An Integrated Artificial Neural Network-based Precipitation Revision Model

Tao Li[1], Wenduo Xu[1], Li Na Wang[1], Ningpeng Li[1], Yongjun Ren[2*] and Jinyue Xia[3]

[1] School of artificial intelligence, Nanjing University of Information Science & Technology
Nanjing, China
[e-mail: lthnxx@21cn.com, 20201220069@nuist.edu.cn, wangln@nuist.edu.cn, 20201249538@nuist.edu.cn]
[2] School of Computer and Software, Nanjing University of Information Science & Technology
Nanjing, China
[e-mail: renyj100@126.com]
[3] International Business Machines Corporation (IBM), NY, USA
[e-mail: junyue.xia@ibm.com]
*Corresponding author: Yonjun Ren

Abstract

Precipitation prediction during flood season has been a key task of climate prediction for a long time. This type of prediction is linked with the national economy and people's livelihood, and is also one of the difficult problems in climatology. At present, there are some precipitation forecast models for the flood season, but there are also some deviations from these models, which makes it difficult to forecast accurately. In this paper, based on the measured precipitation data from the flood season from 1993 to 2019 and the precipitation return data of CWRF, ANN cycle modeling and a weighted integration method is used to correct the CWRF used in today's operational systems. The MAE and TCC of the precipitation forecast in the flood season are used to check the prediction performance of the proposed algorithm model. The results demonstrate a good correction effect for the proposed algorithm. In particular, the MAE error of the new algorithm is reduced by about 50%, while the time correlation TCC is improved by about 40%. Therefore, both the generalization of the correction results and the prediction performance are improved.

---

1. Introduction

$\mathbf{T}$he goal of climate prediction is to infer possible trends in climate development over a certain future period based on how the climate has evolved in the past. In recent years, the potential role for climate prediction in disaster prevention and mitigation has been increasingly recognized, and the demand for such prediction is growing in a range of fields, along with the need for socio-economic development. Improvements in the study of climate prediction are accordingly urgent [1-3]. Global climate models (GCMs) are the most useful tools for studying precipitation assessment and predicting future precipitation trends, as they reflect the characteristics and distribution of regional precipitation. However, a gap still exists between the simulation results and the observed precipitation. Compared with GCMs, RCMs have higher accuracy in terms of simulating precipitation [4]; however, due to the level of boundary conditions on the side of GCMs, it is necessary to apply objective revisions of RCM predictions in order to develop region-specific interpretation application products and thereby improve forecasting accuracy [5-7].

Precipitation revision is an effective method of improving model prediction. The concept of precipitation revision was first proposed because some model data for predicting precipitation in flood seasons already exists in current operations; however, these data are also impacted by certain deviations. It is therefore hoped that revision can decrease the error and improve the accuracy and performance of precipitation forecasting [8-10]. Through the development of meteorological studies, artificial intelligence and data mining research, the use of intelligent computing and data mining techniques for precipitation revision in regional areas has come to provide new and effective methods for improving the quality and prediction accuracy of existing precipitation forecasts, and has therefore become a hot research topic. Zhou Lin et al. proposed applying the probabilistic adjustment method to revise the regional climate model system in order to simulate the national daily precipitation across all seasons, thereby significantly improving the simulation of the mean and extreme precipitation in China [11]. Li Chunhui et al. used the concept of multiscale spatiotemporal projection (MSTP) to establish a prediction method for monthly and seasonal precipitation in the Guangdong region. Periodic decomposition was performed via EOF decomposition, wavelet analysis and the Lanczos filtering method, while prediction was performed using the MSTP method [12]. Wu Qishu et al. used the optimal TS score revision method (OTS) and the optimal ETS score revision method (OETS) to determine the forecast day precipitation revision coefficients at all levels [13]. Lu Xinyu et al. used the 1998–2013 TRMM monthly precipitation products and ground-observed precipitation from 105 meteorological stations in Xinjiang during the same period; these authors further applied stepwise regression and BP neural network methods to select the 1998–2010 data necessary for building a revised precipitation model in Xinjiang, which was tested using the 2011–2013 monthly precipitation. Based on the stepwise regression model and BP, the revised TRMM precipitation products of the neural network model are able to accurately and quantitatively reproduce the precipitation distribution, thereby providing a more practical reference method for improving the quality of TRMM precipitation products [14]. Lu Yao used ensemble systems for precipitation prediction; these systems can take the complex mechanisms in the precipitation prediction process into account and are thus increasingly used in current practical forecasting operations [15]. Kamal et al. used machine learning algorithms to achieve integrated multi-model prediction of precipitation and temperature [16]. Chen et al. predicted extreme precipitation using the K-means clustering algorithm [17]. Gregory et al. applied three different statistical algorithms to predict localized extreme precipitation throughout the Continuous United States (CONUS), using the Random

Forest (RF) training model for precipitation prediction [18]. Chen Jinpeng et al. proposed the time-by-time precipitation forecast revision method based on Convolutional Neural Network. and it has a significant effect on reducing the underreporting rate of classified precipitation forecasts and the nulling rate of clear-rain and weak precipitation forecasts [19]. Dong Xiaoyun et al. used daily rainfall data simulated by CWRF model and observed at 2416 meteorological stations in China from June to August during 1980 －2015 are used to compare correcting effects of Q-lin, Q-tri, RQ -lin, RQ- tri, SSP-lin and CDFt on extreme precipitation of control scheme simulated by CWRF in eastern China [20].

In general, the current forecasting of climate precipitation is based primarily on meteorological models. As artificial intelligence technology has developed, new related studies have emerged that combine these two methods to improve prediction accuracy.

In this paper, machine learning is combined with precipitation in flood season revision to investigate new precipitation in flood season revision algorithms that can potentially improve the core operational capabilities in flood precipitation forecasting. This research is conducted against the background of the revision of precipitation in flood season forecast data obtained from the National Climate Center of China Meteorological Administration (CMA), which is generally performs poorly in terms of flood precipitation forecast performance. This paper presents a precipitation correction model integrated with artificial neural networks, which are modeled via ANN cycle and then integrated by means of weighted accumulation. The validity of the algorithm is experimentally verified, after which the algorithm is applied to the related business of climate forecasting.

The remainder of this paper is structured as follows. The second part of the work explains ANN in detail, along with the theory and basic concept of integrated learning. The third part introduces the organization of the integrated ANN algorithm, while the fourth part outlines the experimental data, methods and results. Finally, in the fifth part, the conclusions drawn from the experimental results are presented.

2. Related algorithmic models

2.1 Artificial Neural Networks

Artificial Neural Network (ANN) systems are connected by a large number of neurons with adjustable connection weights. These systems are characterized by massively parallel processing, distributed information storage, and good self-organizing and self-learning capabilities [21]. An artificial neural network is a model that simulates the biological nervous system and that can theoretically estimate arbitrary nonlinear functions. Artificial neural networks can be used for long-term precipitation forecasting without needing to know the mechanism of precipitation [22]. BP neural networks are used a lot now. BP (Back Propagation) algorithm, also known as the error back-propagation algorithm, is a supervised learning algorithm in artificial neural networks. The basic structure consists of nonlinearly varying units with a strong nonlinear mapping capability. The principle behind this algorithm is that the network sends a series of inputs to the implicit layer through connection weighting, after which the neurons in the implicit layer aggregate all inputs to produce a nonlinear output through the activation function, which is then sent to the output layer through the next connection weighting. The neurons in the output layer subsequently sum up all the inputs and produce an output in response.

2.1.1 Neuron model

Neurons are individual units that are connected to each other, each of which has numerical inputs and outputs that can take the form of real numbers or linear combinatorial functions. First, the network structure has to learn with a guideline before it can work properly. When the network makes a judgment error, it is made possible by learning to avoid the same error. This method has strong generalization ability and nonlinear mapping ability, and can model the system using minimal information. Neural networks can be used for classification, clustering, prediction and so on. Neural networks also need to have a certain amount of historical data; through the training of this historical data, the network can learn the hidden knowledge contained in the data. The first step is to find the features of some problems, along with the consistent relevant data, and use these data to train the neural network.

The artificial neuron model can be expressed as follows:

$$net_i = \sum_{j=1}^{n} w_{ij} x_j - \theta \tag{1}$$

$$y_i = f(net_i) \tag{2}$$

where $X = [x_0, x_1, x_2, ..., x_n]$, $W = \begin{bmatrix} w_{i0} \\ w_{i1} \\ . \\ w_{in} \end{bmatrix}$, $net_i = XW$, $y_i = f(net_i) = f(XW)$ $\tag{3}$
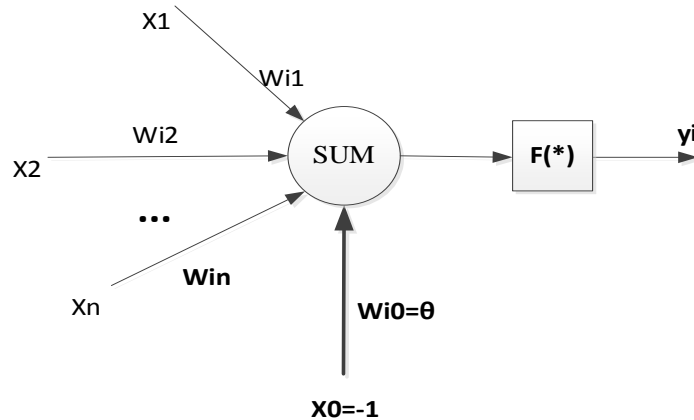

Fig. 1. Neuron model

In **Fig. 1**, $X_1 \sim X_n$ denote the input signals from other neurons, $W_{i1} \sim W_{in}$ are the weights of the incoming signals, and θ represents a bias, which is set to achieve an accurate output and is an important model parameter. The input signals integrated by the neuron and the bias are summed up to produce the final processing signal of the current neuron net. The activation signal serves as the input to the function $f(*)$ in the right half of the circle in the figure above, $f(net)$. f is also referred to as the activation function or excitation function; the activation function is mainly added to the non-linear function in order to avoid problems with the expression and classification ability of linear model. In the above figure, y is the output of the current neuron.

2.1.2 Neural Network Model

Neural networks are composed of a large number of neurons that are interconnected with each other. The output of the network varies depending on the connection method, weight value and excitation function of the network.

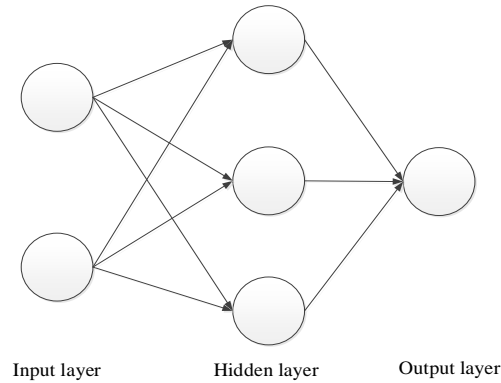The neural network model can be expressed as follows:

Fig. 2. Artificial neural network structure

In **Fig. 2**, for the input layer, each neuron corresponds to a variable feature, while the neuron is equivalent to a container that contains numbers; for the output layer, the regression problem is one neuron, while the classification problems are multiple neurons; for the parameters, all parameters in the network are the weights and biases of the hidden layer neurons.

2.1.3 BP neural network and its improvement

The hidden layer units in the BP network are located between the input and output layers. These units have no direct connection with the outside world, but changes in their state can affect the relationship between the input and output, and each layer can have several nodes.
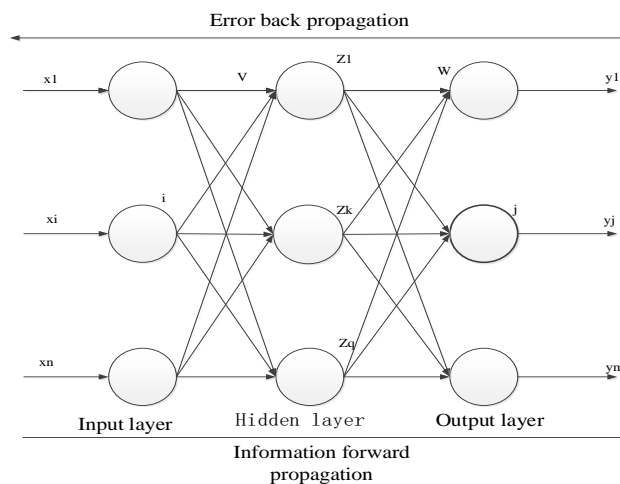

Fig. 3. Three-layer BP neural network

The basic BP algorithm consists of two processes: forward propagation of the signal and backward propagation of the error. The calculation error output direction moves from input to output, while the adjustment weights and bias direction moves from output to input. In forward propagation, the input signal acts on the output node through the hidden layer and produces the output signal through nonlinear transformation. If the actual output is not consistent with the expected output, it turns into an error back-propagation process. The error back-propagation is the layer-by-layer back-propagation of the output error through the implied layer to the input layer; here, the error is assigned to all cells in each layer to adjust the weights of each cell with the error signal obtained from each layer. By adjusting the connection strength between the input node and the hidden layer node, as well as the connection strength and bias between the hidden layer node and the output node, the error is decreased along the gradient direction. Following repeated learning and training, the network parameters corresponding to the minimum error (weights and bias) are determined, and the training will stop. The specific process of the algorithm is as follows:
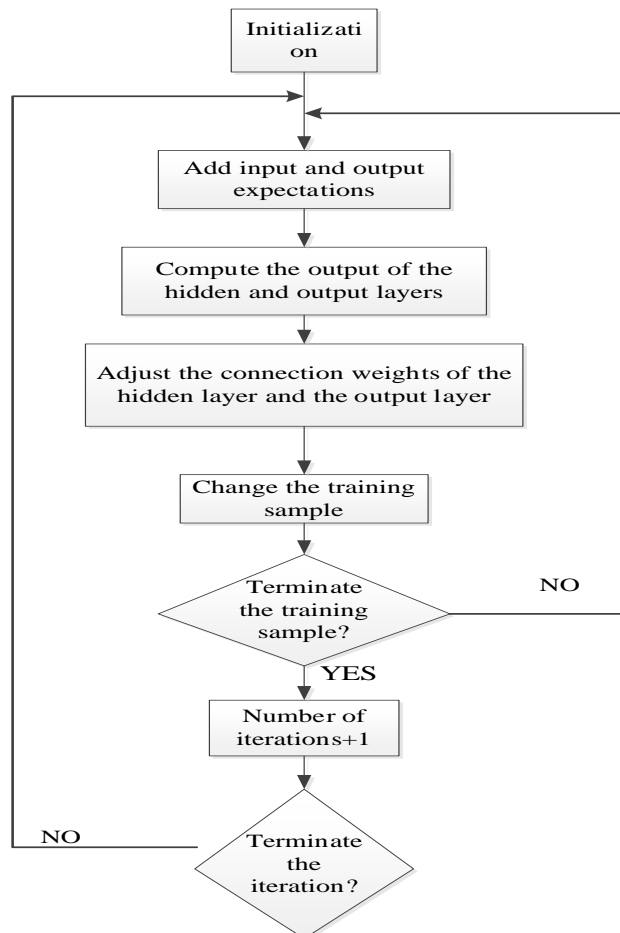

Fig. 4. BP neural network training flow chart

In BP neural network training, the process mainly uses the original dataset for the training of the BP neural network; this eventually forms a specific neural network with a prediction function, and saves the neural network after the training is completed. When considering the

improvement of the BP neural network, the main parameters affecting the performance of BP neural networks are as follows: the number of nodes in the hidden layer, the choice of activation function and the choice of learning rate. The fewer neurons there are in the hidden layer, the worse the BP neural network simulation; moreover, a higher number of neurons in the hidden layer results in a better simulation, although it also makes the training slower. In this paper, the number of hidden layer nodes is improved based on the neural network according to the following empirical formula:

$$\sqrt{N} + X \tag{4}$$

Here, N represents the number of sample features, while the range of values is between one and ten. We first determine the number of hidden layer nodes according to the step-by-step experiment method. This involves setting an initial value, then gradually increasing the number on the basis of this value, comparing the prediction performance of each network, and selecting the number of nodes with the best performance as the number of hidden layer neuron nodes. The following conditions must be met when determining the number of hidden layer nodes: first, the number of nodes in the hidden layer must be less than N-1 (where N is the number of training samples), and if not, the systematic error of the network model is deemed independent of the characteristics of the training samples and tends towards zero, meaning that the established network model has no generalization ability and no practical value. Second, the number of training samples must exceed the number of connection weights in the network model. otherwise, the samples must be divided into several parts and use the "Rotation training" method to obtain a reliable neural network model.

In the BP algorithm, the weights and bias are adjusted once for each training iteration. If the number of nodes in the hidden layer is too small, the network may not be able to train, or the network performance will be poor. Moreover, if the number of nodes in the hidden layer is too large, the systematic error of the network can be reduced; however, on the one hand, the training time of the networks is prolonged, and on the other hand, it is easy for the training to fall into local minima and not reach the optimal point, which is also the inherent reason for "overfitting" during training.

Therefore, an equitable number of hidden layer nodes should be determined by node deletion and expansion methods, with comprehensive consideration given to the complexity of the network structure and error size.

2.2 Bagging

In the shallow regression model, overfitting is the key factor that determines the quality of the model, while the integration model can effectively avoid overfitting by combining shallow regression models. Bagging method is a typical ensemble learning method.

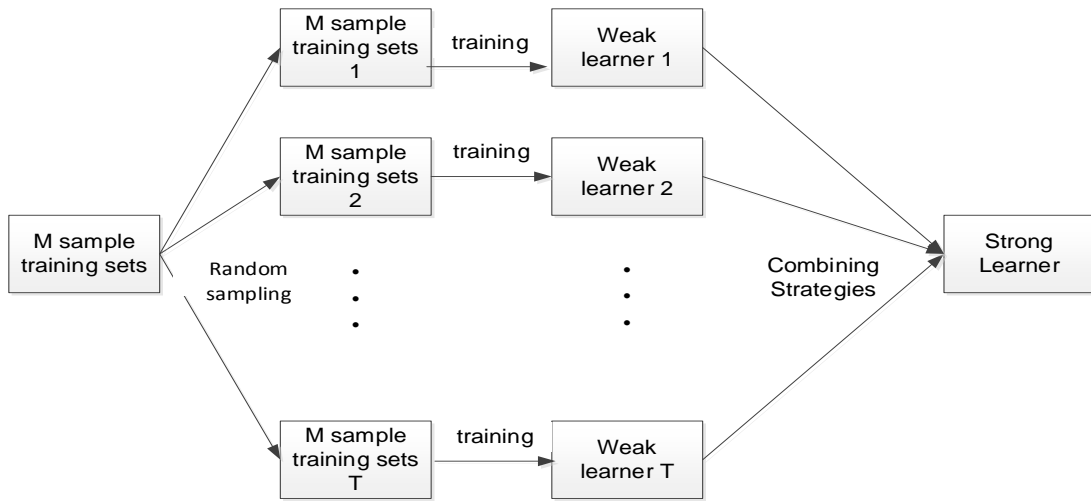The schematic diagram of Bagging is as follows:

Fig. 5. The schematic diagram of Bagging

In **Fig. 5**, the Bagging input is as follows: sample set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, number of weak learners, iteration times T; output: final strong learner $f(x)$.

Bagging features "random sampling". In this approach, M samples are randomly selected and collected T times to obtain the sampling set. For a sample, in a random sampling of m samples, the probability of each sample being collected is $\dfrac{1}{m}$

The probability of not being collected in m samples is as follows:

$$P(\text{Not a single one has been collected}) = (1 - \frac{1}{m})^m \tag{5}$$

Taking the maximum value for m yields:

$$\lim_{m \to \infty}(1 - \frac{1}{m})^m \to \frac{1}{e} \approx 0.368 \tag{6}$$

This means that about 36.8% of the training set is not collected in each random sampling round of Bagging. This 36.8% of data that were not sampled are referred to as "out-of-bag data". While these data do not participate in the fitting of the training set model, they could be used as a test dataset to test the model's generalization ability.

3. Revised precipitation model with integrated artificial neural network

3.1 Basic concept of the algorithm

The basic concept behind the algorithm proposed in this paper involves using CWRF model return and actual observation data as model training data. According to the similarity of climates in neighboring regions and interdecadal influence, precipitation-related meteorological elements are selected for data organization as the input data of the algorithm model. This paper use ANN as the base model for it can be used for long-term precipitation forecasting without needing to know the mechanism of precipitation. Besides, due to the shallow regression model's tendency towards overfitting and similarities in climate over two to six years due to the decadal influence, a single shallow regression model cannot fully use

the data. So, the ANN algorithm model that this paper use is built circularly, and the prediction results of different algorithm models are weighted and integrated. In this way, a better ensemble model is obtained than that of the single shallow neural network.

3.2 ANN Cycle Modeling

Due to the shallow regression model's tendency towards overfitting, as well as similarities in climate over a period of two to six years due to the decadal influence, a single shallow regression model cannot make full use of the data.

In this paper, ANN cycle modeling is selected, and the specific process of modeling is N-year modeling: the first N years of recorded data organization are the input, while N+1 years are the output. The algorithm model under different algorithm parameters is used to predict precipitation during the recorded years. For the second year, N+1 years of recorded data organization are the input, and N+2 years are the output, using the algorithm model under different algorithmic parameters to predict precipitation during the recorded years. Following this rule, each model will produce a prediction result for the flood season of the recorded year under different algorithm parameters, and multiple prediction results are obtained for the flood season of the recorded year under different algorithm parameters. The prediction results of the multiple algorithm models are then integrated to obtain the final precipitation correction results.

Considering the interdecadal influence, the climate exhibits similarity over the last two to six years [23]. This paper uses data from the last 2-6 of the year to be predicted to conduct data organization modeling prediction experiments; the results show that the revised results over the last three years are better than the previous years.

3.3 Weighted integration

As a result of the above, through cyclic modeling, each forecasting model forecasts the flood season for the recorded years under different algorithmic parameters, so that there are multiple different forecasting results for the flood season for the recorded years under different algorithmic parameters. In the shallow regression model, overfitting is the key factor that determines the quality of the model, while the integration model can effectively avoid overfitting by combining shallow regression models. Therefore, this paper opts to use the integration method. There are two main types of integration methods, namely averaging and voting approaches; since this paper studies the numerical class output, the averaging method is used. In this paper, the prediction performance of recorded years is considered, and weighted average similarity is adopted to assess the similarity between the CWRF model returns of recorded years. The formula is as follows:

$$H(x) = \sum_{i=1}^{T} w_i h_i(x) \tag{7}$$

Here, w_i is the weight of the individual learner h_i.

The idea of weighted average similarity in this paper can be expressed as follows. In order to obtain the flood forecasts in a recorded year, the similarity between the CWRF model returns in the years of record and those in the years of the output of different algorithm models is calculated, after which the weights are assigned according to the similarity between the two. Finally, the flood forecasts in the years of record for each algorithm model are accumulated according to the weights, and the flood forecasts in the years of record calculated by the integrated method are obtained.

3.4 Algorithm description

The specific algorithm can be described as follows:

Algorithm 1 ANN Cycle Modeling And Weighted Integration

Input: D: Set of CWRF historical precipitation return data and actual observation data;

         X: tuple of regression prediction ;

         N: Number of years used to model;

         K：Number of models;

         L: Learning Rate;

Output: Prediction value;

1: **for** $i = 1; i < K; i + +$ **do**   ▷ Model creation with cyclic modeling of CWRF precipitation return data and measured data from 1993 to 2019

2:      Organize the CWRF precipitation return data for years i+1992 to 1992+i+N. Take the departure as the model input.

3:      Set weights and thresholds for initializing the neural network

4:      Conduct data normalization so that the training samples obey a 0 mean, 1 variance distribution.

5:      Train the neural network to adjust threshold weights and set learning rates

6:      Export Model M_i

7: **end for**

8: Multiple ANN models modeled via loop were used for regression prediction of tuple X:

9: **for** $i = 1; i < K; i + +$ **do**

10:      Calculating w_i by using the similarity between CWRF model return data from years of record

11: **end for**

12: output:$H(x) = \sum_{i=1}^{K} w_i M_i(x)$

---

4. Experimental results and analysis

4.1 Data sources

This paper will utilize the actual observed flood precipitation data from ground stations for the years 1993–2019 and the regional climate model CWRF historical return flood precipitation data to train the algorithm model. The precipitation return data were obtained from the 1993-2019 flood precipitation return data provided by the CWRF, a regional climate model of the National Climate Center (NCC) of the China Meteorological Administration.

The data parameters are presented in **Tables 1** and **2**.

Table 1. Actual observation data parameters

Parameter	explain
time	year:1993-2019; month:6-8
Longitude range	58.3979E-161.6021E
Latitude range	8.3653N-58.7461N
precipitation	Mainly around 0-16mm/h
Number of checkpoints	231*171

Table 2. CWRF model data parameters

Parameter	explain
time	year:1993-2019; month:6-8
Longitude range	58.3979E-161.6021E
Latitude range	8.3653N-58.7461N
precipitation	unit: mm/h, scope:0-217.9202, Mainly around 0-16mm/h
model type	pattern 01
Number of checkpoints	231*171

4.2 Data preprocessing and organization

In the first step, the precipitation data for the flood season are preprocessed. In order to realize the randomness of the precipitation forecast and avoid the chronological mean precipitation value influencing the forecast, departure is commonly used in meteorological forecasting; thus, this paper also chooses departure. In this paper, the forecast rainfall is compared with the average rainfall over many years, while the predicted rainfall is the value minus the average rainfall of the same period. This approach is generally used for medium- and long-term forecasting and can be used as a reference for flood and drought control:

$$Dpt = Pre - AvgPre \qquad (8)$$

here, Pre represents the precipitation at a certain place in a certain year, month and month, while $AvgPre$ denotes the average of the precipitation at a given location over the years recorded to that point. Dpt represents the precipitation averaged over distance. If $Dpt > 0$, then this is a positive anomaly, and annual precipitation is greater than the cumulative average precipitation. If $Dpt < 0$, then this is a negative anomaly, and the cumulative average precipitation is greater than the annual precipitation.

Next, it is necessary to organize a data format that is suitable for ANN algorithm model training. We will first consider the way in which the data is organized, First, for data selection, two major types of data are used in this paper: CWRF model return data and actual observation data. Moreover, this paper adopts two methods to establish the data revised model: the first is to use only the CWRF mode return data as the input of the algorithm model, while the actual observation data is the output; the second is to use a combination of CWRF mode data and actual observation data as input and the actual observation data as output. We analyze and compare the revised performance of predicted precipitation for two types of input modeling, then conduct feature extraction. First, based on the similarity of the neighboring regional climate, this paper divides the area around the point into a small area of size $M * M$, such that each grid point has $M * M$ feature data. Second, based on the interdecadal effect in the meteorological domain, this paper uses close to N years of data as similar features for each grid point as the modeling input, while the output data is the latter year data. Finally, this paper preprocesses and organizes the data by utilizing the above data organization methods.

This paper then revises the precipitation data in flood season forecast results based on 2132 grid points in the JAC region (YHRB). The above-mentioned data organization methods are used to construct data. In addition, for the features, each grid point in this paper only has CWRF data reported along with actual observed data in a specific region at a specific time.

this paper uses close to N years of data as similar features for each grid point as modeling input, while the output data is the latter year data. If only CWRF model return data is used as input, there are $N*9$ input signals for each grid point; moreover, if the combination of CWRF model data and actual observation data is used as input, there are $N*9*2+9$ input signals for each grid point, containing both CWRF model data and actual observations for the previous N years and CWRF model data for (N+1) years.

4.3 Model parameter selection

In the BP algorithm, the weights and bias are adjusted once for each training iteration. If the number of nodes in the hidden layer is too small, it may not be possible to train the network, or its performance may be poor. Moreover, if the number of nodes in the hidden layer is too large, the systematic error of the network can be reduced, but this will prolong the network training time. Therefore, the reasonable number of hidden layer nodes should be determined by considering the network structure complexity and the error size. Common methods include the node deletion method and extension method [24]. According to the node deletion and expansion methods, three types of node numbers (15, 25, and 35) are used to compare the effects of different numbers of hidden layer nodes on precipitation forecast data revisions.

4.4 Precipitation prediction evaluation index

In the field of machine learning, the mean absolute error (MAE) is a commonly used metric; in the field of climate prediction, the time correlation coefficient (TCC) is also commonly used. Therefore, in this paper, the prediction performance of precipitation prediction data is also evaluated using two evaluation metrics, namely the mean absolute error (MAE) and the time correlation coefficient (TCC), to evaluate the developed machine learning model.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} | pre_i - obs_i | \tag{9}$$

In this situation, pre_i is the model return or precipitation data model prediction for the sample point i, while obs_i is the actual observed precipitation data for sample point i; moreover, N is the number of sample points in the region that actually participate in the assessment.

We next calculate the distance correlation coefficient using the percentage of precipitation distance, which is expressed by the following formula:

$$TCC = \frac{\sum_{i=1}^{N}(\Delta R_f - \overline{\Delta R_f})(\Delta R_0 - \overline{\Delta R_0})}{\sqrt{\sum_{i=1}^{N}(\Delta R_f - \overline{\Delta R_f})^2 \sum_{i=1}^{N}(\Delta R_0 - \overline{\Delta R_0})^2}} \tag{10}$$

Where ΔR_f and $\overline{\Delta R_f}$ denote the predicted value of precipitation departure (or average temperature departure) and its multi-year average value. Moreover, ΔR_0 and $\overline{\Delta R_0}$ denote the corresponding observed value, while N refers to the total number of grid points that actually participate in the assessment.

4.5 Results and analysis

Through the above-mentioned studies, this paper presents an improved artificial neural network based on the "01" pattern precipitation forecast data produced by the CWRF for the

last three years. In order to realize the revision of operational precipitation forecasting in the national region, this section uses the CWRF model data and actual observed flood season precipitation data from 1993 to 2019 in the national regional operational network, then preprocesses these precipitation data and revises them by using an improved ANN algorithm, and then compares and analyzes the performance of the new algorithm in terms of forecast MAE error and forecast time-dependent TCC. After selecting the regional climate model CWRF simulation model 01 flood precipitation data for revision, our experimental comparison found that the revised results of the flood precipitation modeled by the ANN algorithm for the last three years of data organization are superior to the simulation results of the CWRF model.

4.5.1 MAE comparison of revised results

There are seven types of fold lines on the MAE chart, as follows: MAE changes of model results case and actual observed value (obs); changes in MAE between the revised results modeled using only the model results case and the actual observed value (obs) under the corresponding parameters; the change in MAE between the corrected results of the model built with both types of data and the actual observed value (obs).

The "01" pattern experimental results are as follows (MAE of the integrated artificial neural network national regional projections for 1996–2019):
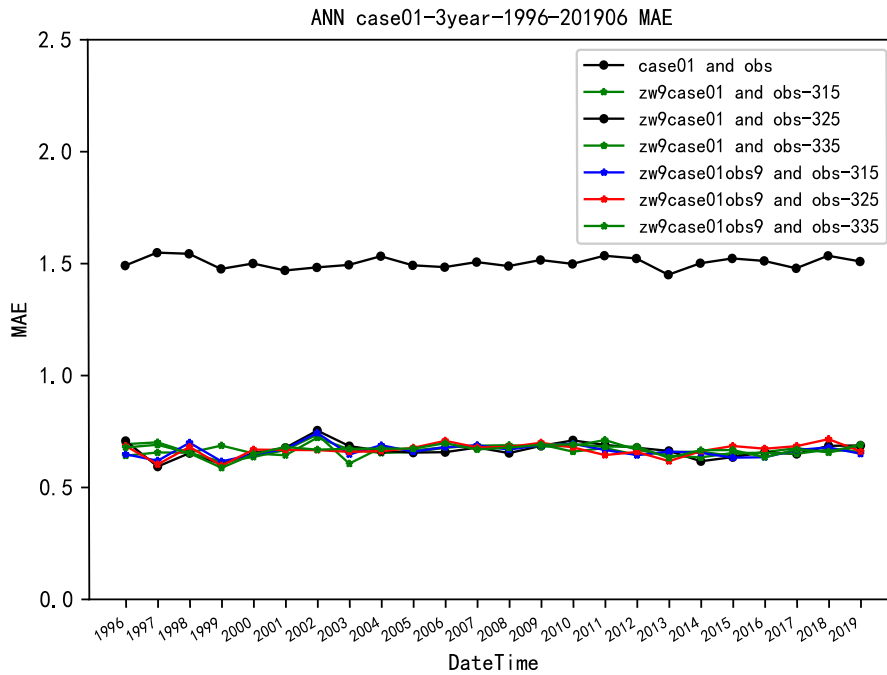

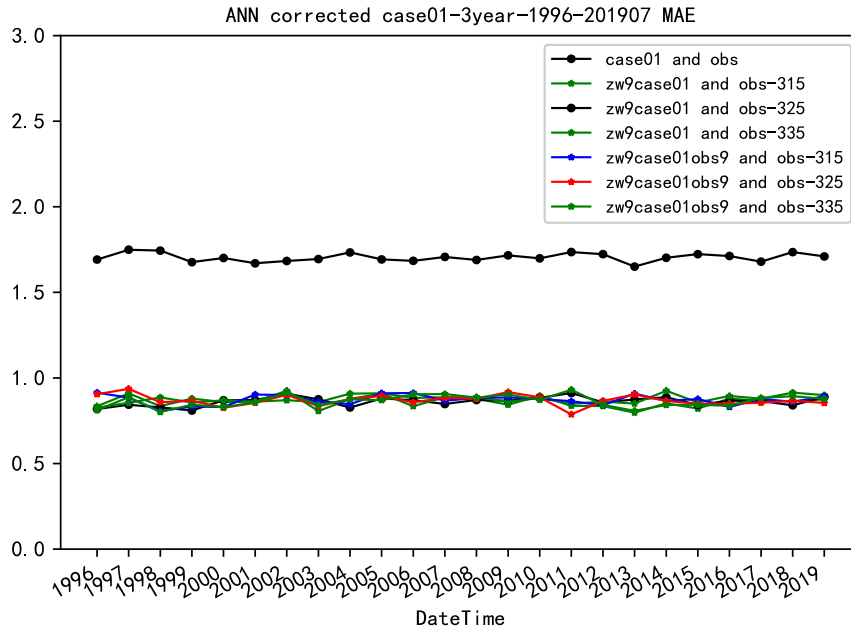**Fig. 6.** MAE of the improved ANN for June forecast data revision results

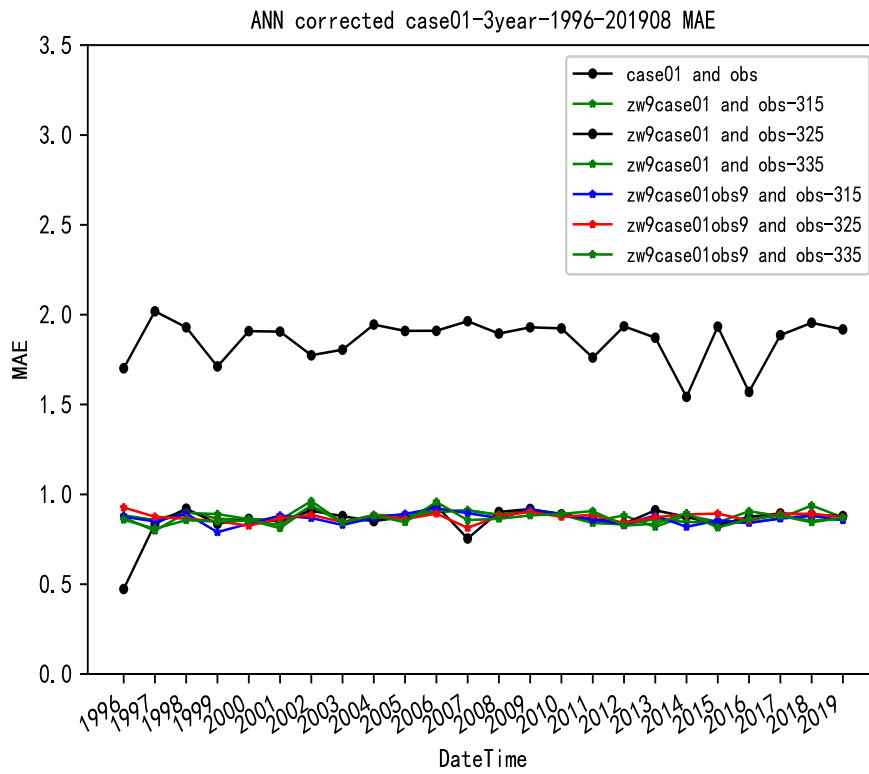**Fig. 7.** MAE of the improved ANN for July forecast data revision results

Fig. 8. MAE of the improved ANN for August forecast data revision results

The following can be seen from **Fig. 6, Fig. 7** and **Fig. 8**:

First of all, the MAE range for CWRF mode is 1.4-2mm/h, while the MAE range for the revised results based on machine learning is 0.5-1mm/h. The MAE between the two is thus reduced by about 1mm/h. Accordingly, the algorithm model proposed in this paper successfully revises the national regional flood precipitation forecast data, and the algorithm thus exhibits some forecasting performance. Next, the CWRF "01" pattern return data based on ANN did not exhibit a large overall fluctuation in the revised results for the months of June, July and August (the national regional watershed flood season), with levels essentially maintained at around 0.4mm/h; therefore, the ANN revision results in the "01" pattern are more stable. Finally, the ANN algorithm exhibits a small fluctuation between the revised results for different model inputs and different numbers of nodes, which remains within 0.2mm/h. In this paper, we experimented with the three nodes 15 25 35, which have less effect on the results. It indicates that the stability of the neural network is good. Therefore, there is no substantial difference between the revised ANN results for different inputs and different numbers of nodes, and the prediction performance is equivalent.

4.5.2 TCC comparison of revised results

We here present the TCC comparison of revised results from regional forecast data in China. There are three pictures presented in the TCC diagram for comparison: TCC changes of model results case and actual observed value (obs); changes in TCC between the revised results modeled using only the model results case and the actual observed value (obs) under the corresponding parameters; change in TCC between the corrected results of the model built with both types of data and the actual observed value (obs).

The "01" pattern experimental results are as follows: TCC of the integrated artificial neural network national regional projections for 1996–2019.
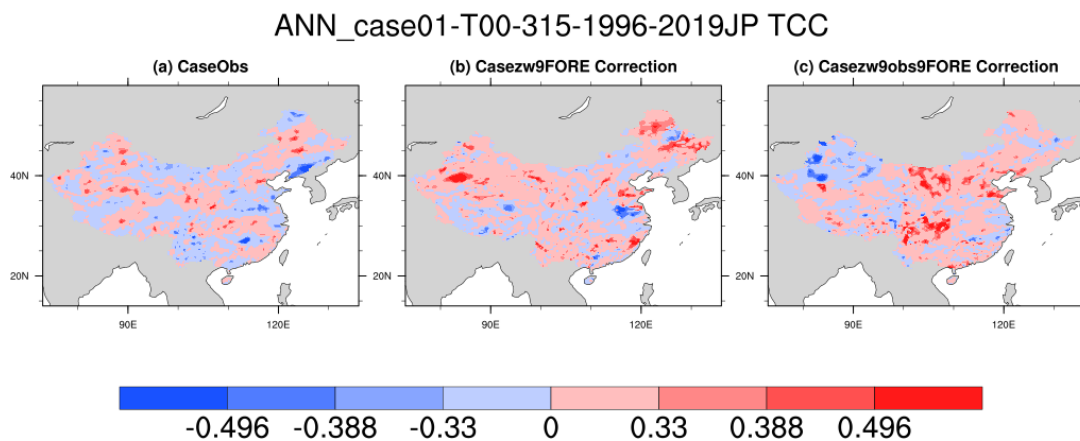

Fig. 9. TCC picture of the improved ANN for last three years forecast data revision results

As can be seen from the figure: first, the TCC range for CWRF mode is 0–0.3, while the TCC range for this paper's algorithm is 0.2–0.6, the difference between the two is 0.1. Thus, the correction effect of the algorithm model presented in this paper is substantial. However, the algorithm proposed in this paper has some shortcomings in modelling precipitation in a completely uniform manner for the national model, which makes it difficult to address regional differences.

5. Conclusion

Previous domestic and foreign studies have utilized a range of precipitation prediction methods, such as SSA, ARIMA, etc. In this paper, the neural network is modeled by looping and weighting into a composite model. The performance of the improved algorithm is then evaluated in terms of the mean absolute error (MAE), and time correlation coefficient (TCC).

The results show that the algorithm presented in this paper is effective for revising the precipitation forecast data for the whole of China. The MAE error of the new algorithm is decreased by about 50%, while TCC is improved by about 40%. It can thus be concluded that the proposed algorithm achieves good prediction performance. However, the algorithm proposed in this paper has some shortcomings in modelling precipitation in a completely uniform manner for the national model, which makes it difficult to address regional differences. Perhaps, can be improving prediction performance by Zoning Revision.

References

[1] Daniel Bannister, Andrew Orr et al., "Bias Correction of High‑Resolution Regional Climate Model Precipitation Output Gives the Best Estimates of Precipitation in Himalayan Catchments," *JGR Atmospheres*, vol. 124, no. 24, pp. 14220–14239, 2019. Article (CrossRef Link)

[2] X. M. Zhang, S. W. Xiong and L. H. Yu, "Discussion on the role of weather forecast in agricultural disaster prevention and reduction," *Agricultural science and technology and information*, vol. 36, no. 22, pp. 28–33, 2019. Article (CrossRef Link)

[3] Y. D Yu, Z. Y Yang, D.Y Qin, et al, "Appropriate spatial scale analysis for the simulation of precipitation by Regional Climate Model," in *Proc. of international Conference on Remote Sensing, Environment and Transportation Engineering. IEEE*, pp. 2967–2970, 2011.
Article (CrossRef Link)

[4] L.H Sun, W. X. Ai, W. X. Song and Y. M. Liu "Assessment Analysis on Winter and Spring Temperature and Rainfall Forecasts over China with Regional Climate Model," *journal of Applied Meteorological Science*, vol. 20, no. 5, pp. 547–553, 2009. Article (CrossRef Link)

[5] Y. F. Zhao, "A study of the regional model dynamic extended medium-term forecasting on persistent severe rainfall in southern China," Ph.D. dissertation, AMS, Bei Jing, 2017.

[6] W. H. Xing, C. L. Li and L. Wang, "Effect analysis of precipitation forecast in Yangtze River Basin Based on regional climate model regcm4," *Express Water Resources & Hydropower Information,* vol. 39, no. 10, pp. 1–15, 2018. Article (CrossRef Link)

[7] Sheau Tich Ngai, Liew Juneng et al., "Future projections of Malaysia daily precipitation characteristics using bias correction technique," *Atmospheric Research*, vol. 240, no. 1, pp. 1–9, 2020Article (CrossRef Link)

[8] X. Y. Li and J. H. Yu, "Comparison of bias correction techniques based on CWRF model for daily precipitation in summer," *Journal of Tropical Meteorology*, vol. 35, no. 6, pp. 842–851, 2019. Article (CrossRef Link)

[9] Z. H. Zhou, C. C. Gao and Y. H. Liu, "Regional grain yield response to climate change in China: A Statistic Modeling Approach," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 11, pp. 4472–4479, 2014. Article (CrossRef Link)

[10] D. X. Wang, "The important function and influence of weather forecast on agricultural production," *Modern Agricultural Science and Technology*, vol. 48, no. 17, pp. 198–203, 2019.
Article (CrossRef Link)

[11] L. Zhou, J. Pan, et al, "Correction Based On Distribution Scaling for Precipitation Simulated by Climate Model," *Journal of Applied Meteorological Science*, vol.25, no.3, pp.302-311, 2014. Article (CrossRef Link)

[12] C. H. Li, W. J. Pan and T. Wang, "A multi-scales patial-temporal projection method for monthly and seasonal rainfall prediction in guangdong," *Journal of Applied Meteorological Science*, vol. 29, no. 2, pp. 217–231, 2018. Article (CrossRef Link)

[13] Q. S. Wu, M. Han, M. Liu and F. J. chen, "A comparison of optimal-score-based correction algorithms of model precipitation prediction," *Journal of Applied Meteorological Science*, vol. 28, no. 3, pp. 306–317, 2017. Article (CrossRef Link)

[14] X. Y. Lu, M. Wang and X. Q. Wang, "Correction of TRMM monthly precipitation data from 1998 to 2013 in xinjiang," *Journal of Applied Meteorological Science*, vol. 28, no. 3, pp. 379–384, 2017. Article (CrossRef Link)

[15] Y. Lu, "Validation and evaluation of quantitative precipitation forecast in summer in China based on ensemble forecast and analysis of its influencing factors," Ph.D. dissertation, Nanjing University, Nanjing, 2018.

[16] Kamal Ahmed, D.A. Sachindra et al., "Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms," *Atmospheric Research*, vol. 236, 2020. Article (CrossRef Link)

[17] X. D Chen, L. Ruby Leung et al., "Predictability of extreme precipitation in western U.S. watersheds based on atmospheric river occurrence, intensity, and duration," *Geophysical Research Letters*, vol. 45, no. 21, pp. 693-701, 2018. Article (CrossRef Link)

[18] Herman G R, Schumacher R S, "Dendrology in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation," *Monthly Weather Review*, vol. 146, no. 6, pp. 1785-1812, 2018. Article (CrossRef Link)

[19] J. P Chen, Y. D Feng et al., "A Correction Method of Hourly Precipitation Forecast Based on Convolutional Neural Network," *Meteorological Monthly*, vol. 47, no. 1, pp. 60-70, 2021. Article (CrossRef Link)

[20] X. Y Dong, J. H Yu et al., "Bias Correction of Summer Extreme Precipitation Simulated by CWRF Model," *Journal of Applied Meteorological Science*, vol. 31, no. 4, pp. 504-512, 2020. Article (CrossRef Link)

[21] J. Sun and H. Y. Wen, "Application practice of artificial intelligence science in construction deformation prediction and control of soft soil underground engineering," *Tunnel construction*, vol. 40, no. 1, pp. 1-8, 2020. Article (CrossRef Link)

[22] UĞURLU M, SEVİM Ş, "Artificial neural network methodology in fraud risk prediction on financial statements; an empirical study in banking sector," *Journal of Business Research-Turk*, vol. 7, no. 1, pp. 60-89, 2015. Article (CrossRef Link)

[23] Q. J Gao and Q. P. Tu, "Comparison of the Climatic Characters of North Pacific SST before and after 1970s," *Journal of Nanjing Institute of Meteorology*, vol. 26, no. 2, pp. 243–254, 2003. Article (CrossRef Link)

[24] Y. M. Li, "Research and application of improved classifier algorithm based on BP neural network," Ph.D. dissertation, China University of Geosciences, 2019.

Tao Li: He is an associate Professor at School of Artificial Intelligence, Nanjing University of Information Science& Technology. He received the B.S. and M.S. degree in Nanjing University of Information Science & Technology, and PHD in Nanjing University of Aeronautics and Astronautics. His research interests mainly include machine learning, big data, data mining.

Wen Duo Xu: He is currently pursuing a master degree at the School of Artificial Intelligence in Nanjing University of Information Science & Technology, Nanjing, China. His current research interests include data mining and machine learning.

Li Na Wang: She is a lecturer at School of artificial intelligence, Nanjing University of Information Science & Technology. she received the B.S. and M.S. degree in Nanjing University of Information Science & Technology, and PHD in Nanjing University of Aeronautics and Astronautics. Her research interests mainly include data mining, artificial intelligence and network security.

Ning Peng Li: He is a graduate student of Nanjing University of Information Science & Technology Nanjing, China. His research interests are data mining and machine learning.
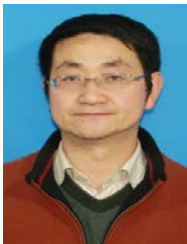
Yong Jun Ren: He obtained the PhD degree in the computer and science department at the Nanjing University of Aeronautics and Astronautics, China, in 2008. Now he is an associate Professor at School of Computer &Software, Nanjing University of Information Science& Technology. His research interests include security of cloud, blockchain and applied cryptography.

Jin Yue Xia: He received his PhD from the University of North Carolina in 2016, and MS degree from Computer and Science Department at the Nanjing University of Aeronautics and Astronautics, China in 2009. His research interests are applied cryptography and network security. He has published over 20 research papers that have appeared at many respected venues such as IEEE Transactions on Information Forensics & Security, ACM Transactions on Information & System Security, etc. Now, he is a senior researcher of IBM. His has served on the technical program committees for several IEEE/ACM conferences on data mining, data management and artificial intelligence.