Information gain-based modular fuzzy neural network to forecast rainstorms

Xiaoyan Huang, Li He, Huasheng Zhao, Ying Huan and Yushuang Wu

ABSTRACT

This study considers large-scale heavy rainfall as a forecast object based on the European central numerical forecast model product and uses a nonlinear fuzzy neural network (FNN) intelligent calculation method to establish a short-term forecast model of rainstorms. The information gain method is introduced to the predictor processing of the forecast model. Then the characteristics of many rainstorm predictors are calculated and screened on the basis of feature weight, information is condensed, some non-correlated forecast information variables are extracted, and the network structure of the forecast model is optimized. The modeled samples are determined and reconstructed by setting thresholds, and the modular forecast models of heavy rainfall and weak rainfall are established. The actual forecast results of the 24 h experimental prediction of the independent samples of large-scale rainstorms in Guangxi in 2012–2016 showed that the information gain-based modular FNN rainstorm forecasting model has higher prediction accuracy and a more stable forecasting effect. The various types of scores of 24 h of rainstorm ($\geq$50 mm) at 89 weather stations in Guangxi from 2012 to 2016 are: threat score (TS) is 0.368, ETS: equal threat score (E) is 0.141, hit rate (POD) is 0.296, empty report rate (FAR) is 0.559, forecast bias (B) is 0.671, and HSS skill score (H) is 0.247. Further comparison and analysis of the European Centre for Medium-Range Weather Forecasts (ECMWF) numerical forecasting model forecast results indicated that the new model performed nonlinear intelligence calculated interpretation modeling on ECMWF numerical forecasting model products, and forecasting accuracy is improved to a certain extent compared with that of the original model. Forecasting techniques are positive and have good release effects, thereby improving the rain forecasting ability of ECMWF to a certain extent and providing a better reference value for business forecasters.

Key words | fuzzy neural network, information gain, interpretation forecast, modular, rainstorm

Xiaoyan Huang (corresponding author)
Li He
Huasheng Zhao
Ying Huan
Yushuang Wu
Guangxi Research Institute of Meteorological Science,
Nanning 530022,
China
E-mail: gx_huangxy@163.com

HIGHLIGHTS

- Based on the European central numerical forecast model product and using a nonlinear fuzzy neural network intelligent calculation method to establish a short-term forecast model of rainstorms.
- The information gain method is introduced to the predictor processing of the forecast model.
- Actual forecast results showed that the new model has higher prediction accuracy and a more stable forecasting effect.

INTRODUCTION

China is located in the East Asian monsoon region. During the onset and duration of the summer monsoon, heavy rainstorms occur frequently, often causing heavy casualties and economic losses. With a complicated process and numerous influencing factors, the summer monsoon has been the focus of weather forecasting and is a difficult

problem in global weather forecasting (Tao *et al.* 2003; Zhao & Sun 2013).

With the continuous enhancement of the space-time resolution of a numerical forecasting model and the continuous improvement of physical modeling, the interpretation and application of numerical forecasting products have become a more effective method of forecasting rainstorms; strengthening them and focusing on the role of forecasters to improve the accuracy of disaster prediction (Reynolds 2003; Tian *et al.* 2008; Yan *et al.* 2008). To improve the forecast level of rainstorms locally and internationally, researchers have developed various rainstorm forecasting methods based on the interpretation of numerical models and numerical model products. Among them, the numerical prediction products of the batching method are common, and the forecasting effect on heavy rainfall during a rainstorm has also been improved (Zhang *et al.* 2010; Liu *et al.* 2015; Huang *et al.* 2016). Linear statistical forecasting modeling also plays an important role in the forecasting of the numerical model products of rainstorms. It mainly uses the statistical method to establish the statistical relationship between predictors and forecasting quantity of the output of a numerical forecasting product to conduct or improve objective factor forecasting (Liu & Ma 1996; Zhang *et al.* 2006; Jing *et al.* 2008; Li & Zhao 2009; Zhao *et al.* 2009; Zhong *et al.* 2009). A lot of study analysis shows that the linear regression method has been greatly improved, and is mainly used in the prediction of natural disasters. For example, the forecasting of precipitation, rainfall intensity, maximum wind speed, and hurricane central pressure which can cause natural disasters and damage through floods, rainstorms, droughts and hurricanes (Murnane & Elsner 2012; Prahl *et al.* 2012; Zhai & Jiang 2014; Lee *et al.* 2016; Choi *et al.* 2017; Choo *et al.* 2017; Kim *et al.* 2017; Choi *et al.* 2018; Mosavi *et al.* 2018; Fung *et al.* 2019). Forecasting factors and rainstorms mostly show nonlinear relationships; therefore, artificial neural network methods with good nonlinear information are also widely used in rainstorm forecasting (Liang *et al.* 2009; Wu *et al.* 2012).

The objective quantitative rainstorm precipitation forecasting method based on the interpretation of numerical forecasting products is an important development in the direction of short-term objective quantitative forecasts. Changes in heavy rainfall are complicated, caused by many influencing factors, and characterized by dynamism,

nonlinearity, and random uncertainty. In prediction, a signal processing algorithm should be modified as much as possible as the external environment changes to predict a storm by continuously adjusting algorithm parameters. In contrast to traditional mathematical models, fuzzy systems and neural networks can imprecisely handle uncertain information with their own characteristics. They can complement each other to good advantage, as a fuzzy system compensates for the abstract problem of neural network output expression, and the neural network compensates for the poor adaptive ability and robustness of the fuzzy system (Tian *et al.* 2002; Zhan *et al.* 2009; Zhang *et al.* 2009). On the basis of these phenomena, we aim to propose a numerical model rainstorm interpretation prediction method based on information gain-based feature optimization and a modular fuzzy neural network (FNN). Heavy rainfall at 89 stations in Guangxi is used as a forecast object. Rainstorm predictors are primarily selected on the basis of the analysis of various physical mechanisms affecting heavy rain, and the information gain calculation method is applied to obtain the optimal combination of factors. Then, sample models are further reconstructed on the basis of the European Centre for Medium-Range Weather Forecasts (ECMWF) value to establish the modular FNN forecasting models for heavy rainfall and light precipitation and to explore a new forecast method for heavy rainfall forecasting.

PRINCIPLE AND METHOD OF INFORMATION GAIN-BASED MODULAR FNN PREDICTION MODEL

The construction method of the forecasting model is essential for an excellent forecasting effect. In this study, the FNN method combined with the information gain predictor selection technique is used to establish a rainstorm forecasting modeling method for numerical forecast products with nonlinear intelligent calculation.

Information gain feature selection method

In machine learning, information gain is a commonly used indicator to measure whether a feature is good (Guo & Ma 2013; Han *et al.* 2017). The measurement standard of information gain is the amount of information the feature

can bring to the forecasting model, and the more information is available, the more important the feature will be. For a certain feature, the amount of information changes when it appears in the forecasting model and when it does not appear. The difference is the amount of information that the feature carries to the forecasting model. In information theory, the amount of information is 'entropy'. When the feature weight is calculated using information gain, the weight of the feature is determined by the amount of information, and the important features are screened by this standard. The formula of information gain is expressed as follows:

$$IG(w) = -\sum_{i=1}^{c} p(c_i) \ln p(c_i)$$
$$+ p(w) \sum_{i=1}^{c} p(c_i/w) \ln p(c_i/w)$$
$$+ p(\overline{w}) \sum_{i=1}^{c} p(c_i/\overline{w}) \ln p(c_i/\overline{w}) \qquad (1)$$

where c is the total number of rainstorm predictors; $p(c_i)$ is the probability that the characteristic c_i rainstorm predictor appears in the total rainstorm predictor concentration; $p(w)$ is the probability of feature item w in the rainstorm predictor; $p(c_i/w)$ is the conditional probability of the rainstorm predictor belonging to the c_i category when it includes the characteristic w; $p(\overline{w})$ is the probability that the rainstorm predictor does not contain the characteristic w; and $p(c_i/\overline{w})$ is the conditional probability of the rainstorm predictor belonging to the c_i category when it does not include w.

Fuzzy neural network structure and learning algorithm

A FNN combines the nonlinear processing ability of fuzzy sets with the advantages of artificial neural network self-organization and powerful self-learning ability; thus, it combines fuzzy computing, fast learning, and difficult nonlinear solving ability and greatly improves abilities. Among them, many methods, such as membership function synthesis, fuzzy clustering, fuzzy discriminant model, and fuzzy priority ratio, are often used in fuzzy mathematical methods (Xi 1987; Chen 1997). This study aims to use this calculation

method to conduct short-term forecasting modeling experiments on large-scale heavy rainfall in Guangxi and to add new forecasting tools for daily work.

FNN structure

The FNN structure of this study is a numerical four-layer feedforward network: input, membership generation, inference, and anti-fuzzy output layers (Wang 1998) (Figure 1).

The fuzzy rule form is:

if x_1 is μ_{1k} and x_2 is μ_{2k}, $\ldots$, then $y = \omega_k$, $k = 1, \cdots, m$

In the membership generation layer of the FNN of this paper, the more common Gaussian function is used as the membership function. This function has higher fitting and membership properties, and is more suitable for use when the amount of data is relatively large and the forecast accuracy is high. The Gaussian membership function is determined by only two parameters, and the formula is expressed as:

$$\mu_{ij} = \exp\left(-(x_i - a_{ij})^2/\sigma_{ij}^2\right), \ (i = 1, 2, \cdots, n \ldots j$$
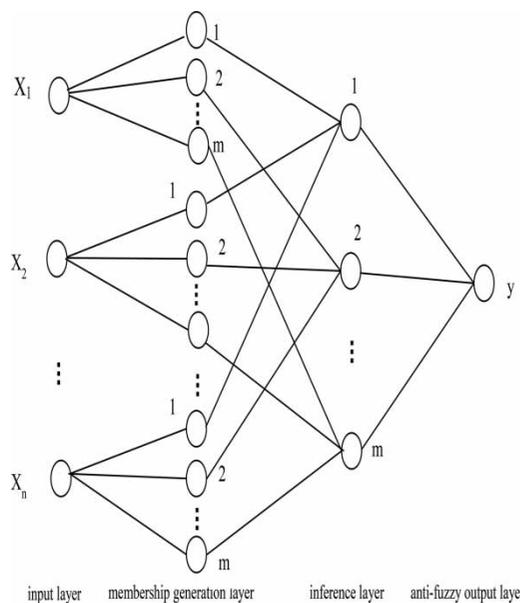$$= 1, 2, \cdots, m) \qquad (2)$$

Figure 1 | Fuzzy neural network structure.

where m is the fuzzy segmentation number, and a_{ij} and σ_{ij} determine the center and width of the membership function, respectively.

In fuzzy logic, according to the 'and' operation, the three calculation methods include: minimum mode; multiplication mode; and addition method. The research results showed that (Zeng & Singh 1996) the performance of the multiplication mode is better than that of the minimum mode. In this study, the method is selected in the inference layer, that is, the output value of each node is the algebraic product of all inputs of the node:

$$\pi_j = \mu_{1j} \times \mu_{2j} \times \cdots \times \mu_{nj} = \prod_{i=1}^{n} \mu_{ij}, \ (j = 1 , \cdots, m) \qquad (3)$$

In the output layer, the following anti-fuzzy network output is used:

$$y = \omega_1 \pi_1 + \omega_2 \pi_2 + \cdots + \omega_m \pi_m \qquad (4)$$

where $\omega_j \ (j = 1 , \cdots, m)$ is the connection weight.

FNN learning algorithm

When the FNN method is used to establish the rainfall forecasting model in Guangxi, it is essentially a multi-layer feedforward network; hence, the feedforward network of the BP (back propagation) algorithm is used to train the adjustment parameters, and the main parameters to be learned are the connection weight of the network, $\omega_j (j = 1, 2, \cdots, m)$, and the central, a_{ij}, and width values of the membership function of the membership generation layer, $\sigma_{ij} \ (1 \leq i \leq n; 1 \leq j \leq m)$.

First, we define the output error of the network (i.e., the objective function) as follows: $E_p = \dfrac{1}{2}(y - Y)^2$, where y is the actual output value, Y is the expected output value, and E_p is the squared error function. In the process of learning and training, the gradient descent method is used to obtain the learning rules of the central value a_{ij}; the width σ_{ij} and the weight ω_j of the membership function

which are as follows:

$$a_{ij}(n + 1) = a_{ij}(n) - \alpha \frac{\partial E_p}{\partial a_{ij}} \qquad (5)$$

$$\sigma_{ij}(n + 1) = \sigma_{ij}(n) - \alpha \frac{\partial E_p}{\partial \sigma_{ij}} \qquad (6)$$

$$\omega_j(n + 1) = \omega_j(n) - \alpha \frac{\partial E_p}{\partial \omega_j} \qquad (7)$$

where α is the learning rate $(0 < \alpha < 1)$. $E_p = \frac{1}{2}(y - Y)^2$ is substituted into Formulas (5), (6), and (7), and the derivation rule of the compound function is used to obtain the updated formulas for the membership function parameters and weights as follows:

$$a_{ij}(n + 1) = a_{ij}(n) - \alpha \frac{\partial E_p}{\partial a_{ij}}$$

$$= a_{ij}(n) - \alpha(y - Y) \frac{\partial y}{\partial a_{ij}}$$

$$= a_{ij}(n) - 2\alpha(y - Y)\omega_j \times \prod_{i=1}^{n} \mu_{ij} \times (x_i - a_{ij})/\sigma_{ij}^2 \qquad (8)$$

Similarly, the following can be obtained:

$$\sigma_{ij}(n + 1) = \sigma_{ij}(n) - 2\alpha(y - Y)\omega_j \times \prod_{i=1}^{n} \mu_{ij}$$

$$\times (x_i - a_{ij})^2/\sigma_{ij}^3 \qquad (9)$$

$$\omega_j(n + 1) = \omega_j(n) - \alpha(y - Y)\pi_j \qquad (10)$$

In the process of network training, the membership function is set to Formula (2). First, the constructed predictive factor is used to learn matrix X and train the network. The main calculation steps are included as follows:

(1) The initial value is set as $i = 1$. The learning rate α is 0.9, and the network training error is e_1.
(2) At the initial moment, the connection weight to the network, the center value of the membership function, and the width value are initialized with a random number.
(3) The BP algorithm is used to train and adjust the parameters and weights.

(4) The error of the actual output of the network and the expected output e_2 is calculated.

(5) When $e_2 > e_1$, step (3) is repeated. Otherwise, when the network training ends, the network parameters and connection rights obtained by the training are used for prediction.

Combined design of information gain and FNN forecasting model

The FNN interpretation forecast modeling of a large-scale rainstorm in Guangxi is mainly designed by combining information gain and the FNN. The introduction of information gain is mainly to control the factor input matrix of the FNN. In general, in the training of the FNN model, the input layer parameters of the network should be sufficient. Otherwise, the network structure is complicated, and the amount of training and learning time are extended, but too few parameter inputs do not reach the predicted accuracy because of the lack of information. Therefore, in the factor input parameter of the FNN input parameters, the appropriate input parameter should be selected. The optimal combination method of the information factor of the information gain is then introduced to select and reconstruct the original storm predictor. On the basis of the calculated information factor on the characteristic of each factor, we determine the role of the attribute, select the feature factor with a large amount of information and a large contribution rate to the prediction accuracy, and eliminate the factor with a small shadow. The information gain method is used to optimize the network structure of FNN and conduct model prediction in accordance with the following steps (Figure 2):

(1) The predictors for numerical forecast products are pre-selected.

(2) The feature extraction of information gain is conducted for each original predictor field.

(3) The extracted information gain characteristic value is correlated with the forecast quantity, and the correlation significance test is performed to obtain several characteristic values condensed with all vector features as the input nodes of the neural network.

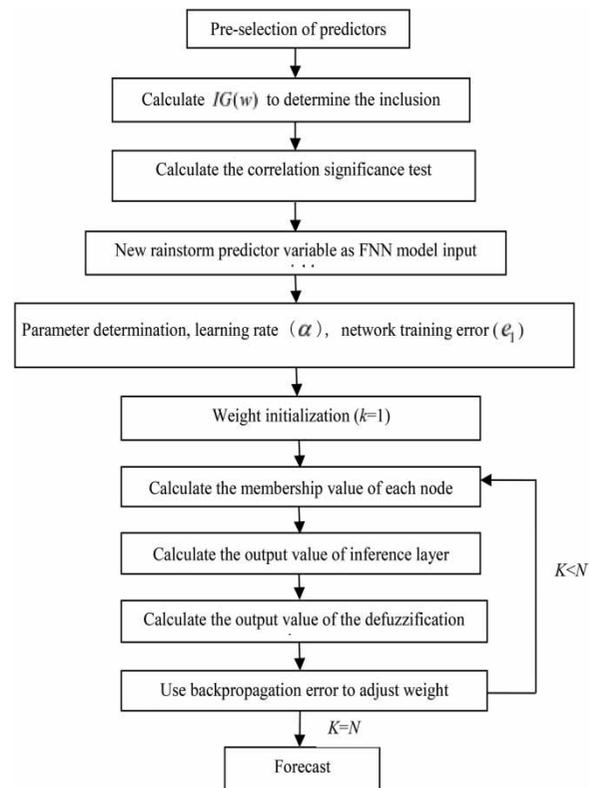(4) The threshold is set on the basis of the forecast result of the European central model product, and the

Figure 2 | Forecast design process of the modular FNN based on information gain for rainstorms.

independent forecast samples are modularized on the basis of the rainfall amount to construct modular forecast models of rainstorm and no rainstorm.

(5) The modular FNN forecasting model is used to predict the heavy rainfall data of the 89 stations in Guangxi, and the prediction results are obtained.

TEST OF INFORMATION GAIN-BASED MODULAR FNN TO FORECAST RAINSTORM INTERPRETATION

Data of forecast test

Guangxi is located in the southern part of China and bordered by the South China Sea. It is affected by middle and high latitudes and tropical ocean circulation. Hence, precipitation is high, and heavy rains often occur, causing floods and disasters, which have a major impact on people's lives and economic assets. In this study, the rainstorm

weather processes of 89 observation stations in Guangxi are used as the forecasting object. When modeling and testing the nonlinear intelligent calculation method for large-scale heavy rainfall in Guangxi, the basic rainstorm precipitation data are daily precipitation in Guangxi from a total of 35 years between 1982 and 2016 (from 20:00 to 20:00, Beijing time, the same as below). Given that this paper is mainly considering the forecasting of rainstorms in the base daily precipitation data of >35 years, the daily precipitation of ≥50 mm is selected, and the precipitation data exceeding this value for 10 stations are taken as a test sample. A total of 690 rainstorm samples are extracted, and the data of 1982–2011 (a total of 579 samples) are used as the basic sample for modular modeling. In the following 5 years, that is, from 2012 to 2016, 111 samples are used as independent samples to conduct prediction experiments consistent with actual business forecasts.

The physical data of numerical forecast products are mainly derived from the global re-analysis data (https://www.ecmwf.int/) of the European Center for Medium-Range Weather Forecasts (ECMWF) ERA interim. The spatial resolution is $0.75° \times 0.75°$ latitude and longitude grid distance, and the time resolution is 6 h, including four world times of 00:00, 06:00, 12:00, and 18:00. Moreover, there are data on temperature, wind field, geopotential height field, sea level pressure field, and a physical index of various precipitation types at four levels (200, 500, 700, and 850 hPa), over a total of 35 years (1982–2016). The selected physical quantity field range was as follows: latitude of $9.75°$ N–$40.5°$ N; longitude of $79.5°$ E–$120°$ E; and a total of 2,310 grid points (Figure 3).

Preselection of rainstorm predictors

The FNN is essentially a statistical forecasting method. Hence, the selection of historical samples and the construction of predictors are essential for the model construction. In a study on the large-scale heavy rainfall forecasting method in Guangxi, the physical quantity field of global re-analysis data from ECWMF ERA interim is used as a primary rainstorm predictor (hereinafter referred to as ECMWF). In the selection of the range of predictors, we know that the atmosphere is a process of continuous motion with time. In general, the occurrence of a weather phenomenon is closely
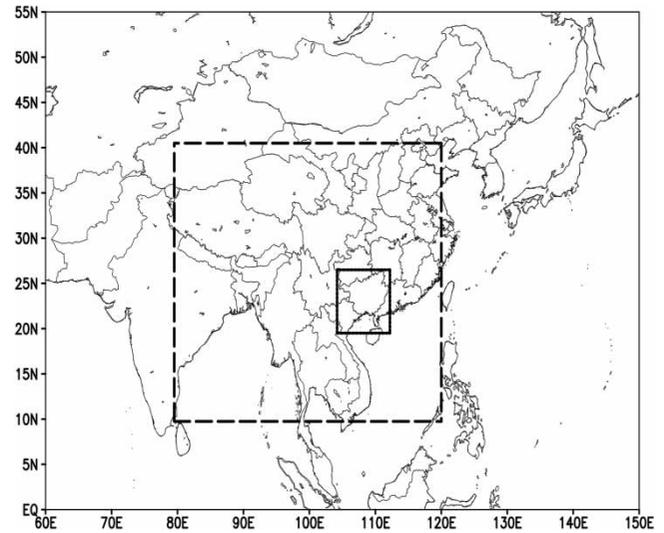

Figure 3 | Range of predictor field values (long broken line box) and forecast area (short broken line box).

related to local atmospheric circulation conditions and large-scale environmental field changes. Guangxi is located in southern South China, and its area is confined to latitude $19.5°$ N–$26.5°$ N; longitude $104.2°$ E–$112.2°$ E, and the range is relatively small for large-scale atmospheric movements (Figure 3). If only the circulation factor in the range is selected as the predictor and the change in atmospheric motion cannot be accurately described, we expand the selection range in the primary selection of factors to $9.75°$ N–$40.5°$ N and $79.5°$ E–$120°$ E, with a total of 2,310 grid points of physical field data as candidate factors.

The most important principle for selecting predictors in the rainstorm forecast experiment here is to judge the correlation with the forecast object (rainstorm). When calculating the correlation coefficients of rainstorm forecast factors, it is found that the most relevant forecast factors are mainly concentrated in the 24 h before the weather phenomenon occurs, followed by 48 h, and finally 72 h. As time moves forward, the correlation becomes lower and lower, and the number of predictors is also very small. If other time-dependent forecasting factors are added, it will take more time to forecast and affect the speed of forecasting. At the same time, it does not significantly improve the accuracy of forecasting. Therefore, in terms of timeliness, the data of 24, 48, and 72 h before the weather phenomenon occurs are mainly selected.

The occurrence and development of rainstorms are inseparable from the water vapor, energy, and thermal conditions. To fully explore various factors causing rainstorms, we extracted 24 types of heavy rainfall precipitation physical quantity forecast factors in the same period of the EC numerical forecasting model, and water vapor factors, including middle and high levels, specific humidity, relative humidity, water vapor flux, water vapor flux divergence, and water vapor advection. Dynamic factors include divergence, vertical velocity, vorticity, and wind field UV component. Thermal conditions mainly include total temperature field, total temperature advection, temperature advection, temperature, and false equivalent temperature. Various physical quantity indexes related to heavy rainfall are preferred: K index, ky index, SI index, and uplifting index. These 24 field factors are used as basic predictor groups, which are further combined and calculated to obtain more physical factors related to rainstorms.

Factor processing and extraction

Our census reveals that a large number of primary rainstorm predictors can be obtained. If all of these predictors are used in the model for modeling and forecasting, the input structure of the forecasting model is too large and easily learned. The information gain method is used to reduce the dimensionality of the predictor group to reduce the input nodes of the forecasting model and retain the forecast information of all predictors as much as possible. The specific practices are as follows:

(1) In the extraction of information gain feature factors, the information gain is calculated for all of the grid points (2,310 grid points) of each physical quantity field, and the first few characteristic components with a high information gain rate are selected to obtain the set of the feature factors of physical quantity.

(2) The feature factors obtained by filtering all the physical quantity fields after information gain are combined to form a new predictor set. The new factor concentration predictors are numerous and produce the overlearning problems when they are directly used in the forecasting model. In the selected feature predictor set, it does not indicate that it has a high

correlation with the forecast amount, although the weight of the information gain rate is high. Therefore, this method should further carry out the correlation calculation analysis with the forecast quantity and screen out highly relevant predictors to form a new forecast factor set. The criteria for the selection of predictors are mainly determined by multiple control experiments on the rainstorm forecast model. When the standard of selected correlation coefficient decreases, it is found that the number of predictors will increase greatly. After increasing the standard value, it is found that the amount of forecast information contained in the forecast factor concentration will gradually weaken, resulting in a decline in forecast accuracy. After many experiments, it is finally determined that the selection criterion of the predictor is the absolute value of the correlation coefficient ≥ 0.3.

(3) The newly formed high correlation predictor group should be further tested for significance to control the complex collinearity relationship between the factors, and the physical significance of the information gain rainstorm precipitation predictor can be decomposed. The noise factor affecting the forecast effect is removed to control the number of predictors selected into the final forecast model. In general, the appropriate predictor matrix structure can obtain a higher forecasting accuracy when the forecasting model is utilized, but the model input is too large and can easily cause overfitting problems (Jin *et al.* 2004).

Modular FNN rainstorm forecast

In this study, in the large-scale heavy rainfall forecast experiment in Guangxi, the FNN modeling method is used to establish the corresponding FNN rainstorm forecasting model for each forecasting station in Guangxi. In the calculation of the FNN algorithm, the setting of the parameters is unified as follows: the number of predictors of each site selected is the input node of the network, the output node is 1, the number of inference layer sections is 3, the number of network training times is set to 500, the learning factor is 0.9, and the overall error is set to

0.0001. The set FNN forecasting model is used to test and inspect 111 independent forecast samples of rainstorm precipitation at 89 stations in Guangxi according to the actual business forecast. By using this model, we predicted and tested 111 independent forecast samples of rainstorms at 89 meteorological observatory sites in Guangxi. When the forecasting model predicted the 580th independent forecasting sample, it first forecasted the first independent sample by using the preceding 579 modeling samples to establish the forecasting model and then added the 579 modeling samples to the first independent sample, 580 samples as a model sample, the second independent sample of the forecast, and so on. Finally, 689 samples were used as a model sample for the last 690 independent samples for forecasting, and the forecasts of the independent samples were in line with the actual forecast.

In the study of rainstorm forecasting modeling for 89 sites in Guangxi, we conducted independent modeling and forecasting for substations, that is, 89 different forecasting models were established to forecast for 89 forecasting stations. In the sample selection, the daily precipitation of ≥ 50 mm and the precipitation data over 10 stations were used as the test sample. However, this finding does not indicate that all of the 89 sites in Guangxi have rainstorm weather. Instead, a certain threshold is set to reclassify the modeling sample set of the station's modeling samples so that each site can be modeled and forecasted more specifically. Specifically, in the actual forecast, we first model the first sample of the 111 independent samples. At present, the forecasting model of the ECMWF numerical forecasting model has stable rainfall forecasting performance and high forecasting accuracy. To control the empty report and false report rate of rainstorms and further improve the forecasting accuracy rate, we assume in the single-station rainstorm forecasting model based on the rain forecast products of ECMWF numerical forecast products that there are two threshold parameters A and B, where A represents the threshold of the precipitation above the rainstorm with a large probability of ECMWF rainfall forecast value. On the contrary, nearly no rainstorm is observed, and B is the actual precipitation. When the value of the ECMWF prediction interpolation to the site reaches A or above, the

real precipitation may occur after a rainstorm. At the same time, the threshold set by parameter B is generally smaller than that by parameter A because an error exists between the predicted value of the model and the actual precipitation. The threshold is set, and a modular classification processing method for the basic modeling samples of the model is adopted, specifically to check the size of the ECMWF prediction value of the station. When the interpolated ECMWF prediction is further interpolated to a value above A, a sample with a historical live precipitation reaching B or higher is searched for in the basic modeling sample to reorganize the sample, and a rainstorm forecast model of the station is established. When the value is less than threshold A, the basic modeling sample is reselected, and the samples with historical live precipitation less than B are searched for in the basic modeling samples to reconstruct the samples and construct a new non-storm heavy rainfall prediction model. In the setting of the threshold, we conducted a number of tests and finally determined that when the threshold A is 20 mm and B is 15 mm, the obtained prediction effect is the best. The specific modeling steps are as follows:

(1) The ECMWF precipitation forecast field is used as a reference standard for rainfall to identify the 24 h rainfall at each site.

(2) The polynomial interpolation method is utilized. The ECMWF 24 h rainfall forecast field is interpolated to 89 meteorological observatories in Guangxi so that each forecast object (weather station) has a rainfall value, which is recorded as Rn.

(3) For the kth ($k = 1, \ldots, 89$) forecast object Y (weather station), if its Rn mean is greater than the threshold A mm, then the sample number set $\Omega_1 (\Omega_1 \subset \Omega)$ whose precipitation is greater than B mm is selected in the historical sample sequence of the forecast object Y in which the sample number is recorded as Ω.

(4) For the modeled sample number set Ω_1 of the forecast object Y, the corresponding factor matrix S and the forecast amount sequence Y' composed of the information gain method selection can be obtained.

(5) The FNN model algorithm is used on the basis of the selected factor matrix and Y' to establish the ensemble

prediction model and to obtain the rainfall forecast value for the station for the next 24 h.

FORECASTING MODEL PREDICTION TEST AND RESULT ANALYSIS

In this study, when the data from 1982 to 2011 were used as the basic modeling samples, a total of 111 samples from the rainstorm process in Guangxi during the period of 5 years from 2012 to 2016 were used as independent forecast samples for the 24 h forecast aging test consistent with the actual business forecast. The model is based on the modeled samples of each site identified above, and the input factors for the FNN forecasting model are selected using the final predictor input model (see sections on 'Fuzzy neural network structure and learning algorithm' and 'Combined design of information gain and FNN forecasting model').

The forecast effect of the new interpretation method, and whether it is suitable for business forecast applications, should be determined; hence, the forecast results should be determined and examined. This study mainly compares the experimental forecast results of the new scheme with the actual precipitation, calculates various scoring indicators, and visually evaluates the precipitation forecasting effect of the new scheme. This study tests the prediction accuracy of this method mainly in terms of threat score (*TS*), ETS: equal threat score (*E*), empty report rate (*FAR*), hit rate (*POD*), forecast bias (*B*), HSS skill score (*H*), and other scoring indicators. The forecast bias (*B*) indicates the ratio of the number of occurrences of the forecast event to the number of times the observed event occurred. The equal threat score is an improvement on the threat score, which can penalize empty or missing reports, making the score fairer than the latter. The HSS skill score can also punish empty or missing reports, and the expected score for the random

Table 1 │ Classification for rainstorm verifications

Actual observation	Forecast	
	Yes	**No**
Rainstorm	N_A	N_C
No rainstorm	N_B	N_D

and constant forecast is 0. The mathematical attribute is linear and progressively fair, which is also one of the fair precipitation tests (Table 1) (Wang & Yan 2007; Wu *et al.* 2017). The specific scoring formulas are as follows:

Threat score: $TS = N_A/(N_A + N_B + N_C)$ (11)

Hit rate: $POD = N_A/(N_A + N_C)$ (12)

Empty report rate: $FAR = N_B/(N_A + N_B)$ (13)

Forecast bias: $B = N_A + N_B/N_A + N_C$ (14)

Equal threat score (ETS):
$$E = N_A - R(a)/(N_A + N_B + N_C - R(a))$$ (15)

HSS skill score: $H = N_A + N_D - R(a) - R(d)/$
$$(N_A + N_B + N_C + N_D - R(a) - R(d))$$ (16)

where

$$R(a) = (N_A + N_B)(N_A + N_C)/(N_A + N_B + N_C + N_D),$$

$$R(d) = (N_B + N_D)(N_C + N_D)/(N_A + N_B + N_C + N_D)$$

where N_A is the number of correct stations (times), N_B is the number of empty report stations (times), N_C is the number of missing report stations (times), and N_D is the correct number of stations (times) for which the forecast and live conditions have not reached the threshold (Table 1). According to the definition, *TS* is the threat score for a rainstorm forecast. The ideal score for the equal threat score is 1, ranging from $-1/3$ to 1, with 0 indicating no skill. However, penalties occur for both empty and missing reports. Hence, the source of the forecast error cannot be distinguished. In general, the score is lower than the threat score, and the equal threat score in the region with more precipitation is significantly lower than the threat score. R_a and R_d are the mathematical expectations of the random forecasts N_A N_D when the number of empty report stations and the number of missing report stations are equal.

The above scoring formula is used to obtain the various types of scores of 24 h of rainstorms ($\geq$50 mm) at 89 weather stations in Guangxi from 2012 to 2016. The new scheme is essentially a method of interpretation and forecasting of numerical forecast products to verify the

advantages and disadvantages of the new scheme for the forecast of heavy rainfall precipitation. To understand the interpretation and prediction ability of this new method proposed in this paper, we calculated the various types of scores of the rainstorm forecasts of the 89 meteorological observatories in Guangxi in the past five years, and the forecast results are compared with those of ECMWF in the same segment and in the same region. First, the data of the ECMWF prediction grid for the station should be interpolated, where bilinear interpolation is utilized for interpolation (the same as below). Table 2 and Figure 4 present scores using the new method for the forecast results in 2012–2016 compared with scores for the corresponding forecast period of the interpolated ECMWF rainfall data, for 89 meteorological stations in Guangxi of the yearly and 5-year average of annual rainstorms of various grades.

The modular FNN rainstorm forecasting model based on information gain has an average TS of 0.37 for the 24 h aging rainstorm ($\geq$50 mm) at 89 weather stations in Guangxi from 2012 to 2016, and ECMWF has a TS of 0.17 for the 24 h aging rainstorm. The score for the new method is 0.20 higher, which is an increase of 120.20%. The accuracy of the new method of rainstorm forecasting is more obvious. The rainstorm forecast results of the new mathod significantly exceed the ECMWF rainstorm forecast results for 5 years through the further analysis of the average rainstorm threat scores of 89 stations in each year from 2012 to 2016. Each year improves by: 79.15%, 123.56%, 107.01%,

152.59%, and 148.02%; except for 2012, the increases in the other four years are more than double (Figure 4(a)).

The equal threat scores were significantly lower than the threat scores. Overall, the equal threat scores of the new method forecasts exceeded the ECMWF forecasts. In 2012–2016, they increased by 6.50%, 90.02%, 22.72%, 67.41%, and 27.31%. These equal threat scores were less improved in 2013. However, in comparison with ECMWF, they were greatly improved in 2013 and 2015 (Figure 4(b)).

The results of the empty report rate (FAR) showed that the ECMWF's empty report rate was high (Figure 4(c)). In our specific analysis, in 2012, the empty report rate of the new method was 0.53, and that of the ECMWF was 0.49, indicating that the new method had more empty reports than ECMWF. In 2013, 2014, and 2015, the ECMWF report rate was higher than that of the new method. In 2016, the results of the empty report rates of the two methods were equivalent.

In the comparison of the torrential rain hit rate (POD), the new hit rate for rainstorms in each year was significantly higher than that in the ECMWF forecast, increasing by 21.33%, 42.10%, 16.46%, 64.86%, and 34.82% in 2012–2016, demonstrating a good release forecast effect (Figure 4(d)).

According to the analysis of the forecast deviation (B) (Figure 4(e)), the new method predicts that the ratio of the number of rainstorms to the actual number of observations increases yearly and is closer to 1 than the ECMWF prediction result. Hence, it is closer to the actual observation value.

Table 2 | Average various scores of the new method (FNN) and ECMWF for the 2012–2016 independent sample forecast of rainstorms at 89 weather stations in Guangxi (unit:%)

| Scoring index	Model	2012	2013	2014	2015	2016	Average
TS	FNN	42.35	36.62	40.78	32.71	32.02	36.84
	ECMWF	23.64	16.38	19.7	12.95	12.91	16.73
ETS	FNN	18.58	16.76	13.83	12.74	9.93	14.08
	ECMWF	17.44	8.82	11.27	7.61	7.80	10.18
FAR	FNN	53.75	57.81	52.58	55.05	61.07	55.88
	ECMWF	49.17	68.85	55.37	58.78	60.91	59.22
POD	FNN	37.19	36.49	30.36	26.18	21.80	29.62
	ECMWF	30.65	25.68	26.07	15.88	16.17	22.11
B	FNN	80.40	86.49	64.03	58.24	56.02	67.14
	ECMWF	60.30	82.43	58.42	38.53	41.35	54.21
HSS	FNN	31.34	28.7	24.31	22.6	18.07	24.69
	ECMWF	29.7	16.21	20.26	14.14	14.47	18.48
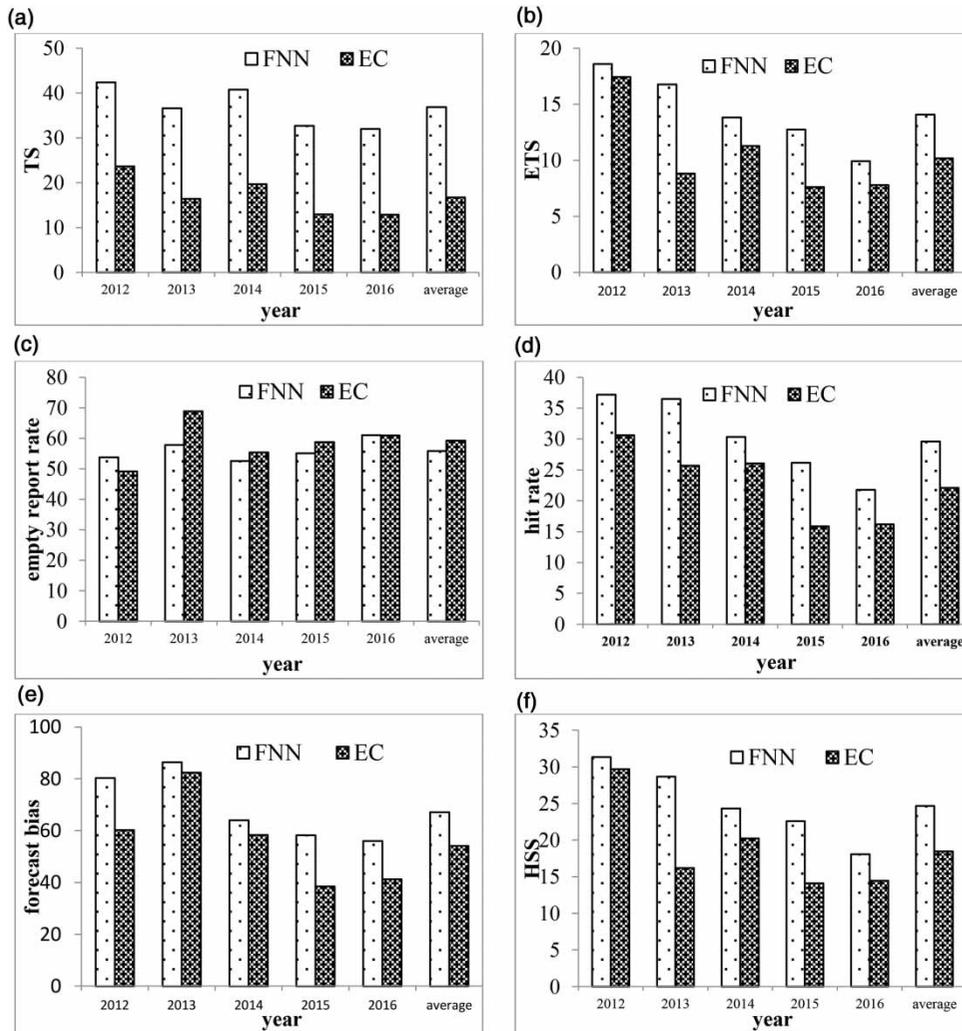
Figure 4 | Average threat score (a), equal threat score (b), empty report rate (c), hit rate (d), forecast bias (e), and HSS skill score (f) (unit: %) for the new method (FNN) and ECMWF (EC) for the 2012–2016 independent sample forecast of the rainstorm level at 89 weather stations in Guangxi.

The HSS forecasting skill score shows that the actual forecast of the new method exceeds the forecasting skill of the ECMWF, increasing by 5.52%, 77.05%, 19.99%, 59.83%, and 24.88% in 2012–2016. The 4-year forecasting skills improved significantly, with an average increase of 33.60% over 5 years (Figure 4(f)).

The results of several torrential rainstorm test statistics show that the forecasting skills of the new model relative to the ECMWF are positive techniques with stable forecast performance and good release effect, which can improve the rainstorm forecasting ability of the ECMWF model to some extent. The improvement is closely related to the processing method of information gain factor, the

reconstruction of modeling samples, and the application of the FNN model.

CONCLUSION AND DISCUSSION

This study considers large-scale heavy rainfall as a forecast object based on the European central numerical forecast model product. The information gain method is introduced to the predictor processing of the forecast model. Then, a short-term forecast model of rainstorms is established by using the FNN intelligent calculation method. Using this model to predict large-scale rainstorms in Guangxi from

2012 to 2016, and comparing the results with the ECMWF model's forecasting results, the following conclusions are obtained:

(1) In this study, the information gain method was introduced to evaluate the feature information of many rainstorm forecasting factors in the primary selection, obtain new characteristic variables, and further acquire comprehensive characteristic information that can describe the original forecasting factors to the greatest extent by calculating the correlation and conducting a significance test. As a new input forecasting feature factor of the FNN model, the model could achieve the purpose of dimensionality reduction and remove redundant information, effectively decreasing the correlation between parameters, optimizing network structure, and enhancing performance.

(2) When the rainstorm forecasting modeling study was carried out on 89 forecasting stations in Guangxi, thresholds were set on the basis of ECMWF forecasting products to reclassify the modeled samples for rainstorm and non-storm samples so that the module could be better modeled for prediction.

(3) Rainstorm forecasting still has problems such as the inability to accurately evaluate the degree of influence of various factors. In order to solve the problem of rainstorms being affected by the interaction and coupling of multiple factors, many scholars currently use artificial neural network methods for prediction and analysis. However, due to the limitations of the neural network, the selection of weights and thresholds was relatively random, resulting in a low probability of obtaining the global optimal value. The difference between a fuzzy neural network and a traditional artificial neural network is that the FNN combines fuzzy theory with a neural network, and integrates the advantages of a neural network's nonlinear processing ability and fuzzy theory's logical reasoning ability. It is a hybrid intelligent optimization method. In contrast to the implementation processes of other feedback neural networks, fuzzy neural networks fuzzify the weights to make them have logical inference meaning. The characteristic of this network is that its network parameters are defined by fuzzy theory, which can handle fuzzy

information and has strong fault tolerance. This fuzzy information processing method can break through the limitation of domain search, realize distributed collection, and is beneficial to quickly search for the optimal value. In our experiment, the fuzzy logic reasoning ability of the fuzzy neural network algorithm and the nonlinear processing ability of the neural network were used to construct the main influencing factors of the rainstorm and its own network optimization model to realize the accurate description and prediction of the rainstorm process. From the comparative analysis of the forecasted rainstorm value and the actual value in the experiment, it can be seen that the forecast accuracy was significantly higher than the forecast result of the numerical forecast model, indicating that the fuzzy neural network model has good reliability and accuracy, and has good performance for the weather forecast business. Moreover, the method can be widely used in the forecast of other weather disaster elements, such as the forecast of wind, temperature and other elements.

(4) The actual forecast results of the 24 h timeliness test of large-scale heavy rainfall in Guangxi in 2012–2016 showed that the new method has better forecast results and a stable forecasting effect.

(5) The forecasting model of heavy rainfall proposed in this paper was based on the numerical forecasting products. Therefore, the good or poor precipitation forecasting ability above the rainstorm level of the numerical forecasting model would directly affect the forecasting ability of the forecasting model proposed in this paper. However, using the forecast model in the rainstorm level forecasts of 89 stations in Guangxi showed that the forecasting ability of the forecasting product was more accurate than the interpolation of the numerical forecasting model, and had a better reference value for business forecasting personnel, through the method proposed in this paper.

ACKNOWLEDGEMENTS

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Chen, G. F. 1997 Fuzzy mathematics and weather forecast. *Meteorological Monthly* **6** (6), 8–11. https://doi.10.7519/j. issn.1000-0526.1979.06.008.

Choi, C. H., Kim, J. S., Kim, J. H., Kim, H. Y., Lee, W. J. & Kim, H. S. 2017 Development of heavy rain damage prediction function using statistical methodology. *Journal of Korean Society of Hazard Mitigation* **17** (3), 604–612. doi:10.9798/ kosham.2017.17.3.331.

Choi, C. Y., Kim, J. W., Kim, J. S., Kim, J. S., Kim, D. Y., Bae, Y. H. & Kim, H. S. 2018 Development of heavy rain damage prediction model using machine learning based on big data. *Advances in Meteorology*. Article ID 5024930. 1–10. https:// doi.org/10.1155/2018/5024930.

Choo, T. H., Kwak, K. S., Ahn, S. H., Yang, D. U. & Son, J. K. 2017 Development for the function of wind wave damage estimation at the western coastal zone based on disaster statistics. *Journal of the Korea Academia-Industrial Cooperation Society* **18** (2), 14–22. Available from: http://www.dbpia.co.kr/journal/ articleDetail?nodeId=NODE07199910&language=ko_KR.

Fung, K. F., Huang, Y. F., Koo, C. H. & Soh, Y. W. 2019 Drought forecasting: a review of modelling approaches 2007–2017. *Journal of Water and Climate Change (2020)* **11** (3), 771–799. https://doi.org/10.2166/wcc.2019.236.

Guo, S. & Ma, F. 2013 Improving the algorithm of information gain feature selection in text classification. *Computer Applications and Software* **30** (8), 139–142. https://doi:10.3969/j.issn. 1000-386x.2013.08.037.

Han, J. B., Halidan, A., Gulnur, A. & He, Y. 2017 Improved information gain algorithm based on Uyghur feature selection. *Computer Engineering and Applications* **53** (23), 34–38. https:// doi.10.3778/j.issn.1002-8331.1607-0312.

Huang, Y. S., Ding, G. Y. & Yu, J. H. 2016 Forecasting heavy rains of raining seasons of Nanping city based on a burdening method with products of physical quantities. *Guangdong Meteorology* **38**, 37–40. https://doi.10.3969/j.issn.1007-6190. 2016.04.009.

Jin, L., Kuang, X. Y., Huang, H. H., Qin, Z. N. & Wang, Y. H. 2004 Study on the over fitting of the artifical neural network forecasting model. *Acta Meteorologica Sinica* **62**, 62–70. https:// doi.10.3321/j.issn:0577-6619.2004.01.007.

Jing, L. H., Wang, N. G. & Zhao, Y. J. 2008 The typhoon depression rainstorm forecasts in Xianning of Hubei province. *Torrential Rain and Disasters* **27**, 149–153, 165. Available from: http://www.byzh.org.cn/CN/article/ searchArticleResultByKeyword.do.

Kim, J. S., Choi, C. H., Lee, J. S. & Kim, H. S. 2017 Damage prediction using heavy rain risk assessment: (2) development of heavy rain damage prediction function. *Journal of Korean Society of Hazard Mitigation* **17** (2), 371–379. doi: 10.9798/ kosham.2017.17.2.371.

Lee, J. S., Eo, G., Choi, C. H., Jung, J. W. & Kim, H. S. 2016 Development of rainfall-flood damage estimation function using nonlinear regression equation. *Journal of the Korean Society of Disaster Information* **12** (1), 74–88. doi:10.15683/ kosdi.2016.3.31.74.

Li, B. & Zhao, S. X. 2009 Development of forecasting model of typhoon type rainstorm by using SMAT. *Meteorological Monthly* **35** (6), 3–12. https:10.7519/j.issn.1000-0526.2009.6.001.

Liang, Z. Q., Jin, L. & Gong, Y. F. 2009 The research on methods of forecasting local rainstorms in low latitude area using the neural network. *Journal of Tropical Meteorology* **25**, 458–464. https://doi.10.3969/j.issn.1004- 4965.2009.04.011.

Liu, Y. J. & Ma, K. Y. 1996 A study of heavy rain prediction in typhoon over the south china sea with statistic-method. *Scientia Meteorological Sinica.* **16**, 173–177. Available from: http://www.jms1980.com/ch/reader/view_abstract.aspx? file_no=19960225&flag=1.

Liu, Y., Guo, D. M., Yao, J., Qu, L. W. & Li, M. 2015 Applications of ingredients-based forecasting methodology to refine rainstorm forecast. *Journal of Arid Meteorology* **33**, 514–520. https://doi.10.11755/j.issn.1006-7639(2015)-03-0514.

Mosavi, A., Ozturk, P. & Chau, K. W. 2018 Flood prediction using machine learning models: literature review. *Water* **10** (11), 1536. https://doi.org/10.3390/w10111536.

Murnane, R. J. & Elsner, J. B. 2012 Maximum wind speeds and US hurricane losses. *Geophysical Research Letters* **39** (16), 1–5. Available from: http://www.docin.com/p-1423983645.html.

Prahl, B. F., Rybski, D., Kropp, J. P., Burghoff, O. & Held, H. 2012 Applying stochastic small-scale damage functions to German winter storms. *Geophysical Research Letters* **39** (6), 1–6. doi:10.1029/2012GL050961.

Reynolds, D. 2003 Value-added quantitative precipitation forecasts: how valuable is the forecaster? *Bulletin of the American Meteorological Society* **84**, 876–878. https:// doi.10. 1175/BAMS-84-7-876.

Tao, S.-Y., Zhao, S.-X., Zhou, X.-P., Ji, L. R., Sun, S. Q., Gao, S. T. & Zhang, Q. Y. 2003 The research progress of the synoptic meteorology and synoptic forecast. *Chinese Journal of Atmospheric Sciences* **27**, 451–467. https://doi.10.3878/j.issn. 1006-9895.2003.04.03.

Tian, R., Wen, X. & Tian, C. T. 2002 Characteristics and comparison of fuzzy systems and neural networks. *Computer*

Measurement and Control **10** (2), 71–73. https:// doi. 10. 3969/j.issn.1671-4598.2002.02.001.

Tian, X. Y., Huo, Y., Dong, Q., Jing, Y. L. & Jiao, Z. F. 2008 The discriminate prediction on rainstorm using the physical amounts of T 213 and the principle of similarity. *Scientia Meteorological Sinica.* **28**, 456–461. https://doi.10.3969/j.issn. 1009-0827.2008.04.018.

Wang, S. T. 1998 *Neuro-fuzzy System and its Application.* Beijing University of Aeronautics and Astronautics Press, Available from: http://www.doc88.com/p-53076128770.html.

Wang, Y. & Yan, Z. H. 2007 Effect of different verification schemes on precipitation verification and assessment conclusion. *Meteorological Monthly* **33** (12), 53–61. https:// doi.10.3969/j.issn.1000-0526.2007.12.008.

Wu, M. G., Jiang, C. Y., Zhang, X. H. & Lai, R. Q. 2012 Application of BP neural network using cross-entropy to 96 hours forecast of heavy precipitation event in northern Fujian province. *Journal of Nanjing University of Information Science and Technology: Natural Science Edition* **4**, 220–225. https://doi.10.13878/j.cnki.jnuist.2012.03.003.

Wu, Q. S., Han, M., Liu, M. & Chen, F. J. 2017 A comparison of optimal-score-based correction algorithms of model precipitation prediction. *Journal of Applied Meteorological Science* **28** (3), 306–317. https:// doi.10.11898/1001-7313. 20170305.

Xi, L. H. 1987 The application and development of fuzzy mathematics the statistical interpretation of the NWP products. *Meteorological Monthly* **13** (12), 3–7. https:// doi. 10.7519/j.issn.1000-0526.1987.12.001.

Yan, M. L., Wang, M., Yu, B. & Fan, G. Q. 2008 A heavy rain fall forecast method based on fuzzy cluster typing by using application and interpretation of NWP. *Journal of the Meteorological Science* **28**, 581–585. https://doi.10.3969/j. issn.1009-0827.2008.05.020.

Zeng, X. J. & Singh, M. G. 1996 Approximation accuracy analysis of fuzzy systems as function approximators. *IEEE Trans. on Fuzzy Systems* **4** (1), 44–63. https:// doi. 10.1109/91.481844.

Zhai, A. R. & Jiang, J. H. 2014 Dependence of US hurricane economic loss on maximum wind speed and storm size. *Environmental Research Letters* **9** (6). article 064019. doi:10. 1088/1748-9326/9/6/064019.

Zhan, H., Liu, Y. P., Li, J. H., Luan, J. P. & Rong, Y. 2009 Application and study of forewarning model for tunnel construction safety based on fuzzy theory. *China Safety Science Journal* **19** (4), 5–10. https:// doi.10.3969/j.issn.1003-3033.2009.04.001.

Zhang, D. S., Shao, M. X., Mu, Q. Z., Zhang, B. R., Liu, M. H. & Liu, M. 2006 A short-range forecast method of heavy rainfall in Miyun reservoir basin. *Meteorological Monthly* **32**, 61–66. https:// doi.10.3969/j.issn.1000-0526.2006.11.010.

Zhang, W. Q., Sun, M., An, W. & Ma, Y. F. 2009 Study on the discrimination of water inrush from deep-well floor based on fuzzy neural network. *China Safety Science Journal* **19** (12), 61–65. https:// doi.10.3969/j.issn.1003-3033.2009.12.010.

Zhang, X. L., Tao, S. Y. & Sun, J. H. 2010 Ingredients-based heavy rainfall forecasting. *Chinese Journal of Atmospheric Sciences* **34**, 754–766. doi: 10.3878/j.issn.1006-9895.2010.04.08.

Zhao, S. X. & Sun, J. H. 2013 Study on mechanism and prediction of disastrous weathers during recent years. *Chinese Journal of Atmospheric Sciences* **37**, 297–312. https:// doi.10.3878/j. issn.1006-9895.2012.12317.

Zhao, S. R., Zhao, C. G. & Zhao, M. X. 2009 Regression estimate of event possibility and precipitation categorical forecast. *Journal of Applied Meteorological Science* **20**, 521–519. https://doi.10.3969/j.issn.1001-7313.2009.05.002.

Zhong, Y., Yu, H., Teng, W. P. & Chen, P. Y. 2009 A dynamic similitude scheme for tropical cyclone quantitative precipitation forecast. *Journal of Applied Meteorological Science* **20**, 17–27. https://doi.10.3969/j.issn.1001-7313.2009.01.003.