Official Journal of the World Water Council
Water Policy

Water Policy Vol 23 No 4, 1075 doi: 10.2166/wp.2021.247

Research on stage–discharge relationship model based on information entropy

Lin Hao[a,b,c], Jiang Zhu[a,b,c,*], Liu Boxiang[a,b,c] and Chen Ying[d]
[a] Key Laboratory of Fluid and Power Machinery, Ministry of Education, Xihua University, Chengdu 610039, China
[b] School of Energy and Power Engineering, Xihua University, Chengdu 610039, China
[c] Key Laboratory of Fluid Machinery and Engineering, Sichuan Province, Xihua University, Chengdu 610039, China
[d] Shenzhen Deeproute.ai Co. Ltd, Shenzhen 518048, China
*Corresponding author. E-mail: 397470723@qq.com

LH, 0000-0002-1508-8055

ABSTRACT

In order to improve the estimation accuracy of stage–discharge relationship model, the back propagation neural network optimized through the genetic algorithm (GA-BP) based on information entropy was proposed. Firstly, the information entropy and hierarchical clustering were used to quickly cluster the hydrological sample data and get the optimal number of clusters. Secondly, the k-nearest neighbor algorithm was used to divide the new stage data into the most appropriate clustering categories. Finally, the river daily discharge was estimated. Some measured data collected from a hydrological station were used to test the model, and the simulation results showed that the method proposed by this paper can get higher estimation accuracy than the classical analytical model, BP neural network algorithm and GA-BP neural network algorithm, which provided a new effective method for parameter estimation of the stage–discharge relationship model.

Key words: GA-BP, Hierarchical clustering, Information entropy, k-Nearest neighbor, Neural network, Stage–discharge relationship

HIGHLIGHTS

- A GA-BP algorithm based on information entropy is proposed.
- Information entropy and hierarchical clustering were used to obtain the optimal number of clustering.
- Through the model comparison, the GA-BP model based on information entropy has high accuracy.
- This paper presented a new method for flow estimation.

INTRODUCTION

The stage–discharge relationship model is a curve describing the relationship between the water level of the basic section at the hydrological measuring station and the flow through the section. It plays an important role in compiling hydrological and water resources data, hydrological prediction, engineering design and construction (Maghrebi *et al.*, 2016). The current measuring equipment for river flow is not only of great error but also of great cost (Nezamkhiavy & Nezamkhiavy, 2014). Therefore, the most effective and common method is to establish a stage–discharge curve and obtain flow data by converting the average water level, which highlights the importance of the stage–discharge curve (Roushangar & Alizadeh, 2019).

Researchers at home and abroad have put forward a large number of study protocols on the research method of the stage–discharge relationships. Jain & Chalisgaonkar (2000) took the lead in applying a three-layer

feed-forward neural network (ANN) to the modeling of river flow. The results show that compared with the traditional curve fitting method, the ANN has obvious advantages and can estimate the discharge well. Lohani *et al.* (2006) applied Takagi-Sugeno (TS) to the stage–discharge curve and compared it with the traditional least square method and artificial neural network (ANN). The results show that TS is superior to the traditional modeling method based on ANN. Wolfs & Willems (2014) used ANN and M5 model tree to train the hydrological data and compared it with the traditional stage–discharge curve. The results show that the ANN can accurately simulate the calibration data, but when a small amount of data is used for training, there will be over-fitting situation. The architecture of M5 model tree is easier to interpret than ANN and can obtain higher fitting accuracy. Birbal *et al.* (2021) proposed a Gene Expression Programming (GEP) as an extension of Genetic Programming (GP) and applied it to stage–discharge curves. This method was compared with the traditional stage–discharge relationship curves (SRC) and regression methods, and the results show that the performance of GEP model is significantly better than that of GP model and traditional model. Kashani *et al.* (2015) compared the performance of ANN, adaptive neuro-fuzzy inference system (ANFIS), gene expression program design and traditional conventional methods (water level-flow relationship curve and regression method) using the water level and flow data in the Kizlemak River, Turkey. The results show that the related coefficient (r), root mean square error (RMSE) and mean absolute error (MAE) of the machine learning method have high accuracy. In addition, the ANFIS model has the best performance of all methods. Alizadeh *et al.* (2021) used the hybrid preprocessing method of empirical mode decomposition (EEMD), wavelet transform (WT), mutual information (MI) and support vector machine to predict the river flow. The results show that the proposed WT-EEMD-MI method can effectively improve the prediction accuracy of river flow.

Due to the influence of natural conditions and human activities, the process of water change becomes more and more complex. Obviously, the classical analytical model of stage-discharge cannot describe the dynamic relationship between stage and discharge well (Petersen-Øverleir, 2006). In recent years, the rapid development of automatic control technology, computer technology and image display technology has provided new methods and means for the establishment of stage–discharge model and the study of hydrological parameter prediction. Hence, in order to improve the accuracy of stage–discharge curve, a back propagation neural network optimized through the genetic algorithm (GA-BP) based on information entropy is proposed in this paper. This method is a non-parametric learning method based on machine learning. It does not have to form a clear hypothesis to define the complete objective function on the entire sample space, and it can form a different local approximation of the objective function for each query sample. This method first uses information entropy and hierarchical clustering (HAC) to set up a non-analytical relationship model between stage and discharge samples, then quickly clusters hydrological data samples and obtains the optimal number of clusters; then uses K-nearest neighbor (KNN) method to classify the new water level data into the most appropriate cluster category; finally, the daily flow of the river is estimated by using the newly established relationship model. To verify the effectiveness of the proposed method, the classical analytical model of stage–discharge, the BP neural network algorithm and the GA-BP neural network algorithm are compared and tested with the proposed method.

STAGE–DISCHARGE RELATIONSHIP MODEL

Classical analytical model

The stage–discharge relationship refers to the empirical relationship between the discharge Q of a basic section and the corresponding water level h. Different researchers have different opinions on the expression of stage–discharge relationship curve, among which there are two common forms: polynomial model and power-law model.

The expression of the polynomial model is as follows:

$$Q = a_0 + a_1 h_1 + a_2 h_2 + \cdots + a_m h_m \tag{1}$$

where Q is the flow through the basic section (m^3/s); $a_0, a_1, \cdots, a_m$ are the undetermined coefficients, h_i is the water level (m).

The power–law relationship is as follows:

$$Q = kh^b \tag{2}$$

where Q is the flow through the section (m^3/s); k is the constant coefficient to be estimated; h is the average water level (m); b is the index.

a_i, k, b in the above formulas can generally be obtained by the linear regression, the least squares and other methods. In this paper, the least square (LS) is used to calculate the analytical model.

The least square

LS is a tradition mathematical optimization design method (Ruiz *et al*., 1996). LS is used to obtain the unknown data conveniently, and the sum of the square of the error between the estimated data and the measured data is minimized. LS is widely used in mathematical statistics, mathematical optimization and prediction estimation. Based on the theory of error, LS has high reliability and solves the problem of how to obtain credible reliability from a set of measured data. In this paper, the LS is used to calculate the polynomial form of stage-discharge curve mentioned above, and the estimated discharge will be obtained through the stage. The main steps are as follows:

Step 1. Determine the number n of polynomial fitness.
Step 2. Determine the coefficients of polynomials according to the principle of LS and minimize the sum of squares of errors when the function is m polynomials.
Step 3. Fit discharge.

The structures of the above classical analytical model are simple and easy to solve. But, the value of them is greatly affected by many complex factors, such as flood fluctuations, variable backwater and sedimentation. Therefore, the traditional stage–discharge relationship model established by the mathematical method cannot take these complex factors into consideration. So, the above empirical relationships cannot accurately reflect the true stage–discharge relationship in the river (Ajmera & Goyal, 2012).

Based on the above analysis, a GA-BP network algorithm based on information entropy is proposed in this paper to establish a non-parametric relationship between stage and discharge, which aims to more accurately and truly reflect the hydrological situation in the river, and provide reliable data for water conservancy planning and design, as well as hydrological prediction. To verify the effectiveness of the proposed algorithm, BP network algorithm and GA-BP network algorithm are compared with the novel method. The details are as follows.

BP NEURAL NETWORK ALGORITHM

Back propagation neural network was proposed by Rumelhart (1986), which is a multi-layer feed-forward neural network trained according to the error back propagation algorithm. At present, BP neural network has made great contributions to computer science, information science and mathematical statistics and has been the most widely used neural network so far. The basic principle is showed by Figure 1: In forward propagation, the sample is passed from the input layer and processed by the hidden layer to the output layer. If the actual

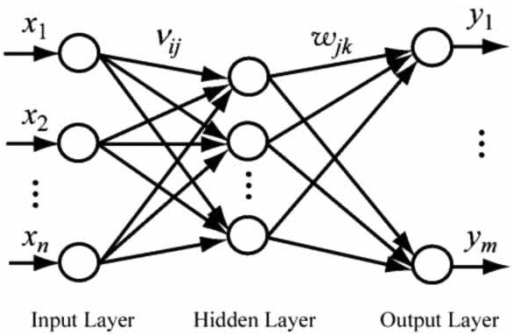

Fig. 1. | The structure diagram of three layers of BP neural network.

output of the output layer is inconsistent with the expected output, it enters the back propagation. Error back propagation is to transfer the output error layer by layer through the hidden layer to the input layer in some form and allocates the error to all the units of each layer, so as to obtain the error signal of each layer unit. This kind of signal adjusts the weight of each layer in forward and back propagation repeatedly until the error of network output is reduced to an acceptable degree or the number of learning times is set, among which the process of weight adjustment is the process of network learning (Tang *et al.*, 2020; Wang *et al.*, 2020).

Studies have shown that a three-layer BP neural network model can simulate any complex nonlinear problem, that is, one input layer, one hidden layer and one output layer (Peter, 2019). In this paper, the following parameters are set: input layer node is 1; hidden layer node is 5; output layer node is 1; Select S-type transfer function (logsig) as transfer function; Levenberg–Marquardt calculation method is used in the training function; the number of training cycles is set to 5,000 times, the number of iterations is 500 times, the learning rate is 0.1, and the allowable error is 0.001.

GA-BP NEURAL NETWORK ALGORITHM

BP (back propagation) neural network can learn independently, has strong nonlinear mapping ability and rigorous derivation process, but it has the disadvantages of slow convergence speed and weak generalization ability. To solve this problem, researchers propose to use GA to optimize the BP neural network (Chen *et al.*, 2019). GA can effectively solve the problems of BP neural networks by selecting, crossing and mutating operations to finally obtain the optimal threshold and initial weight of the network (Jan, 2019). The concrete steps are as follows:

1. Establish BP model. The setting of BP neural network is the same as that of BP model mentioned above, that is, the input layer node is 1, the hidden layer node is 5 and the output layer node is 1.
2. Initialize the cluster. According to the number of nodes of the BP neural network in the previous step to encode the chromosome, this paper uses a real number code scheme, where the individual coding length is $1 \times 5 + 5 \times 1 + 1 + 1 = 12$.
3. Determine the fitness function. The reciprocal of the sum of squared errors between the actual output value and the expected output value is taken as the fitness function value F of each individual:

$$F = \frac{1}{\sum_{i=1}^{n} (y_i - x_i)^2} \tag{3}$$

where n is the number of samples; x_i is the actual output value of the i-th node; y_i is the expected output value of the *i-th* node.

4. Random traversal sample method is used for selection.
5. Cross operation. The real number crossing method is used, and the crossover operation of the *k-th* chromosome m_k and the *l-th* chromosome m_l at the j position are as follows:

$$\begin{cases} m_{kj} = m_{kj}(1-n) + m_{lj}n \\ m_{lj} = m_{lj}(1-n) + m_{kj}n \end{cases} \tag{4}$$

where m is a random number and $m \in [0, 1]$.

6. Mutation operation. The *j-th* gene of the *i-th* individual is selected for mutation, the specific operation are as follows:

$$m_{ij} = \begin{cases} m_{ij} = m_{ij} + (m_{ij} - m_{\max}) \times f(g), r > 0.5 \\ m_{ij} = m_{ij} + (m_{\min} - m_{ij}) \times f(g), r < 0.5 \end{cases} \tag{5}$$

where $f(g) = k(1 - g/G_{\max})^2$, and g is the current iterations, and k is a random number. $G_{\max}$ is the maximum number of evolution, and r is a random number and $r \in [0, 1]$; $m_{\min}$ and $m_{\max}$ are the lower bound and the upper bound of m_{ij}.

7. Replace the original chromosome with a new chromosome and calculate the fitness. If the condition is satisfied or the number of iterations is reached, Step (8) will be performed; otherwise, go to Step (3) to continually optimize.
8. Assign optimized weights and thresholds to the BP neural network and train the data until the requirements set by the network are reached.

GA-BP BASED ON INFORMATION ENTROPY

In 1948, the founder Claude Elwood Shannon of information theory introduced the concept of entropy to information theory (Hasan & Rai, 2020). Information entropy is often used as a quantitative indicator of the information content of the system, which can be further used as the goal or parameter selection of system equation optimization Criterion (Capozziello & Luongo, 2017). Shannon proved mathematically that the uncertainty function of random variables satisfying monotonicity, nonnegativity and accumulation has a unique form:

$$H(x) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{6}$$

where x_i means various random events; $P(x_i)$ represents the probability of random event x_i; $H(x)$ is entropy.

If the data are divided into a set of k clusters named $C = \{C_1, C_2, \cdots, C_k\}$, the information entropy can be expressed as:

$$H(x) = \sum_{i=1}^{k} \sum_{x_{ij} \in C_j} P(x_{ij}) \log_2 (P(x_{ij})) \tag{7}$$

where $P(x_{ij})$ is the probability of event i in set C_j.

It is worth explaining that with the increase of the number of clusters, the amount of data in each class decreases, the probability that each datum belongs to one class increases, and the information entropy of the whole class becomes larger. In the process of increasing the number of classes, class division is carried out in the order of disorder → order → disorder. The initial disorder is because the clustering is too general to see the characteristics of the data set, and the final disorder is that the clustering is too fine and lacks the overall understanding. Therefore, the transition value of the total information entropy of the data set can be used to determine the optimal number of clusters (Mahata & Sing, 2020). We define the state of k class as the k state of the dataset, l_k as the information entropy of the k state, $l_k - l_{k-1}$ as the information entropy jump value. The information entropy transition value is the difference between the jump from the $k-1$ to the k state entropy value and the jump from the k to the $k+1$ state entropy value, that is $|(l_{k+1} - l_k) - (l_k - l_{k-1})|$. When the transition value of information entropy reaches the minimum value, it shows that the entropy jump from the k-th state to the $(k+1)$-th state is the smallest among all the jumps compared with the k-th to $(k-1)$-th state, namely the minimum increase in uncertainty across the data set (Su *et al.*, 2010). Now there is no need to increase the number of clusters, which eventually identified as k clusters.

Clustering belongs to unsupervised learning, that is, the marking information of training samples is unknown, and the goal is to expose the attributes, structure and information of training samples and to provide the basis for further data mining. The essence of clustering process in machine learning is an optimization process, that is, the objective function of the system reaches a minimum value through a fast operation. Clustering mainly includes density-based method, model-based method, partition-based method and hierarchy-based method, in which HAC is to create a hierarchical nested clustering tree by calculating the similarity between different categories of data points (Bonetto & Latzko, 2021). It calculates the distance between each category of data points and all data points to determine their similarity. The smaller the distance, the higher the similarity, and combines the nearest two data points or categories to generate the cluster tree.

A GA-BP algorithm based on information entropy is proposed in this paper in order to reduce the estimation error of stage–discharge curve and improve the estimation accuracy of discharge. The implementation process is as follows.

Firstly, the number of clusters is determined. In order to obtain the optimal clustering scheme, the HAC is used to divide the hydrological data and the information entropy is introduced to judge the optimal number of clusters. The details are as follows:

Step 1, determine the initial clustering number range $[C_{\min}, C_{\max}]$, generally take $C_{\min} = 2$, $C_{\max} = \sqrt{n}$, where n is the number of samples.
Step 2, select k cluster centers from each n training object (initial value is set to 2);
Step 3, each sample is classified into one class (initializing data), the distance between each two classes is calculated, that is, the similarity between each two samples is calculated;
Step 4, according to the principle of minimum variance, select the category that meets the distance requirement, and complete the inter-class merger;
Step 5, recalculate the distance between the newly generated classes and the old classes (similarity) in Step 2;
Step 6, repeat Step 2 and Step 3 until all objects finally merge to form k clusters.
Step 7, calculate information entropy and the transition value of information entropy. When $k < C_{\max}$, returned to Step 1; otherwise determine the optimal number of clusters according to the entropy transition value of the data set.

After that, this paper uses the optimal clustering number to cluster and uses the KNN algorithm to divide the new data (for testing) into classes divided by HAC rules (Sharma & Shamkuwar, 2019). KNN belongs to supervised learning in machine learning. This method determines the category of samples to be divided according to

the category of one or more nearest samples, that is, when a new water level data is given, KNN searches one or more samples closest to the new water level data in the water level samples that have been clustered. The Euclidean distance is used between the sample x_i and x_j:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \tag{8}$$

where $< a_1(x), a_2(x), \cdots, a_n(x) >$ is the proper vector for the new sample $\boldsymbol{x}$. x_i is the stage and discharge on day i.

Finally, when each new sample is classified, the river flow is estimated using the GA-BP algorithm in the previous section.

METHOD TEST

Experimental data

In this paper, the algorithm is tested with the measured data from 2007 to 2010 in some basin of Minjiang River. Minjiang River originates from Langjialing, Songpan County, the south foot of Minshan Mountain, and it is the largest tributary of Yangtze River Runoff input. From north to south, it flows through the west of Sichuan Basin, converges with Jinsha River in Yibin City and converges into Yangtze River. The drainage basin above Dujiangyan is called the upper reaches of Minjiang River, which is located between Sichuan Basin and Qinghai Tibet Plateau, and most of them are alpine canyon areas. It is located between 30°45′–33°10′ N and 102°35′–130°57′ E. The total length of the main stream of Minjiang River is 337 km, and the drainage area is about 22,612 km^2. The geographical location of the study area is shown in Figure 2.

The paper collected annual and monthly data on runoff, precipitation and temperature in the upper reaches of the Minjiang River from 2007 to 2010. The distributions of meteorological and hydrological stations are shown in Figure 3. The data from 2007 to 2009 are used for training the relationships of stage and discharge, and the data collected in 2010 are used to test the trained model. Part of the measured data is shown in Table 1.

Parameter estimation for relationship model of stage–discharge

When the GA-BP algorithm based on information entropy is used to test, the clustering center of the method is judged and determined by the information entropy. The results of clustering are shown in Figure 4. Figure 4(a) is the distribution of the measured data, and Figure 4(b) is the distribution of the information entropy transition value.

It can be seen from Figure 4 that when the transition state is 3, the transition value of information entropy from (3 classes → 4 classes) to (4 classes → 5 classes) reaches the lowest, and then it starts to increase again, which shows that the number of 4 clusters can make the overall information amount the largest, which is the optimal number of clusters. So, the number of clusters is 4 for the next operation in this paper. The clustering results are shown in Figure 5.

In this paper, absolute mean error and average absolute percentage error are used as the criteria to test the performance of each method.

Absolute mean error:

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \tag{9}$$

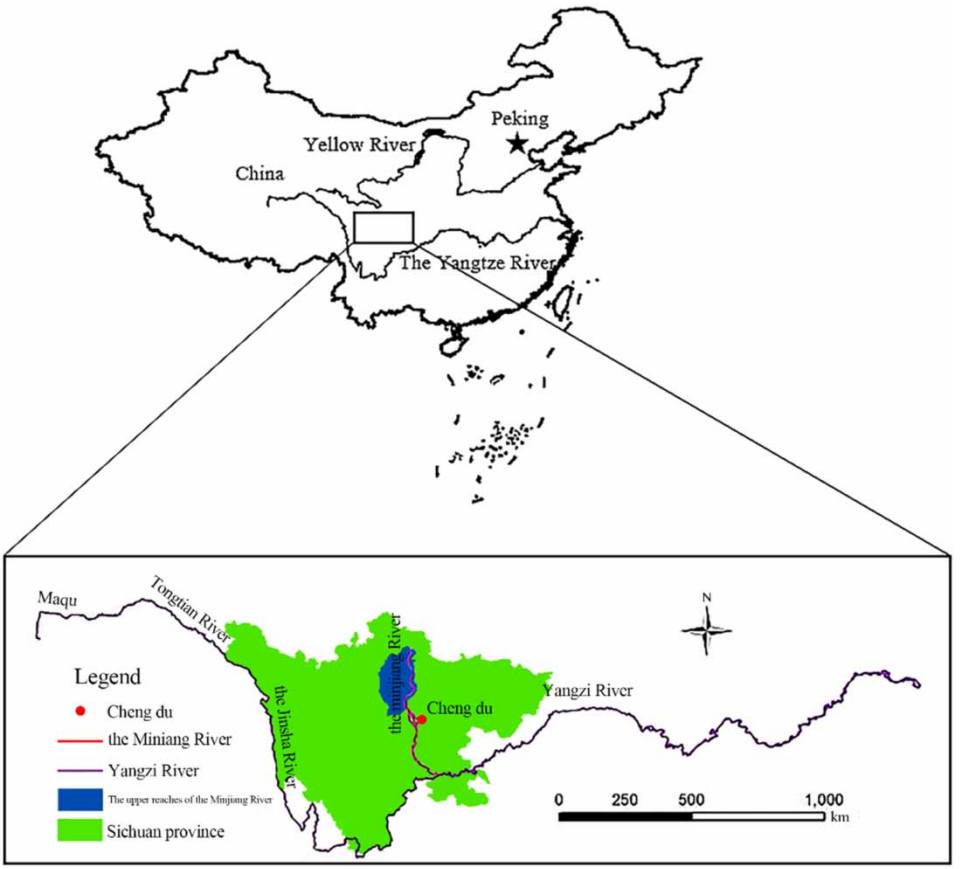

Fig. 2. | Geographical location of the study area.

Mean absolute percentage error:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - y_i|}{y_i} \times 100\ \% \tag{10}$$

where x_i is the estimated river flow on the i-th day, y_i is the actual river flow on the i-th day and n is the number of test samples.

The results of MAE and MAPE are shown in Figure 6.

The analysis of Figure 6 shows that the MAE and MAPE of the proposed method are lower than those of the other three methods. The proposed method can effectively reduce the error of parameter estimation and improve the precision of model estimation.

Based on the above calculations, the results are counted as in Table 2.

By comparing Table 2, it can be seen that MAE and MAPE of the method proposed by this paper (GA-BP model based on information entropy) are superior to classical analytical model, BP model and GA-BP model. The MAE and MAPE of the method proposed by this paper are 25.15 and 16.38% less than those of the classical analytical model, 24.21 and 18.05% less than those of the BP, 24.53 and 15.43% less than those of the GA-BP.

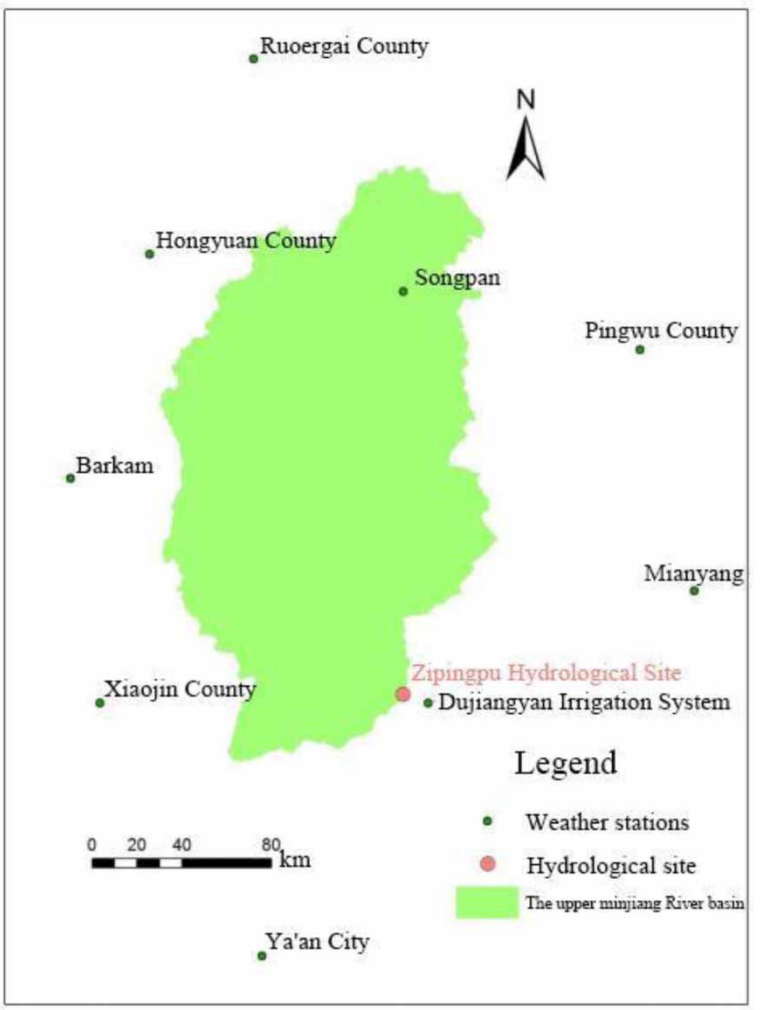

Fig. 3. | Distributions of meteorological and hydrological stations.

Figure 7 shows the estimation results of River daily discharge by several methods.

Figure 7 shows the estimated results of several methods for daily river flow. We can see from Figure 7 that the classical analytical model, the BP model and the GA-BP model basically agree with the estimation results of the river flow, but the BP model has a large error in estimating the large flow rate, so it cannot estimate the large flow rate better. The GA-BP model based on information entropy proposed in this paper has higher estimation accuracy and can better estimate the flow. By comparing Figure 7, we can see that the proposed method can effectively improve the phenomenon of large deviation between the measured flow rate and the estimated daily flow rate under normal climatic conditions.

Comparing the four methods mentioned above, the method proposed in this paper can obtain a smaller estimation error than the other three methods. Combined with the daily flow distribution diagram shown in Figure 4(a), compared with Figure 6, it can be seen that in the dry season from January to mid-April, the daily

Table 1. | Some experimental data.

Day	Average water level (m)	Average flow (m^3/s)	Average sediment concentration (g/m^3)
1	530.12	629	53
2	530.02	586	54
3	529.75	472	55.7
4	529.82	501	57.7
5	529.82	497	57.9
6	529.79	486	52.6
7	529.85	511	46.9
8	530.02	581	41.6
9	530.22	671	35.8
10	530.10	616	31.4

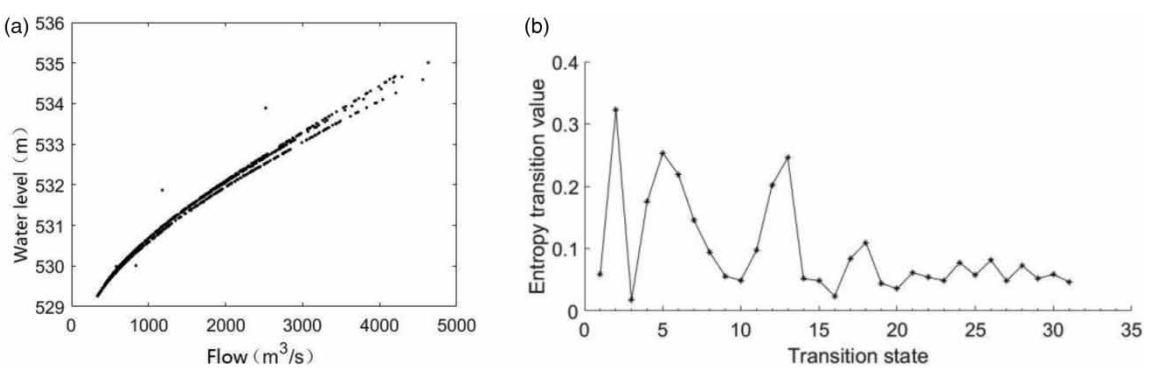

Fig. 4. | Calculation of clusters number. (a) Observed data. (b) Entropy transition value.

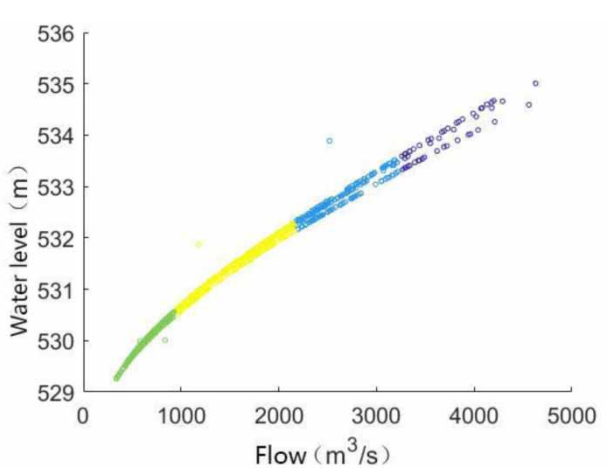

Fig. 5. | Cluster results.

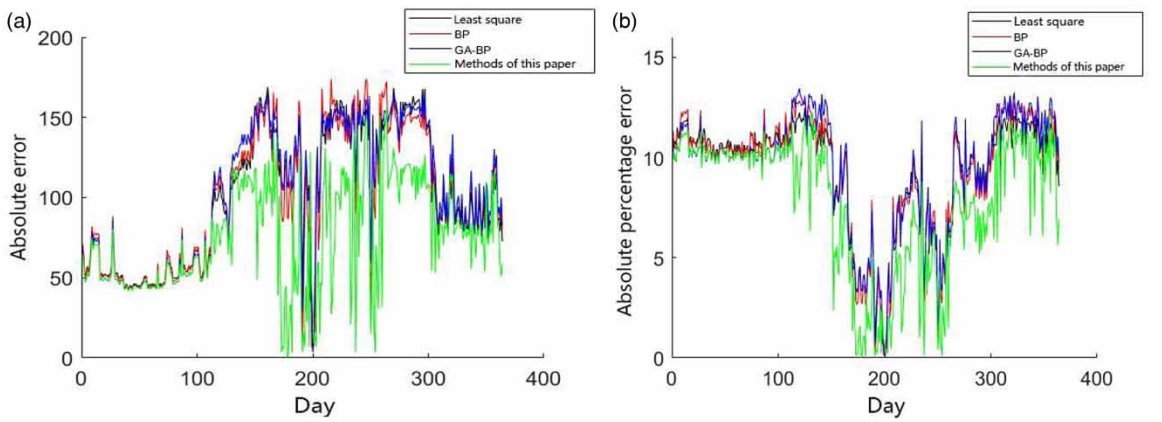

Fig. 6. | Model error comparison. (a) MAE of four models. (b) MAPE of four models.

Table 2. | Error comparison of algorithms.

Model	*MAE*	*MAPE* (%)
Classical analytic model	100.62	9.28
BP	104.64	9.47
GA-BP	99.78	9.17
Method proposed by this paper	75.31	7.76

flow of the river is relatively small, the estimated daily flow is basically the same as the actual observed value, and the accuracy of the flow estimation by the four schemes is not much different; from the middle of April to September, the daily flow of river flow in this period is relatively large, and there will be many unpredictable factors that affect the flow of the day, such as flood, rainfall, weather changes, which cause the measured daily flow to have many different values and large deviations under the same water level. Obviously, the traditional methods of estimating river flow will cause a great many errors, and during this time, the scheme proposed in this paper can significantly improve the accuracy of river daily flow estimation; from October to December, with the passing of the rainy season, the river flow gradually decreases, and the uncontrollable influencing factors decrease. The accuracy of the classical analytical model, BP, and GA-BP algorithm to estimate the daily flow of the river increases. The GA-BP algorithm can further improve the estimation accuracy of river daily flow. Through the analysis of Figures 6 and 7, it can be seen that this method can get more accurate estimates than the analytic model, BP model and GA-BP model, and effectively improve the phenomenon that there is a large deviation between the estimated value and the measured value of River daily flow under normal climate conditions. This is because the river hydrological data are used as training samples for clustering, and then KNN method is used to cluster the flow related data and the new river data samples are classified into appropriate classes, which can avoid the interference of other irrelevant information, thus improving the efficiency and accuracy of River daily flow data estimation. In the case of no extreme climate, the method proposed in this paper has a strong practical significance to capture the change of river hydrology more accurately.

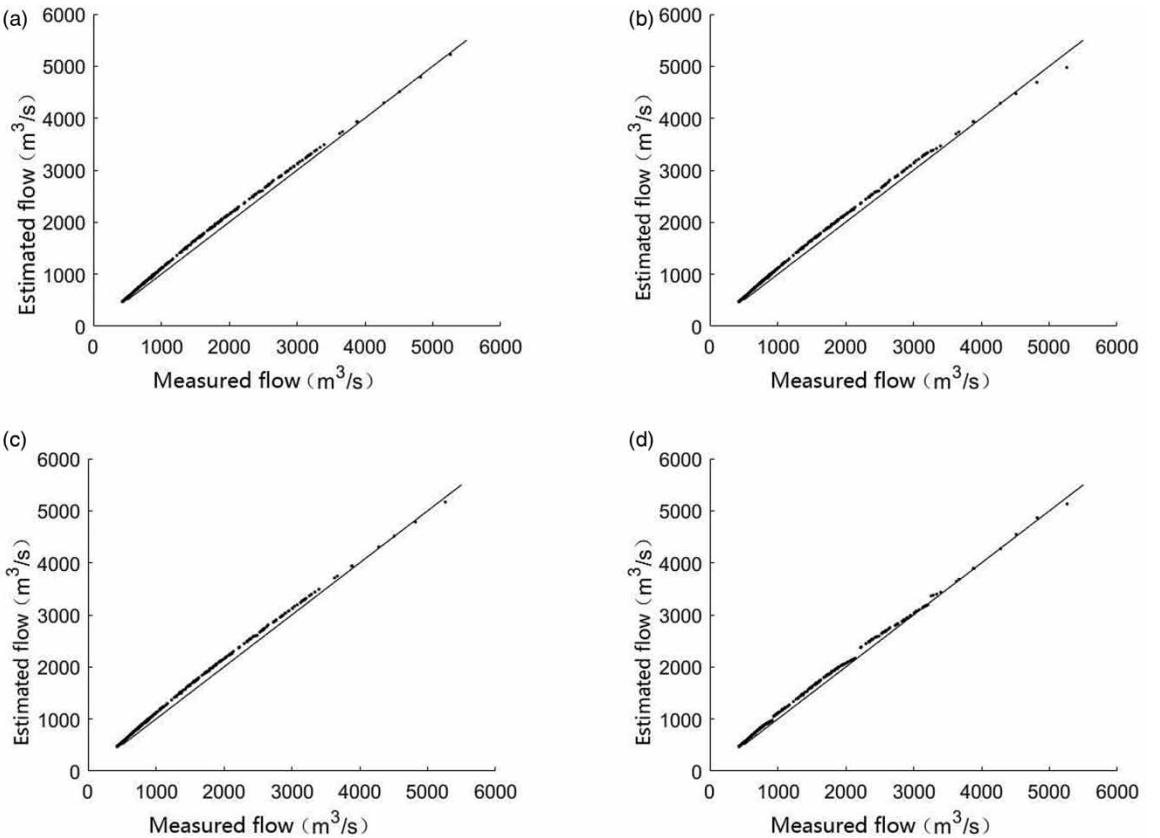

Fig. 7. | Flow estimation results. (a) Estimated results of classical analytical model. (b) Estimated results of BP. (c) Estimated results of GA-BP. (d) Estimated results of GA-BP based on information entropy.

CONCLUSION

Most of the classical SRCs are based on empirical regression, which cannot be well applied to the study of flow characteristics of complex rivers. In this paper, the GA-BP model based on information entropy is proposed to estimate the parameters of the river water level and flow curve in the Minjiang River Basin, and the estimation results are compared with those of the classical analytical model, BP model and GA-BP model. The method is verified by the measured data from the hydrological station in the Minjiang River Basin, and the simulation results show that:

1. Compared with the classical analytical model, the model based on the neural network can estimate the runoff flow better and has a higher estimation accuracy. Obviously, the model based on neural network can better capture the change characteristics of dynamic flow.
2. The GA-BP model based on information entropy proposed in this paper can control the average absolute error of flow estimation below 80, which is reduced by 25.15, 24.21 and 24.53% compared with the classical analytical model, BP model and GA-BP model. It is obvious that this method can obtain higher estimation accuracy than the classical analytical model, BP model and GA-BP model.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

ACKNOWLEDGEMENTS

This work is supported by the Key Scientific Research Project of Xihua University (Grant No. Zl120413), the Key Laboratory of Automotive Engineering of Sichuan Province (Grant No. szjj2017-014), and the Innovation Fund of Postgraduate, Xihua University (ycjj2019043).

REFERENCES

Ajmera, T. K. & Goyal, M. K. (2012). Development of stage-discharge rating curve using model tree and neural networks: an application to Peachtree Creek in Atlanta. *Expert. Syst. Appl. 39*, 5702–5710.

Alizadeh, F., Gharamaleki, A. F. & Jalilzadeh, R. (2021). A two-stage multiple-point conceptual model to predict river stage-discharge process using machine learning approaches. *J. Water Clim. Chang. 12*, 278–295.

Birbal, P., Azamathulla, H., Leon, L., Kumar, V. & Hosein, J. (2021) Predictive modelling of the stage-discharge relationship using Gene-Expression Programming. *Water Supply* ws2021111. https://doi.org/10.2166/ws.2021.111.

Bonetto, R. & Latzko, V. (2021). Chapter 8. Machine learning. In: *Computing in Communication Networks*. Fitzek, F. H. P., Granelli, F. & Seeling, P. (eds). Academic Press, pp. 135–167.

Capozziello, S. & Luongo, O. (2017). Information entropy and dark energy evolution. *International Journal of Modern Physics D 27* (03), 1850029.

Chen, N., Xiong, C., Du, W., Wang, C., Lin, X. & Chen, Z. (2019). An improved genetic algorithm coupling a back-propagation neural network model (IGA-BPNN) for water-level predictions. *Water (Switzerland) 11* (9), 1795.

Hasan, M. S. U. & Rai, A. K. (2020). Groundwater quality assessment in the Lower Ganga Basin using entropy information theory and GIS. *Journal of Cleaner Production 274*, 123077.

Jain, S. K. & Chalisgaonkar, D. (2000). Setting up stage-discharge relations using ANN. *J. Hydrol. Eng. 5*, 428–433.

Jan, J. (2019). A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chemical Science 12*, 3567–3572.

Kashani, M. H., Daneshfaraz, R., Ghorbani, M. A., Najafi, M. R. & Kisi, O. (2015). Comparison of different methods for developing a stage -discharge curve of the Kizilirmak River. *Journal of Flood Risk Management 8* (1), 71–86.

Lohani, A. K., Goel, N. K. & Bhatia, K. K. S. (2006). Takagi-Sugeno fuzzy inference system for modeling stage-discharge relationship. *J. Hydrol. 331*, 146–160.

Maghrebi, M. F., Ahmadi, A., Attari, M. & Maghrebi, R. F. (2016). New method for estimation of stage-discharge curves in natural rivers. *Flow Meas Instrum. 52*, 67–76.

Mahata, N. & Sing, J. K. (2020). A New fuzzy clustering algorithm by incorporating constrained class uncertainty-based entropy for brain MR image segmentation. *Computer Vision and Image Processing 1147*, 301–310.

Nezamkhiavy, K. & Nezamkhiavy, S. (2014). Estimate stage-discharge relation for rivers using artificial neural networks – case study: dost Bayglu hydrometry station over Qara Su River. *Acad. Journals 6*(9), 232–238.

Peter, H. (2019). Artificial intelligence and machine learning. *Handchirurgie, Mikrochirurgie, plastische Chirurgie : Organ der Deutschsprachigen Arbeitsgemeinschaft fur Handchirurgie: Organ der Deutschsprachigen Arbeitsgemeinschaft fur Mikrochirurgie der Peripheren Nerven und Gefasse: Organ der V 51* (1), 62–67.

Petersen-Øverleir, A. (2006). Modelling stage-discharge relationships affected by hysteresis using the Jones formula and nonlinear regression. *Hydrol. Sci. J. 51*, 365–388.

Roushangar, K. & Alizadeh, F. (2019). Scenario-based prediction of short-term river stage-discharge process using wavelet-EEMD-based relevance vector machine. *J. Hydroinformatics 21*, 56–76.

Ruiz, L. M., Valenciano, F. & Zarzuelo, J. M. (1996). The least square prenucleolus and the least square nucleolus. Two values for TU games based on the excess vector. *Int. J. Game Theory 25*, 113–134.

Rummelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. *Nature 323*, 318–362.

Sharma, N. & Shamkuwar, M. (2019). Big data analysis in cloud and machine learning. In: *Big Data Processing Using Spark in Cloud*. Mittal, M., Balas, V. E., Goyal, L. M. & Kumar, R. (eds). Springer Singapore, pp. 51–85.

Su, X., Wang, X. & Wang, Z. (2010). New fuzzy clustering algorithm based on information entropy. *Journal of Tianjin University of Technology 26* (05), 57–60 (China).

Tang, Y., Su, J. & Khan, M. A. (2020). Research on sentiment analysis of network forum based on BP neural network. *Mob. Networks Appl. 26*, 174–183.

Wang, Q., Dai, Y., Wu, Y. *et al.*, (2020). Data model for smart electricity meter comprehensive verification based on BP neural network. *J. Phys. Conf. Ser. 1486*, 022028. (9pp).

Wolfs, V. & Willems, P. (2014). Development of discharge-stage curves affected by hysteresis using time varying models, model trees and neural networks. *Environ. Model Softw. 55*, 107–119.

First received 11 November 2020; accepted in revised form 4 June 2021. Available online 30 June 2021