. The results show that CLPPFA optimised ELM has high effectiveness and can effectively improve the prediction performance of ELM algorithm.

(3) In this paper, the ARIMA algorithm is used to correct the error values of the initial prediction of the CLPPFA_ELM model, and then the CLPPFA_ELM-ARIMA model is established. It is found through example validation that: the prediction effects of the adopted CLPPFA_ELM-ARIMA model are all better than those of ridge regression, SVR, RF and BP models. The experiments showed that the CLPPFA_ELM-ARIMA model based on information gain and error correction can effectively and accurately predict grassland soil moisture.

ACKNOWLEDGMENT

This research was funded by the National Natural Science Foundation of China, grant number 71771111.

REFERENCES

[1] L Huang, J H Li, H Y Zhang, et al. Connotation, accounting and assessment of grassland ecological value. Journal of Grassland Industry, vol:33, no:06, pp:47-63, 2024.

[2] Y W Chang, Y H Wu, X Liu, et al. Characteristics of soil moisture cycle in desert grassland based on wavelet analysis. China Grassland Journal, vol:45, no:09, pp:87-97, 2023.

[3] Y Q Liu, L Lu, P B Liu, et al. Population genetic diversity and population genetic structure of Brandt's vole in three regions of Inner Mongolia Autonomous Region. Chinese Journal of Vector Biology and Control, vol:34, no:03, pp:291-297, 2023.

[4] Q Zhang, L Shi, Holzman M, et al. A dynamic data-driven method for dealing with model structural error in soil moisture data assimilation. Advances in Water Resources, vol:132, pp:1-17, 2019.

[5] Abioye E A, Abidin M S Z, Mahmud M S A, et al. IoT-based monitoring and data-driven modelling of drip irrigation system for mustard leaf cultivation experiment. Information Processing in Agriculture, vol:8, no:2, pp:270-283, 2021.

[6] Q Zhang, S Z Huang, X H Chen. Research on soil moisture simulation and prediction based on support vector machine. Journal of Soil Science, vol:50, no:01, pp:59-67, 2013.

[7] Taneja P, Vasava H K, Daggupati P, et al. Multi-algorithm comparison to predict soil organic matter and soil moisture content from cell phone images. Geoderma, vol:385, pp:1-15, 2021.

[8] L P Yang, Cl Hou, Z Q Su, et al. Soil moisture inversion in arid areas based on machine learning and fully polarised radar data. Journal of Agricultural Engineering, vol:37, no:13, pp:74-82, 2021.

[9] H S Wu, L L Liu, C Y Zhang, et al. GNSS-IR soil moisture inversion combining wavelet noise reduction and BP neural network. Remote Sensing Information, vol:37, no:02, pp:119-125, 2022.

[10] Prasad R, Deo R C, Li Y, et al. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma, vol:330, pp:136-161, 2018.

[11] Adeyemi O, Grove I, Peets S, et al. Dynamic neural network modelling of soil moisture content for predictive irrigation scheduling. Sensors, vol:18, no:10, pp:1-22, 2018.

[12] Filipović N, Brdar S, Mimić G, et al. Regional soil moisture prediction system based on Long Short-Term Memory network. Biosystems engineering, vol:213, pp:30-38, 2022.

[13] Bates J M, Granger C W J. The combination of forecasts. Journal of the operational research society, vol: 20, no:4, pp: 451-468, 1969.

[14] Andrawis R R, Atiya A F, El-Shishiny H. Combination of long term and short term forecasts, with application to tourism demand forecasting. International Journal of Forecasting, vol:27, no:3, pp:870-886, 2011.

[15] J Yu, S Tang, Z Z L, et al. A deep learning approach for multi-depth soil water content prediction in summer maize growth period. IEEE Access, vol:8, pp:199097-199110, 2020.

[16] ElSaadani M, Habib E, Abdelhameed A M, et al. Assessment of a spatiotemporal deep learning approach for soil moisture prediction and filling the gaps in between soil moisture observations. Frontiers in artificial intelligence, vol:4, pp:1-14, 2021.

[17] B Sun, Y Liang, M T Han, et al. GNSS-IR soil moisture inversion method based on GA-SVM. Journal of Beijing University of Aeronautics and Astronautics, vol:45, no:03, pp:486-492, 2019.

[18] X Q Li, L L Liu, Z L Liu, et al. PSO-ELM-assisted GNSS-IR soil moisture inversion method. Radio Engineering, vol:53, no:06, pp:1368-1374, 2023.

[19] S M Xing, X L Gao, Z M Lin, et al. A prediction model of out-of-plant moisture content of corn drying system based on limit learning machine. Journal of Shenyang Agricultural University, vol: 54, no:05, pp:619-626, 2023.

[20] Yapici H, Cetinkaya N. A new meta-heuristic optimizer: the Pathfinder algorithm. Applied soft computing, vol: 78, pp:545-568, 2019.

[21] Gandomi A H, Yang X S. Evolutionary boundary constraint handling scheme. Neural Computing and Applications, vol:21, pp:1449-1462, 2012.

[22] Mantegna. Fast, accurate algorithm for numerical simulation of Levy sTable stochastic processes. Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics, vol: 49, no:5, pp:4677-4683, 1994.

[23] D P Tian, T X Zhao. Adaptive particle swarm optimisation algorithm based on population fitness variance. Computer Engineering and Applications, vol:46, no:18, pp:24-26+39, 2010.