Computer Fraud and Security
ISSN (online): 1873-7056

Deep Learning Prediction Model of Drought and Flood in Summer Based on Random Forest and Attention Mechanism

Jie Li[1], Hong Lu[2,*], LinLi Jiang[1,*], Long Jin[2]
[1]Guangxi Science & Technology Normal University, Laibin, China
[2]Guangxi Zhuang Autonomous Region Climate Center, Nanning, China
**Corresponding Author.*

Abstract:
The traditional linear statistical forecasting approach is frequently utilized to tackle the significant interannual variability and nonlinear traits of summer drought and flood patterns in Guangxi. However, its forecasting precision tends to be inadequate. In this research, the year-to-year change in the average precipitation for the month of August was employed as a means to predict trends in droughts and floods. The correlation factor between the forecast and the previous 500 hPa monthly average height field was calculated, yielding 81 preliminary forecast factors for the early monthly average circulation field. First, the random forest algorithm was utilized to assess and prioritize the significance of the 81 predictive elements and indicators. Subsequently, the top six crucial variable components were chosen to feed into the deep learning LSTM network's forecasting framework. Second, an attention mechanism was employed to provide different attention values to the input variables of the model. Third, a forecasting model for the year-to-year change in average summer rainfall in Guangxi was constructed, incorporating a random forest and an attention mechanism-enhanced LSTM network (RF-LSTM-Attention). Upon applying this predictive model to forecast the average summer rainfall for the eight-year period in Guangxi (June–August) from 2013 to 2020, the model exhibited an average absolute percentage error of 9.49%. The relative error of prediction in the six years, especially in the years with maximum and minimum precipitation in the eight-year return sample, did not exceed 15%. The RF-LSTM-Attention model was approximately doubled and showed better forecast accuracy in qualitative and quantitative forecasting.

Keywords: drought and flood prediction model, Long Short-Term Memory, attention mechanism, Random Forest (RF)

INTRODUCTION

Droughts and floods are among the climate disasters that exert the greatest impact on the national economy. Flood disasters occur yearly in Guangxi and tend to increase with time. Therefore, accurate identification, research, and analysis of the influencing factors of summer precipitation in Guangxi and determining the characteristics of precipitation data from historical data are essential to improve the accuracy of summer precipitation forecasting in Guangxi and have practical guiding importance for the effective prevention and mitigation of disaster impacts.

However, short-term climate forecasting of droughts and floods caused by precipitation changes on monthly and seasonal time scales is a key, difficult issue in atmospheric science [1-3]. This difficulty is mainly due to the insufficient understanding of the physical mechanisms that affect the inter-annual variation of average precipitation on monthly and seasonal time scales. Therefore, the forecast products predicted by short-term climate models cannot easily meet the requirements of operational forecasting. At present, short-term climate forecasting of drought and flood disasters on monthly and seasonal time scales still relies heavily on various statistical forecasting methods [4-6] or combined dynamic–statistical forecasting methods [7,8]. In addition to the calculation and analysis using these forecasting methods and the utilization of physical quantity forecasting factors, such as the average circulation and sea temperature fields in the early stage of drought and flood disasters on the corresponding time scale [9,10], many researchers in recent years have carried out research on the use of the inter-annual increment of the mean precipitation series on monthly and seasonal time scales [11,12] as a new forecast object to forecast the monthly and seasonal mean precipitation in different regions [13,14]. Fan and Wang et al. conducted the earliest research on short-term climate prediction of precipitation in the middle and lower reaches of Yangtze River (June–August) by analyzing the correlation between the interannual increment of summer average precipitation and the previous winter and spring atmospheric circulation[11]. Six key prediction factors were identified, and a regression prediction model was established. In

the 10-year post-forecast from 1997 to 2006, the average root-mean-square error was 18%, and the model had a good prediction effect on the trend. Zheng et al. also utilized the inter-annual increment of precipitation in summer (June–August) in southwestern China as the forecast value to determine the five factors affecting the average atmospheric circulation in the previous period and established a regression forecast model[12]. The average root-mean-square error of the 2011–2017 precipitation anomaly hindcast test was 16%, and the forecast accuracy was higher than the local operational forecast comprehensive score. Deng et al. and Wang et al. used interannual increments as prediction objects to study the short-term climate prediction of precipitation in the flood season and the start date of the rainy season and achieved good results[12,13]. However, the linear regression statistical forecasting method was used in the forecast modeling of these studies. Short-term statistical climate forecasting methods generally need to consider different early-stage physical quantity factors, and each factor may find a physical mechanism explanation related to the forecasting quantity. However, the comprehensive influence of these multiple factors on forecasting is not likely to have a simple linear relationship. Hence, nonlinear forecast modeling methods must be used to conduct short-term climate forecast modeling of drought and flood disasters on monthly and seasonal time scales and inter-annual increments.

With the rapid development of various machine learning methods, such as neural networks and support vector machines [15,16], increasingly complex nonlinear deep learning models have been applied to forecast modeling research in many disciplines[17]. Fan et al. used deep long short-term memory (LSTM) to predict and interpolate PM2.5 in the Beijing–Tianjin–Hebei Urban Agglomeration[18]. BT et al. combined auto-encoding and recurrent neural network (RNN) to predict environmental pollution by PM2.5[19]. All of these studies captured the characteristics of sequence data through LSTM training and achieved good effects. In atmospheric science, extensive research has been conducted on the forecasting of meteorological disasters, such as typhoons and cold damage, by using machine learning methods, including fuzzy neural network and genetic neural network [20-22].

When using the intelligent calculation method of deep learning for prediction modeling, we often face the problems of how to analyze the importance of different physical quantity prediction factors from many primary prediction factors and how to assign different weights to different prediction factors. The random forest algorithm [23,24] has attracted attention because it can extract important feature sets with the least redundancy from the original data set as the model input. Xiong et al. used the random forest algorithm to select factors for a high-dimensional teleconnected climate factor set, identified the predictors that had a great impact on runoff on the basis of the importance of the variables, and built a support vector machine model for forecasting[23]; they obtained good forecasting results. Li et al. used random forest to select key prediction factors from the circulation index, sea surface temperature, air pressure, and early monthly runoff[24]. They optimized parameters based on particle swarm optimization and a cross-validation algorithm, established a random forest and support vector machine model, and conducted inflow runoff prediction of Longjiang Reservoir. They achieved good prediction accuracy. However, these methods do not consider the changing characteristics of eigen factors' attention to different prediction information.

There is a close relationship between heavy precipitation and precipitation characteristics. Feng and Zhai used the Butterworth filtering method to analyze the low-frequency precipitation characteristics in China and the relationship between regional sustained heavy precipitation and low-frequency precipitation in different regions of eastern China in the summer half year. The computational methods of various attention mechanisms [25-27]) have also attracted the attention of prediction researchers in different disciplines and have been applied to the classification and prediction problems of neural networks with remarkable results. By incorporating an attention mechanism into the deep learning neural network model, the model can pay more attention to the information that has a greater impact on the prediction results and pay less attention to the information that has less impact on the classification prediction, thereby improving the accuracy of model prediction. Li and Lu proposed a short-term load prediction method based on a dual attention mechanism and GRU (Gated Recurrent Unit). The attention mechanism layer was established for features and time series. The attention mechanism feature independently analyzed the correlation between historical information and input features and extracted important features. The time series attention mechanism independently selected the historical information of the GRU network key time points to improve the long-term stability of the prediction effect and increase the prediction accuracy of the model. Peng Y and Qiao Y et al. proposed a PM2.5 prediction model incorporated

with an attention mechanism, which effectively utilizes the advantages of long-term and short-term memory to memorize long-term information in time series[25]. The model learns the weight distribution between multi-feature factors and PM2.5 through the local attention mechanism to focus on information that has a great impact on the PM2.5 value, thus improving the final prediction effect.

Based on the analysis provided above, the present study adopted the interannual increment series of the summer (June–August) average precipitation of 87 stations in Guangxi as the prediction quantity, calculated and selected the physical quantity prediction factors of the average annular flow field in the early stage of inter-annual increment change, and attempted to identify the characteristic factors of the high-dimensional physical quantity prediction factor set by using the random forest algorithm. Then, an LSTM deep learning neural network prediction model that incorporates the feature attention mechanism and can assign different weights to the predictors of different importance physical quantities was built to explore the prediction effect of this model and verify its validity and applicability in applied research on short-term climate drought and flood disaster prediction.

MATERIALS AND METHODS

Materials

The precipitation data is the average precipitation observation data for 63 years from June to August, 1959 to 2022, from 87 national meteorological observation stations in Guangxi; The atmospheric circulation data is reanalysis data from the National Center for Environmental Forecasting and the Center for Atmospheric Research (NCEP/NCAR) in the United States, including monthly potential height fields, wind fields, and a horizontal resolution of 2.5°X2.5°, a total of 81 forecast factors.

Method

This article refers to the interannual increment method defined by Fan et al. using the average precipitation interannual increment (the difference between the average precipitation from June to August of the current year and the average precipitation from June to August of the previous year) of 87 basic stations in Guangxi from 1960 to 2020 (June to August) as the forecast quantity, and 81 data from reanalysis of atmospheric circulation data as the forecast factors[11]. Random forest is used for important feature factor selection, and deep learning long short-term memory (LSTM) network is used as the basic model, And further embed attention mechanisms to examine the relationship between different feature factors and output, assign different weights, and dynamically train and adjust them. Establish a deep learning (RF-LSTM Attention) prediction model for summer precipitation increment sequence using random forest combined with attention mechanism, and predict the summer precipitation in Guangxi.

PRINCIPLES AND METHODS OF DEEP LEARNING MODELS FOR FORECAST MODELING

Basic Model Principles of Forecast Modeling

In this study, the LSTM deep learning model, which is widely used in many disciplines, was employed as the basic model in establishing the prediction model of deep learning with random forest combined with an attention mechanism (RF-LSTM-Attention) for predicting summer precipitation increment sequences. The model is an improved recurrent neural network (RNN). The LSTM model can transfer information from the previous unit to the next unit through the transfer of the state of the cell unit. Unlike RNN, each cell unit of LSTM introduces three special gating units to control the memorization and forgetting of information [25] and to solve the problem of gradient disappearance and explosion that may occur in RNNs during long sequence training. Therefore, LSTM can perform well on long sequences. The internal configuration of an LSTM cell unit is illustrated in Figure 1.

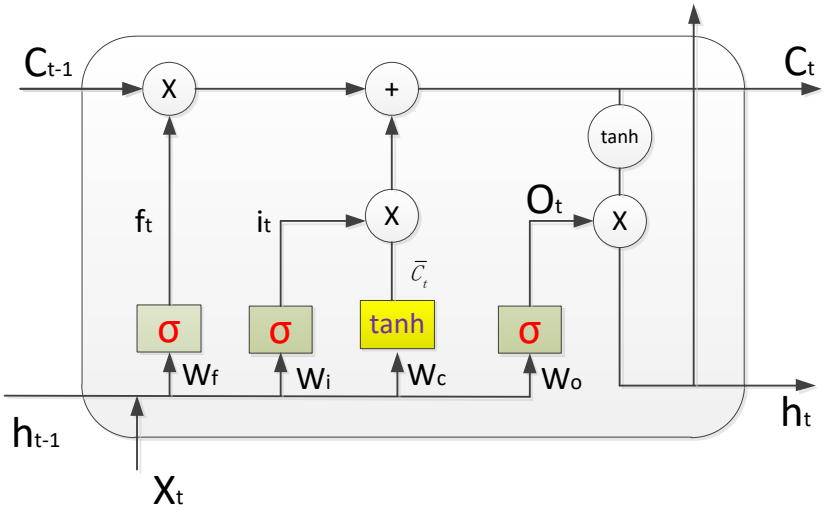

Figure 1. Internal structure of an LSTM unit

Compared with RNN that has only one transfer state h_t, LSTM has two transfer states h_t (hidden layer state) and C (unit state). In Figure 1, X_t and h_t are the input and hidden layer output of the LSTM unit at time t, h_{t-1} is the hidden layer output of the unit at the previous moment, C_{t-1} is the unit state at the previous moment, and C is the cell state of LSTM at time t. The three special gates of LSTM receive these information to realize the memorization and forgetting of the information.

The forget gate reads the output h_{t-1} of the previous moment and the input X_t of the current moment and outputs a forgetting factor f_t in the range of 0–1 through the sigmoid activation function, where 0 means "completely discarded" and 1 means "complete retention," which is used to control the degree to which unit state C_{t-1} is forgotten at the previous moment. The calculation formula is as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{1}$$

where W is the weight matrix input to the LSTM unit, b_f is the corresponding offset, and σ is the sigmoid activation function.

Then, the input gate is also calculated by the sigmoid activation function i_t to control the influence of the current input on the unit state, and activation function $\tanh$ is used to calculate a new candidate memory unit vector $\tilde{C}_t$ for the current moment to determine how much new information is added.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c). \tag{3}$$

Next, the new unit state C at the current moment is calculated, and the unit state at the previous moment is dot-multiplied by the forgetting factor one by one. When the dot-multiplication value is close to 0, the new information is discarded in the unit state. The output of the input gate is then obtained and added point by point to derive the current new unit state value. The calculation formula is as follows:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t. \tag{4}$$

Afterward, the output gate calculates the current output information through a sigmoid activation function then transforms it through the tanh function to obtain the final output of the hidden layer of the unit. The specific procedure is shown as

$$\begin{aligned} O_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t &= O_t \odot \tanh(C_t) \end{aligned}. \tag{5}$$

The LSTM neural model completes the training and learning of the training data with the neural network by inputting a large amount of training data, specifying the target function, adjusting the connection weights between neurons, and gradually reducing the output error function to a specified value [26].

Importance Analysis and Selection of Input Variables of Forecast Models

Usually, when establishing a linear or nonlinear mean precipitation drought and flood disaster forecasting model, importance analysis and selection of a large number of physical quantity predictors, such as the mean circulation field and sea temperature field, must be performed. This study adopted a random forest learning algorithm using a bagging algorithm to integrate multiple decision trees [27-29]. This method has a strong nonlinear processing ability and can effectively select the features of multivariable, high-dimensional feature data sets. By calculating the importance of a single feature variable and ranking the importance of the feature variable, the most important feature variable factor set for the prediction model is obtained. The important idea of its algorithm is to form a training subset by randomly sampling n training samples as bagged data with playback on the original high-dimensional set of data samples, and the samples that are not sampled for sampling as out-of-bag samples and as a test set. The random sampling of L rounds of samples is repeated to form L training subsets so that L decision trees can be built in the original high-dimensional feature data set. After the L decision trees have grown, the random forest algorithm integrates the predictions of each individual tree and votes on the results of several weak learners to form a strong learner. By integrating numerous weak learnershigher generalization performance can be obtained than that for a single learner.

For the random forest algorithm, the variable importance measure (VIM) is mainly used to evaluate the contribution of each feature assigned to each tree within the random forest, and the contribution can be assessed through evaluation indicators, such as out-of-bag data error or Gini coefficient. Formula (6) can be used to calculate the mean-square error of out-of-bag data for each individual decision tree [30].

$$MSE_j^{OOB} = \frac{\sum_{i=1}^{n_{OOB_j}} (y_i^j - y_i^{OOB_j})^2}{n_{OOB_j}}, \quad j = 1, 2, \ldots, L \tag{6}$$

where MSE_j^{OOB} is the OOB mean-square error of the jth tree, y_i^j is the actual OOB value observed for the jth tree, $y_i^{OOB_j}$ is the predicted OOB value of the jth tree, and n_{OOB_j} is the amount of data outside the bag of the jth tree.

Noise is randomly added to each OOB feature of out-of-bag data $X_k, k = 1, 2, \ldots, m$, and the out-of-bag data error of each decision tree is calculated again to obtain the mean-square deviation $\left\{MSE_{kj}^{OOB}, k = 1, 2, \ldots, m; j = 1, 2, \ldots, L\right\}$ as follows:

$$\begin{bmatrix} MSE_{11}^{OOB} & MSE_{12}^{OOB} & \cdots & MSE_{1L}^{OOB} \\ MSE_{21}^{OOB} & MSE_{22}^{OOB} & \cdots & MSE_{2L}^{OOB} \\ \vdots & \vdots & \vdots & \vdots \\ MSE_{m1}^{OOB} & MSE_{m2}^{OOB} & \cdots & MSE_{mL}^{OOB} \end{bmatrix}. \tag{7}$$

Thus, the calculation formula of the K-th feature variable importance score is obtained as

$$VIM_k = \frac{\frac{1}{L}\sum_{j=1}^{L}(MSE_j^{OOB} - MSE_{kj}^{OOB})}{S_E} \tag{8}$$

where S_E is the root-mean-square error of all L regression trees, as follows:

$$S_E = \sqrt{\frac{\sum_{j=1}^{L} E_j^2}{L}} \tag{9}$$

Formula (8) is used to calculate and select the importance of feature variables for the high-dimensional early physical quantity factors.

Feature Attention Mechanism

In the actual prediction and modeling of summer drought and flood disasters, when some important characteristic variables are selected and determined as the input matrix of the forecast model by different methods, the values of these characteristic factors change with time and have different importance. The effects of the characteristic variables on the forecast results of the forecast model also differ. Therefore, the feature attention algorithm was added in this study [25,30]. The LSTM network was used to analyze the output of the hidden state at the previous time and the input of the characteristic factors at the current time, construct the characteristic attention mechanism layer based on the LSTM neural network, calculate and analyze the attention weight to reflect the importance of the input characteristics at the current time to the prediction data, and optimize the influence weight coefficient of each characteristic factor by means of neural network training in order to improve the learning ability of the LSTM neural network model. The principle of the feature attention mechanism algorithm is shown in Figure 2.

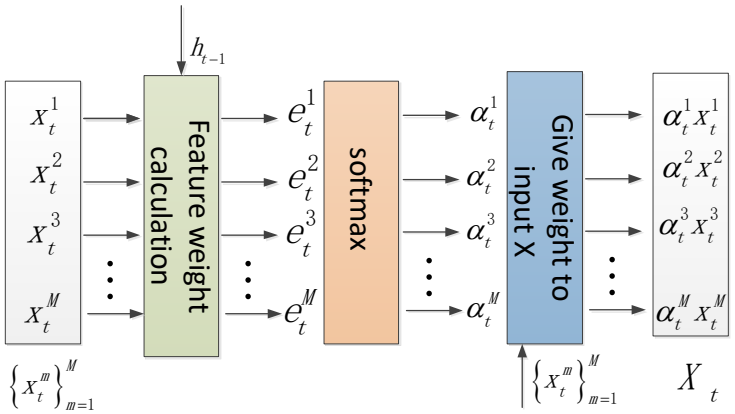

Figure 2. Feature attention mechanism

In the calculation for integrating the LSTM model into the feature attention mechanism, hidden state h_{t-1} outputted by the LSTM network at the previous moment and input feature x_t^m at the current moment were used as the input of the feature attention mechanism layer. Formula (10) was employed to calculate the attention weight of each feature factor at current moment e_t^m to assess the extent of correlation between the current input feature data and the output at the previous moment. Then, Formula (11) was used to perform softmax normalization to obtain attention weight coefficient α_t^m for enhancing or weakening the expression of the input feature data. Finally, in accordance with Formula (12), the weight and corresponding feature were multiplied to obtain the output of the attention value of feature X_t. The influence degree of the input feature was adaptively optimized.

$$e_t^m = \tanh(W_e h_{t-1} + U_e x_t^m + b_e) \tag{10}$$

$$\alpha_t^m = soft\ max(e_t^m) = \frac{\exp(e_t^m)}{\sum_{i=1}^{M} \exp(e_t^i)} \tag{11}$$

$$X_t = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \cdots, \alpha_t^M x_t^M)^T \tag{12}$$

where e_t^m is a measure of the correlation between the current input and the output of the hidden layer at the previous moment. W_e, U_e, and b_e are the weights and biases of the attention mechanism and need to be trained by the LSTM network model. Through a certain number of training, the weights were trained by the historical state output at the previous moment and the feature vector input at the current moment to obtain the importance of different feature factors affecting the average precipitation in the input data set.

PREDICTION EXPERIMENT AND PERFORMANCE ANALYSIS OF THE RF-LSTM-ATTENTION PREDICTION MODEL

This study adopted the method incorporating the random forest algorithm, feature attention mechanism, and LSTM deep learning neural network to construct the average precipitation increment prediction model for Guangxi summer (June–August). The method is mainly composed of an input layer, random forest factor selection, a feature attention layer, an LSTM network hidden layer, and a fully connected output layer. The input vector of the primary physical quantity factor is selected by the random forest algorithm to choose the important eigen factors as the model input. The input feature sequence combined with the hidden state output of the LSTM network at the previous moment is calculated by the feature attention layer to obtain the weight coefficients of each feature factor for the current prediction. The optimized input features improved by combining the attention are used to calculate the output of the LSTM hidden layer and finally inputted to the fully connected layer to derive the final prediction result.The architecture of the forecast model is presented in Figure 3.

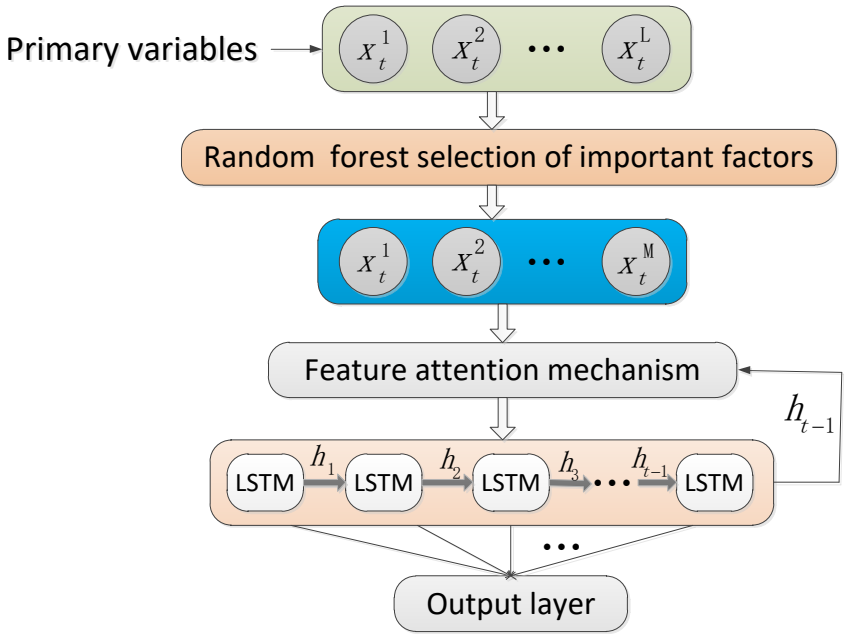

Figure 3. LSTM network model structure based on random forest combined with feature attention

The calculation process of the forecast model can be summarized as follows. Important factors are selected from the original input data through the random forest algorithm then combined with the output of the LSTM network at the previous moment to calculate the weight of the current precipitation forecast, with each feature factor affecting the precipitation input at the current moment through the feature attention mechanism layer. The corresponding weights are assigned to the input features to obtain a new feature input weighted optimization value, and feature learning is performed through LSTM network training while focusing on the importance of the key features of this prediction. Then, the final prediction result of the model is obtained through the fully connected layer.

Performance Evaluation Index of the Prediction Model

When the established RF-LSTM-Attention forecast model of the Guangxi summer precipitation increment sequence was used to carry out forecast experiments, accuracy calculation, and analysis, this study adopted the same evaluation index calculation formulas in the work of Fan et al.[13] and Zheng et al.[14].

(1) Mean-square error (MSE) is calculated as

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{13}$$

(2) Mean absolute error (MAE) is derived as

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right| \tag{14}$$

(3) Relative error percent (REP) measures how far each predicted value deviates from the current observed value, and the formula is

$$RPE = \frac{\hat{y}_i - y_i}{y_i} \times 100\% \tag{15}$$

(4) Mean absolute percent error (MAPE) measures the deviation of the overall predicted value from the observed value as follows:

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{16}$$

where m represents the number of predicted samples, y_i represents the actual observed value of the original data, and $\hat{y}_i$ represents the predicted value.

Forecast Quantity and Forecast Physical Quantity Factor

In the actual forecast calculation, the forecast was based on the inter-annual increment calculated from the average precipitation data of 87 national base stations in Guangxi from 1960 to 2020 (June–August). The calculation method used the average precipitation from June to August of the current year minus the average precipitation from June to August of the previous year. The inter-annual increment, as the prediction object, can obtain better prediction results than using the average precipitation in the past as the prediction quantity [11-13]. Furthermore, a correlation survey was performed between the inter-annual increment of precipitation from June to August and the monthly average 500-hPa geopotential height field from April of that year to early December of the previous year. The correlation area with adjacent correlation grid points greater than 25, that is, the correlation area with an absolute value of correlation coefficient of each grid point greater than 0.25, was adopted as a basic correlation prediction factor. Combination factor calculation was further carried out for the two adjacent large correlation areas. When the correlation coefficients of the two adjacent areas had different signs, the average values of the representative correlations of the two areas were subtracted; when they had the same sign, they were added to improve the correlation coefficients of the relevant physical quantity factors. Through such calculation, 81 predicted physical quantity factors of the 500-hPa average annular flow field in the previous period were obtained (Table 1). They were used as the primary predictor set for the input of the RF-LSTM-Attention prediction model.

Table 1. Correlation coefficient between predicted physical quantity factors and predicted interannual increment of 500-hPa mean annular flow field in 81 early stages

Factor	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
Correlation coefficient	0.49	0.49	0.47	-0.36	-0.39	0.48	0.59	0.42	0.41	0.38	0.37	-0.32	-0.34	-0.43	-0.5	0.54	0.4
Factor	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32	X33	X34
Correlation coefficient	0.36	-0.31	-0.31	0.36	-0.46	0.43	0.54	0.45	0.45	-0.41	0.48	0.51	0.39	0.36	0.52	-0.39	0.43
Factor	X35	X36	X37	X38	X39	X40	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50	X51
Correlation coefficient	0.31	-0.34	-0.38	-0.5	0.45	0.37	-0.4	-0.45	0.41	0.33	0.29	-0.33	-0.39	0.39	0.34	0.33	0.32
Factor	X52	X53	X54	X55	X56	X57	X58	X59	X60	X61	X62	X63	X64	X65	X66	X67	X68
Correlation coefficient	-0.31	0.47	0.45	0.37	0.33	-0.44	-0.45	0.47	0.55	0.46	0.42	0.41	0.4	0.39	0.32	-0.4	-0.45
Factor	X69	X70	X71	X72	X73	X74	X75	X76	X77	X78	X79	X80	X81	Y			
Correlation coefficient	0.36	0.33	-0.3	-0.52	0.5	0.3	0.41	0.42	-0.39	-0.4	-0.4	-0.45	-0.5	1			

Prediction Experiment and Error Analysis of the Prediction Model

On the basis of the 81 primary physical quantity forecasting factors in Section 3.2 (Table 1), the precipitation increment sequences from June to August every summer from 1960 to 2012 was used as a training sample for the learning and training of the forecast model. The data from 2013 to 2020 were used as the test sample for the return prediction experiment. The random forest calculation method was applied to select the six most important characteristic variables (x60=0.55, x53=0.47, x72=-0.52, x7=0.59, x28=0.48, and x58=-0.45) from the 81 predictors as the input of the forecast model, and the attention mechanism was added. The calculation and analysis provided different important factors and different weight coefficients for prediction test calculation.

The prediction model uses the deep neural network framework Keras to build each layer, employs Formula (13)'s MSE as the objective function, and selects the deep neural network learning framework optimizer Adam to adaptively calculate the learning rate of each parameter. With the parameter determined by specifying the TensorFlow fixed seed value of 1008, a forecast calculation analysis of the important hyperparameter combinations in deep learning, including time_ stept, unit_ dim, batch_ size, learning_ rate, and (epochs), was performed.

Usually, in a neural network model, all training samples complete one forward and backward propagation, which is regarded as a training round epoch. Many times of iterative training are needed to learn and obtain the complex relationship in the data. Increasing the training rounds can improve the accuracy of the model, but too many rounds lead to over-fitting. Figure 4 shows the MAE changes of the RF-LSTM-Attention model under different training rounds and the error curve of the network model learning and training performed. When the epochs of the training round reached 100, the average absolute error MAE was the smallest. Therefore, the epoch number was determined to be 100. Figures 5 and 6 show the variation curves of the predicted values of the training samples from 1960 to 2012, and the variation curves of the actual average precipitation when the combination of hyperparameters is [time_ stept=1, unit_ dim=4, batch_ size=1, learning_ rate=0.001, epochs=100], respectively.

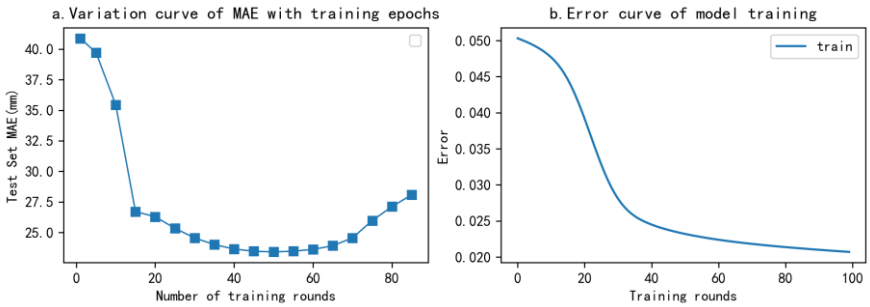

Figure 4. Variation curve of MAE with training epochs and Error curve of model training

From Figure 5, it can be seen that the predicted value calculated by the forecast model for the training samples from 1960 to 2012 is similar to the actual observed value.

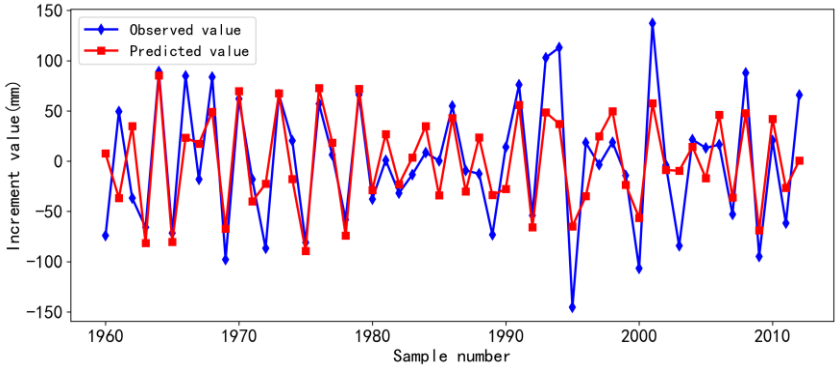

Figure 5. Comparison of prediction results and observed values of interannual increment in 53 years (training set) from 1960 to 2012

With the established RF-LSTM-Attention prediction model for the inter-annual increment of summer precipitation in Guangxi, a return prediction test was further performed on the inter-annual increment of summer precipitation from 2013 to 2020.

Figure 6 presents the six feature variables selected by the random forest algorithm in the LSTM deep learning network prediction model incorporating the attention mechanism and the different average weight coefficients of each variable in the process of predicting eight independent samples (average attention). The first and fourth feature variables selected by the random forest algorithm, namely, x53 and x7, had the highest average attention, and the first feature variable x1 had the lowest average attention. For the correlation coefficient between these variables and the forecast (x60=0.55, x53=0.47, x72=-0.52, x7=0.59, x28=0.48, x58=-0.45), the characteristic variable with the largest linear correlation coefficient did not have the highest attention weight coefficient. If multiple variables affect the forecast at the same time, the nonlinear relationship of the combined effects of the multiple variables may be revealed. Figure 7 shows the changes in the attention coefficients of the six feature variables per year for the post-forecast eight-year independent sample. For each year's post-forecast, the attention coefficient of each feature was affected by the correlation between the output at the previous moment and the input at the current moment.

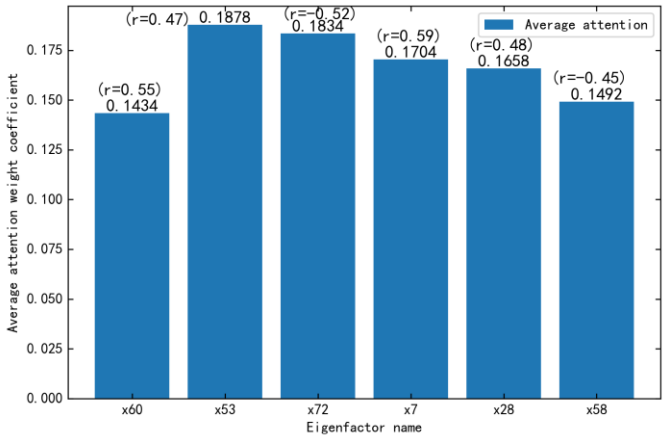

Figure 6. Histogram of the average attention coefficients of six features in the post-forecast eight-year interannual increment

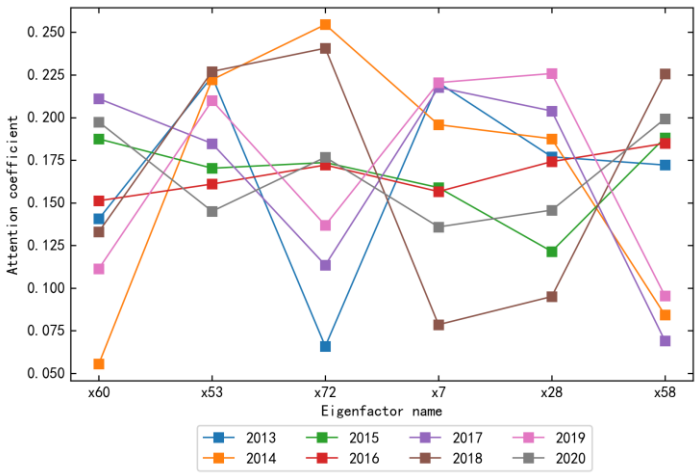

Figure 7. Variation of the attention coefficients of six features in the post-forecast eight-year interannual increment from 2013 to 2020

Next, the precipitation in each year was calculated using the inter-annual increment predicted by the model and the precipitation in the previous year. A comparison of the predicted value of the eight-year precipitation and the actual observed value is shown in Figure 8. The eight-year predicted value curve had good agreement with the actual value curve. The same error evaluation formulas (13–16) as those of Zheng et al.[12] and Fan et al.[13] were used to calculate the statistical return precipitation forecast error, and the results are shown in Table 2. The

relative error of the eight-year return forecast did not exceed 15% in six years. However, the value was 30.74% for 2014 and 19.65% for 2013, and the predicted MAE percentage was 9.49%. The model predicted an average absolute error of 23.96 mm from 2013 to 2020 and had good prediction accuracy in the year with the maximum precipitation (2017) and the year with the minimum precipitation (2016) in the eight-year return sample. This prediction relative error can meet actual business requirements. Evidently, the proposed RF-LSTM-Attention model has good accuracy in predicting the average precipitation of each year and shows a good application prospect for short-term climate drought and flood trend prediction.

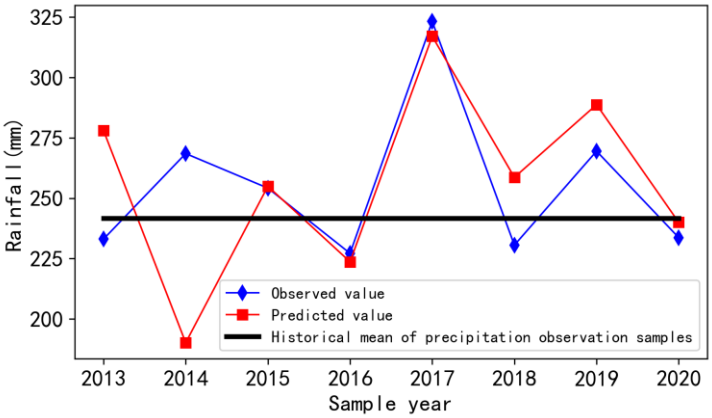

Figure 8 Comparison of predicted results and observed values of precipitation in eight years (test set) from 2013 to 2020

Comparative Analysis of the RF-LSTM-Attention Prediction Model and Other Methods

The effect of adding the attention mechanism in the RF-LSTM prediction model was further investigated to analyze the prediction performance of the summer precipitation prediction model established in this study and its difference with other nonlinear and linear prediction methods.

Table 2. List of precipitation forecast results and statistics of relevant indicators of RF-LSTM-Attention and RF-LSTM models from 2013 to 2020

Data Forecast year	Precipitation observed value (mm)	RF-LSTM-Attention Model precipitation forecast value (mm)	RF-LSTM-Attention model prediction error value (mm)	RF-LSTM-Attention Model relative error percentage REP (%)	RF-LSTM Model precipitation forecast value (mm)	RF-LSTM model prediction error value (mm)	RF-LSTM Model relative error percentage REP (%)
2013	233.1	278.91	45.81	19.65	277.3	44.2	18.96
2014	268.5	185.97	-82.53	-30.74	152.56	-115.94	-43.18
2015	254.2	254.55	0.35	0.14	274.31	20.11	7.91
2016	227.2	223.26	-3.94	-1.73	232.19	4.99	2.2
2017	323.2	317.54	-5.66	-1.75	310.63	-12.57	-3.89
2018	230.6	258.72	28.12	12.2	241.42	10.82	4.69
2019	269.5	288.28	18.78	6.97	273.33	3.83	1.42
2020	233.7	240.19	6.49	2.78	244.81	11.11	4.75
Mean Absolute Error (mm)		23.96			27.95		
Mean Absolute Percentage Error (%)		9.49			10.88		

First, the prediction performance of the RF-LSTM prediction model with and without the attention mechanism was compared and analyzed. For an objective comparative analysis, in the RF-LSTM prediction model without the attention mechanism, the same six important characteristic variables selected by the random forest method in Section 3.2 were used as the model input. The training and return prediction samples were also 53 samples from 1960 to 2012 and 8 samples from 2013 to 2020, respectively. Under the same TensorFlow fixed seed value, hyperparameters, such as the optimal training batch of the RF-LSTM model, were determined, and the hyperparameter combinations were set as [time_ stept=1, unit_ dim=4, batch_ size=1, learning_ rate=0.001,

epochs=200]. The calculation results showed that the prediction model without the attention mechanism had an average absolute error of 27.95 and an average absolute percentage error of 10.88 in the eight-year return prediction. The results are shown in Table 2. The two error evaluation indices were greater than those for the prediction model integrated with the attention mechanism possibly because the RF-LSTM-Attention prediction model pays more attention to the more important characteristic variables after integrating the attention mechanism, thus improving the generalization performance of the model.

To further compare the prediction performance of the RF-LSTM-Attention prediction model with that of a linear statistical model, the 81 primary mean annular flow field prediction factors in Table 1 were adopted as the factor set, and the stepwise regression method was applied. In addition, 5, 6, and 7 prediction factors were selected from the 81 factors to establish three stepwise regression prediction equations.

$$Y1 = 1.04x_{29} - 0.31x_{68} + 0.42x_{7} + 0.72x_{24} - 0.25x_{81} - 0.86 \tag{17}$$

$$Y2 = 1.08x_{29} - 0.30x_{68} + 0.29x_{7} + 0.63x_{24} - 0.27x_{81} + 0.32x_{23} - 1.05 \tag{18}$$

$$Y3 = 0.95x_{29} - 0.32x_{68} + 0.25x_{7} + 0.57x_{24} - 0.25x_{81} + 0.32x_{23} - 0.21x_{14} - 0.80 \tag{19}$$

The modeling samples of the three stepwise regression forecast equations were also the 53-year samples from 1960 to 2012, and the return forecast test samples were also the eight-year samples from 2013 to 2020. The forecast errors of the historical modeling sample (1960–2012) and eight-year return forecast sample (2013–2020) were calculated using the same error evaluation indicators shown above (Table 3). The inter-annual incremental fitting results of the three multiple regression equations are shown in Figs. 9.

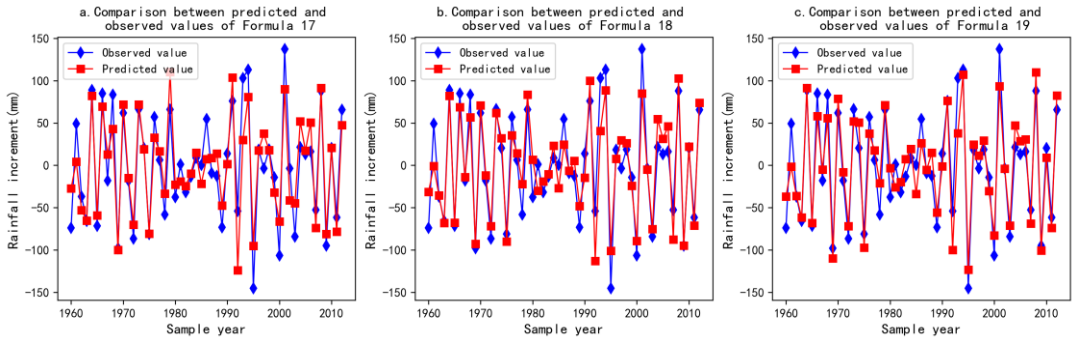

Figure 9. Comparison of prediction results and observed values of inter-annual increment in 53 years (training set) from 1960 to 2012 by using Formula (17,18,19)

Table 3. List of precipitation forecast results and statistics of relevant indicators of three regression equation from 2013 to 2020

Data Forecast year	Precipitation observed value (mm)	Regression formula (17) predicted value of precipitation (mm)	Regression formula (17) prediction error value (mm)	Regression formula (17) relative error percent REP (%)	Regression formula (18) predicted value of precipitation (mm)	Regression formula (18) prediction error value (mm)	Regression formula (18) relative error percent REP (%)	Regression formula (19) model predicted value of precipitation (mm)	Regression formula (19) prediction error value (mm)	Regression formula (19) relative error percent REP (%)
2013	233.1	237.65	4.55	1.95	208.72	-24.38	-10.46	209.84	-23.26	-9.98
2014	268.5	220.42	-48.08	-17.91	232.22	-36.28	-13.51	230.21	-38.29	-14.26
2015	254.2	325	70.8	27.85	315.44	61.24	24.09	327.6	73.4	28.87
2016	227.2	170.23	-56.97	-25.07	166.58	-60.62	-26.68	156.25	-70.95	-31.23
2017	323.2	351.93	28.73	8.89	353.11	29.91	9.25	353.89	30.69	9.5
2018	230.6	276.74	46.14	20.01	289.07	58.47	25.36	284.09	53.49	23.2
2019	269.5	196.01	-73.49	-27.27	185.39	-84.11	-31.21	200.66	-68.84	-25.54
2020	233.7	250.8	17.1	7.32	257.5	23.8	10.18	262.19	28.49	12.19
Mean Absolute Error (mm)		43.23			47.35			48.43		
Mean Absolute Percentage Error (%)		17.03			18.84			19.35		

The calculation results of the three stepwise regression forecast equations showed that the complex correlation coefficients of the three forecast equation modeling samples reached 0.80, 0.844, and 0.87. Figure 9 indicate that the forecast equations had a good fitting effect on the historical samples. The three prediction equations were used to calculate the return prediction error evaluation index for eight years from 2013 to 2020, and their results were compared with the calculation results on the return prediction error index of the same eight years obtained by the RF-LSTM prediction model with and without the attention mechanism proposed in this study (Table 4). The forecast MAE and MAPE of the linear regression method were about twice as large as those of the nonlinear RF-LSTM-Attention and RF-LSTM forecasting models.

Table 4. Comparison of return prediction errors of different prediction methods

Model	Stage	MAE (mm)	MAPE (%)
Regression formula 17	Return forecast	43.23	17.03
Regression formula 18	Return forecast	47.35	18.84
Regression formula 19	Return forecast	48.43	19.35
RF-LSTM	Return forecast	27.95	10.88
RF-LSTM-Attention	Return forecast	23.96	9.49

Further analysis showed that the six important characteristic variables selected from the 81 prediction factors in Table 1 by the random forest method were different from the five, six, and seven variable factors selected by the three stepwise regression methods, except that one characteristic factor (factor X7) was similar. In addition, the six important characteristic variables selected by random forest were not the predictors with the highest correlation coefficient. The random forest algorithm selected the factors according to the importance score of the characteristic variables, and the variables with the highest importance score were not the factor variables with the highest correlation coefficients. Once the feature variable with the highest score (the largest VIM value) calculated by the random forest algorithm was determined, the VIM importance score of the subsequent feature variable in the ranking decreased rapidly, as shown in Figure 10, possibly because when the random forest method selected the first important feature variable, it effectively reduced the same influence of the other variables, which was different from the method that uses the highest linear correlation coefficient as the important selection criterion. This result indicates that the random forest algorithm can reflect the nonlinear correlation between multi-factors and forecast objects through the importance ranking of feature variable scores, thereby improving the forecast performance of the forecast model effectively.

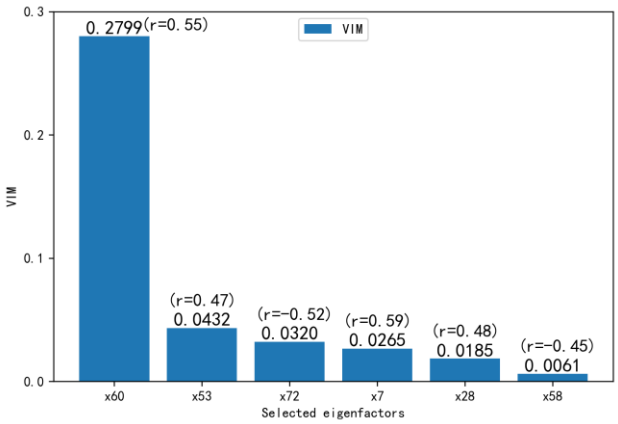

Figure 10 Prediction factor of the top six variables in random forest importance ranking

To examine further the practicability of the RF-LSTM-Attention model, after obtaining the six average circulation field predictors in 2021, the established RF-LSTM-Attention model was used to calculate the actual forecast of the average precipitation in Guangxi from June to August in the summer of 2021. At the same time, an actual forecast test was performed with the four other methods in 2021, and the comparison results are shown in Table 5. The proposed RF-LSTM-Attention model had the best prediction accuracy in the actual prediction test in 2021, and the predicted value was close to the actual value. However, a large error still existed between the value predicted by the three regression forecast methods and the actual value.

Table 5. Comparison of actual forecast precipitation and error results in 2021

Actual forecast model	Actual value of precipitation in 2021 (mm)	Model precipitation forecast value (mm)	Model prediction error value (mm)	Model relative error percentage REP (%)
RF-LSTM-Attention	172.4	177.83	5.43	3.15
RF-LSTM	172.4	182.04	9.64	5.59
Regression formula 17	172.4	253.93	81.53	47.29
Regression formula 18	172.4	269.75	97.35	56.47
Regression formula 19	172.4	253.62	81.22	47.11

CONCLUSION AND DISCUSSION

In this study, the average precipitation increment sequence in summer (June–August) in Guangxi was used as the forecast object, and a new short-term climate prediction model of summer drought and flood trends was established using the random forest algorithm and long-term memory deep learning neural network model integrated with an attention mechanism. The main findings are as follows:

(1) In the theoretical method of prediction modeling, the combination of LSTM deep learning neural network model with the attention mechanism and random forest algorithm can maximize the advantages of the respective algorithms and play different roles in the deep learning of data information features, selection of important characteristic variables, and setting of the attention importance of different variables. Therefore, in experimental research on short-term climate prediction of summer precipitation, prediction accuracy can be considerably improved relative to the corresponding linear regression prediction method, thereby providing a new method for the investigation and application of short-term climate prediction by comprehensively using various machine learning algorithms.

(2) When faced with a large number of high-dimensional data sets of primary average physical quantity factors in the short-term climate prediction statistical forecast modeling of average precipitation, high-dimensional factor data can be randomly extracted many times with playback to construct decision trees for L training subsets consisting of in-bag data and corresponding out-of-bag data. Therefore, the calculation method of the index feature important score of the random forest algorithm can be obtained, and the most important feature variable is objectively and quantitatively selected from the high-dimensional data factor set as the model input. The forecast test results in this study show that this new method, which is different from the linear regression method, is more effective.

(3) After selecting the important input feature variables of the deep learning neural network prediction model, by integrating the attention mechanism layer into the LSTM network model, we can quantitatively and objectively calculate and analyze the importance of the selected characteristic variables in the predictor and give the important characteristic variables a high degree of attention. Thus, the generalization performance of the deep learning neural network prediction model is improved.

(4) In this study, when various intelligent calculation methods were used to forecast and model the average precipitation increment sequence in Guangxi in summer, the previous physical quantity factor of the forecast was mainly the circulation factor of the previous 500 hPa monthly average geopotential height field, and the matching of the spatiotemporal scales of predictors was considered. Many studies have proven that the incremental series of mean precipitation provides good forecast information. However, in the research on the theoretical methods for prediction modeling of short-term climate events via deep learning intelligent computing, because this new prediction model that combines the deep learning network with other intelligent computing methods is essentially a “black box model,” explaining the physical mechanism of the influence of the prediction characteristic variables on the prediction object, especially the comprehensive influence of this multivariable characteristic, is difficult. Although some deep learning forecast modeling studies have explored the interpretability of the deep learning model’s learning and training contents, they were still limited to analyzing the difference in the importance of different variable combinations in forecast results.

ACKNOWLEDGMENTS

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. National Natural Science Foundation of China(42065004); Guangxi Innovation Driven Development Special Project (Major Science and Technology Special Project) (Guike AA21077018, Guike AA21077018-2), Guangxi Normal University of Science and Technology College Student Research Fund Project (GXKS2022DXSA009).Key Laboratory of Guangxi Normal University of Science and Technology (GXKSKYPT2024006).We acknowledge the meteorological information center of Guangxi Zhuang Autonomous Region for its support of this research data. We sincerely appreciate the anonymous reviewers' helpful comments and the editor's efforts in improving this manuscript.

REFERENCES

[1] Huang R, Zhang Q, Ruan S. Prediction and early warning of meteorological disasters and scientific disaster prevention and reduction countermeasures in China. Beijing Meteorological Publishing House, 2005: 94-107 (in Chinese).

[2] McGregorS, GallantA, Rensch VP. Quantifying ENSOs Impact on Australia's Regional Monthly Rainfall Risk. Geophysical Research Letters,2024,51(6).

[3] Du L, Ke Z, Liu C et al. Prediction model of summer precipitation in China based on cluster zoning. Meteorology,2016,42 (01): 89-96. (in Chinese) https: //kns. cnki. net/kcms/detail/detail. aspx? FileName =QXXX201601011&DbName= CJFQ2016

[4] Qin ZN, Jin L, He H et al. Chebyshev polynomial diagnosis and prediction method of April rainfall anomaly in Guangxi. Plateau Meteorology, 2006, (06):1184-1189 (in Chinese). https: //kns. cnki. net/kcms/detail/detail.aspx?FileName=GYQX200606027&DbName=CJFQ2006

[5]F Wei, Yu X, Mann ME. Probabilistic trend of anomalous summer rainfall in Beijing Role of interdecadal variability. Journal of Geophysical Research Atmospheres,2008, 113(D20):1-9. https: //doi. org/10. 1029/2008JD010111

[6] Liu K, Pan J, Zhang R. Research progress of summer precipitation prediction methods in China. People's Yellow River, 2007, 40(01): 18-22(in Chinese). https: //kns. cnki. net/kcms/detail/detail. aspx?FileName= RMHH201801006&DbName=CJFQ2018

[7] He H, Jin L, Qin Z et al. Application of dynamic extension forecast products in monthly precipitation forecast in Guangxi. Journal of Applied Meteorology, 2007, (05): 727-731. https: //kns. cnki. net/kcms/ detail/detail.aspx? FileName= YYQX200705020&DbName=CJFQ2007

[8] Liu Y, Ren H, Zhang P et al. Improve the prediction of summer precipitation in South China by a new approach with the Tibetan Plateau snow and the applicable experiment in 2014. Chinese Journal of Atmospheric Sciences,2017,41(2):313-320(inChinese). https://doi.org/10.3878/j

[9] Wei F, Huang J. A study of downscaling factors of atmospheric circulations in the prediction model of summer precipitation in Eastern China. Chinese Journal of Atmospheric Sciences, 2010, 34(1): 202-212(in Chinese). https://kns.cnki.net/kcms/detail/detail.aspx?FileName=DQXK201001019&DbName=CJFQ2010

[10] Liu Y, Wang L, Zhao Z et al. Review of China's summer precipitation forecast in 2006. Progress in Climate Change Research, 2007, (04): 243-245. https: //kns. cnki. net/kcms/detail/detail. aspx? FileName= QHBH200704014&DbName=CJFQ2007

[11] Fan K, Wang H, Choi Y. A physical statistical prediction model of summer precipitation in the middle and lower reaches of the Yangtze River. Science Bulletin, 2007,52(24): 2900-2905(in Chinese). https://kns.cnki.net/kcms/detail/detail.aspx?FileName=KXTB200724015&DbName=CJFQ2007

[12] Zheng R, Liu J, Ma Z. Application of interannual increment method in summer precipitation prediction in Southwest China. Journal of Meteorology,2019,77(3):489-496(in Chinese). https: //doi. org/10. 11676/ qxxb2019.027

[13] Deng SG, Hu YJ et al. Prediction of Precipitation in the Southern Qinghai-Xizang Plateau During the Flood Period Based on the Interannual Increment Approach. Plateau Meteorology, 2021,40(4):737-746(in Chinese). https://doi.org/10. 7522/j

[14] Wang X, Tian B, Fan K. Climate prediction at the beginning of Pu'er rainy season in Yunnan Province based on interannual increment method. Journal of Atmospheric Science, 2019,42 (06): 801-813. https://doi.org/10.13878/j.cnki.dqkxxb.20181230001

[15] Dai Y, He N, Fu Z et al. Beijing Intelligent Grid Temperature Objective Prediction Method (BJTM) and Verification of Forecast Results. Journal of Arid Meteorology, 2019,37(2):339-344(in Chinese). https://doi.org/10.11755 /j

[16] AK Sharma, S Chaurasia, DK Srivastava. Supervised Rainfall Learning Model Using Machine Learning Algorithms. International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham,2018, 275-283. https://doi.org/10.1007/978-3-319-74690-6_27

[17] Cao Y, Wang C, Wang X et al. Urban road short-term traffic flow prediction based on spatiotemporal node selection and deep learning. Computer Applications, 2020,40 (05) (in Chinese): 1488-1493(in Chinese). https://doi.org/10. 11772/j

[18] Fan JX, Li Q, Zhu Y et al. Research on spatiotemporal prediction model of air pollution based on RNN. Surveying and Mapping Science,2017,42 (07): 76-83+120(in Chinese). https://doi.org/10.16251/j.cnki.1009-2307.2017.07.013

[19] BT Ong, K Sugiura, K Zettsu. Dynamic pre-training of deep recurrent neural networks for predicting environmental monitoring data. IEEE International Conference on Big Data (Big Data), 2015, pp. 760-765. https://doi.org/10.1109/BigData.2015.7004302

[20] Jin L et al. Modeling theory, method and application of neural network weather forecast. Beijing Meteorological Publishing House 2004:168-180 (in Chinese).

[21] Lu H, Ou Y, Qin C et al. A fuzzy neural network bagging ensemble forecasting model for 72-h forecast of low-temperature chilling injury. Natural Hazards (Springer), 2020, (6):1-14. https: //doi. org/10. 1007/s11069-020-04393-y

[22] Huang Y, Jin L, Lu H et al. A Combined Qualitative and Quantitative Prediction Scheme for Cold-Wet Extreme Weather in Guangxi Based on Intelligent Computing. Chinese Journal of Atmospheric Sciences, 2019,43(6): 1424-1440(in Chinese). https://doi.org/10.3878/j

[23] Xiong Y, Zhou J, Jia B et al. Monthly runoff prediction based on random forest teleconnection factor selection. Journal of Hydropower,2021,1-14(in Chinese). http: //kns. cnki. net/kcms/detail/11.2241.TV.20211116.1426.004.html

[24] Li LJ, Wang YT, Hu QF et al. Long-term inflow forecast of reservoir based on random forest and support vector machine. Hydro-Science and Engineering,2020, (4): 33-40 (in Chinese). https: //doi. org/10. 12170/20190626001

[25] Peng YQ, Qiao Y et al. PM2.5 prediction model fuses attention mechanism. Sensors and Microsystems,2020,39(07):44-47(in Chinese). https://doi.org/10.13873/J.1000-9787(2020)07-0044-04

[26] Liu JC, Qin XL, Zhu R. Location prediction of RFID moving objects based on LSTM attention. Computer Science,2021, 48 (03): 188-195(in Chinese). https://doi.org/10.11896/jsjkx.200600134. ISSN: 1002-137X.

[27] Fu Y, Cao W, Zhang J. High Arsenic Risk Distribution Prediction of Groundwater in the Hetao Basin by Random Forest Modeling. Rock and Mineral Testing, 2021,40 (06): 860-870(in Chinese). https://doi.org/10.15898/j.cnki.11-2131/td.202108170099

[28] Zhi Y, Jin Z, Lu L et al. Improving atmospheric corrosion prediction through key environmental factor identification by random forest-based model. Corrosion Science, 2021,178:109084. https: //doi. org/10.1016/j.corsci.2020.109084.

[29] Ghosh D, Cabrera J. Enriched Random Forest for High Dimensional Genomic Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, pp 2035-2044. https: //www. researchgate.net/publication/339292359_Enriched_Random_Forest_for_High_Dimensional_Genomic_Data

[30] Li MY, Kong F. Combined self-attention mechanism for named entity recognition in social media. Journal of Tsinghua University (NATURAL SCIENCE EDITION), 2019, 59(06): 461-467(in Chinese). https: //doi. org/10.16511/j.cnki.qhdxxb.2019.25.005