

Effects of Training Data on the Learning Performance of LSTM Network for Runoff Simulation

Anbang PENG

State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering

Yuanyang TIAN (✉ yuanyangt@163.com)

Chongqing Jiaotong University <https://orcid.org/0000-0003-4471-2985>

Wei XU

Chongqing Jiaotong University

Xiaoli ZHANG

North China University of Water Resources and Electric Power

Research Article

Keywords: LSTM, Rainfall runoff, Data amount, Over-fitting

Posted Date: January 27th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1252947/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Effects of Training Data on the Learning Performance of LSTM Network for Runoff Simulation

Anbang PENG¹ · Yuanyang TIAN^{1,2} · Wei XU² · Xiaoli ZHANG³

Abstract With the rapid development of Artificial Intelligence (AI) technology, an increasing number of intelligent algorithms have been used for simulating and forecasting hydrological process, among which the Long Short-Term Memory (LSTM) network is widely studied. The training of artificial intelligence networks often entails a large amount of training data, which contradicts the limitation of hydrological data. In this study, the effect of training data amount on the performance of LSTM network for runoff simulation are evaluated. First, the runoff series of 130 years are randomly generated by K-Nearest Neighbour (KNN) algorithm and SWAT model. The K-Nearest Neighbour (KNN) algorithm is employed for generating the meteorological data series based on the observed data, and the SWAT model is used to obtain the runoff series with the generated meteorological data series. Then, the LSTM models are developed and evaluated, with the 5-year, 10-year, 20-year, 40-year and 80-year data series of rainfall and runoff as training data respectively, and the 50-year data serves as validating data.

Anbang Peng, E-mail: pengfei110ab@126.com

✉ Yuanyang Tian, E-mail: yuanyangt@163.com

¹ State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing, China

² Chongqing Jiaotong University, Chongqing, China

³ School of Water Conservancy, North China University of Water Resources and Electric Power, Zhengzhou, China

The results obtained in Yalong River, Minjiang River and Jialing River show that (1) increasing the training data amount can effectively reduce the over-fittings of LSTM network; (2) increasing the training data amount can also improve the prediction accuracy and stability of LSTM network.

Key words: LSTM; Rainfall runoff; Data amount; Over-fitting

1 Introduction

The reliability of runoff prediction is of great significance to water resource management. But the evident non-linear nature and randomness of runoff process makes it challenging to predict runoff. The models available for runoff prediction fall into three categories: (1) conceptual model, which is simple and easy to use, but gives a less detailed description of runoff generation and flow concentration; (2) physical model, which enjoys a good physical foundation but involves complicated calculation structure and large amount of basic data to build model. (3) data-driven model, which just needs to create the relationship between input and output data, without considering the hydrological mechanism of the basin and having a good simulation. Among them, the data-driven model, e.g., the convergence of the conventional back propagation (BP) neural network is too slow and thus has limited learning ability. With the rapid development of deep learning over the years (Hinton and Salakhutdinov 2006), the learning abilities of the Deep Neural Networks (DNN) are advancing by leaps and bounds, demonstrating super learning abilities in learning knowledge of various industries. However, the traditional DNN network cannot be used for time series modelling.

In order to learn the changing rule of the time series, Recurrent Neural Network (RNN)

came into being. Based on conventional RNN, Long-Short Term Memory (LSTM) Network was first proposed by Schmidhuber and Hochreiter in 1997, which has overcome such problems as gradient explosion and gradient vanishing in RNN. Due to the strong abilities in learning the time series data by digging out the long-term dependence relationship in time series, LSTM has been widely used in many fields, such as, image processing (Jiang and Liu 2021; Lin et al. 2021), transportation (Zhi et al. 2021; Huang et al. 2020), geological disasters relief (Ad et al. 2021), financial economy (Dami and Esterabi 2021) and so on. Moreover, its applicability for hydrological prediction has been demonstrated by many researchers. For example, XU et al. (2020) analyzed the performance of the LSTM network, SWAT model, Xin'anjiang Model, Multiple Regression Model and Back-propagation Neural Network (BPNN), and concluded that LSTM network has better learning abilities except for the problem of over-fitting. YANG et al. (2018) used the LSTM network to simulate the relationship between rainfall and runoff in catchment area with different climatic conditions, concluding that LSTM network is more advantageous than RNN in terms of learning time series data, and that it can deliver higher quality prediction than Xin'anjiang Model. Huang (2020) predicted the runoff of some sub-basins of the Hanjiang River System using the LSTM network, concluded that the LSTM network's super parameters have a big effect on the runoff prediction results and that, if an appropriate super parameter is chosen, Nash-Sutcliffe efficiency coefficient (NSE) can reach 0.93 or above. All the aforementioned researches indicate that the LSTM network has strong capability in simulating runoff. In order to improve the learning ability of LSTM network, some scholars have conducted in-depth research (Yin et al. 2021; Wang et al. 2021). However, there are little studies focusing on the issue of over-fitting and the effect of training data amount on

the performance of LSTM model.

At present, the Deep Learning Network training needs to take big data as the foundation (Hu et al. 2018; Mao et al. 2020), but the hydrological data available often fail to meet the needs. If the network is trained with inadequate hydrological data, its reliability would deserve more concern. Therefore, studying the influence of training data amount on the LSTM and finding the future breakthrough point is the key to the application of deep learning in the field of hydrology.

2 Research methodologies

2.1 Research Design

In order to study the effect of training data amount on the LSTM network performance, a scheme is proposed in this study. Firstly, the K-Nearest Neighbour (KNN) algorithm (Upmanu and Ashish 1996) is used to derive rainfall data with long series. Secondly, the rainfall data is used as the input for SWAT model to simulate runoff data (served as experimental real data). Rainfall and runoff data are divided into training data and verification data. Then, the effect of the training data amount on the LSTM network performance are analyzed by varying the training data amount. Finally, the better parameters choices and data volume are found out, and the learning ability of LSTM network are tested based on this. Research flow chart of this article is shown in Fig. 1.

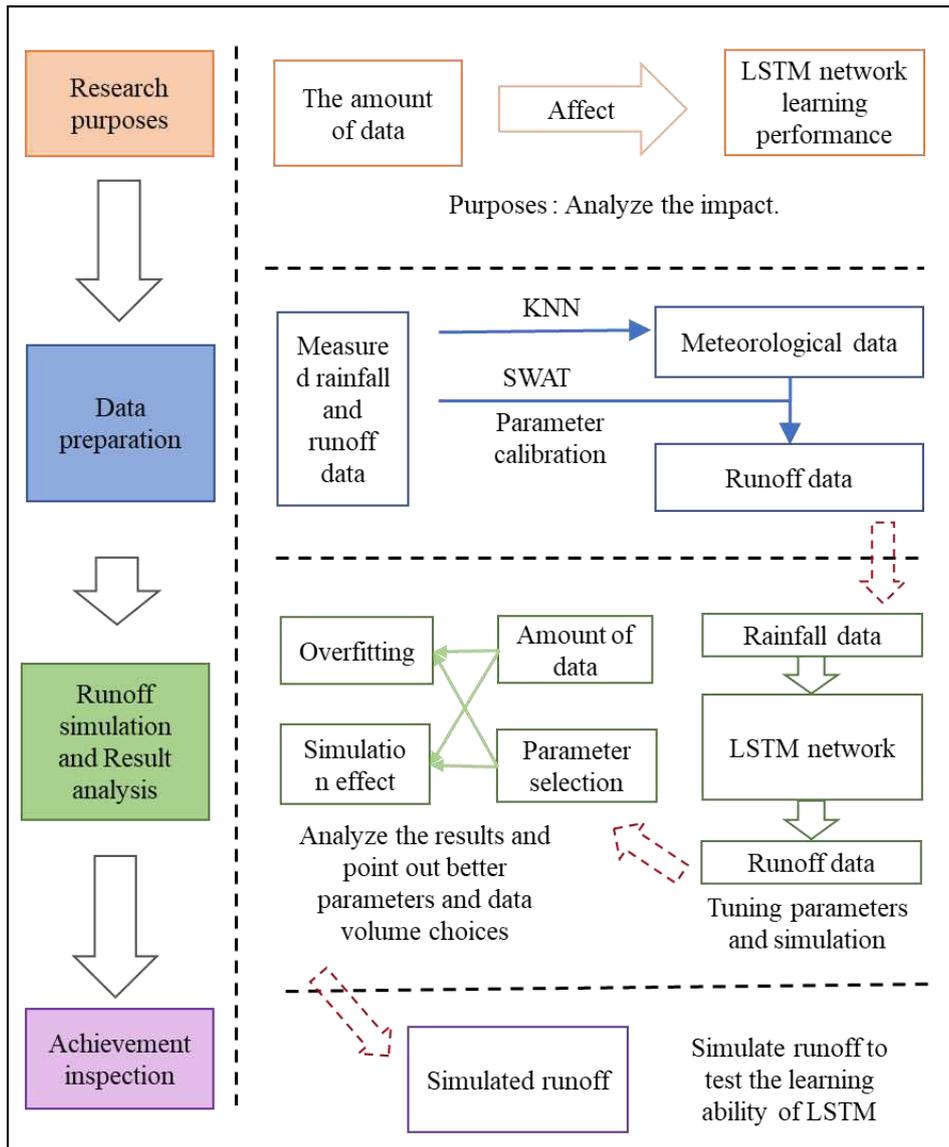


Fig. 1 Research flow chart

2.2 LSTM network

The key parts of the LSTM network are comprised of fully connected layer and LSTM cells. As shown in Fig. 2(a), the LSTM network includes four kinds of layers: (1) input layer, which is to receive input series data; (2) fully connected layer, which acts as a bridge between input layer and the LSTM cell layer to transmit the dimension of the input data to the dimension of LSTM

cell; (3) LSTM cell layer, which provides n cells with different memory capacity; (4) output layer, which is to pass the output of LSTM cells.

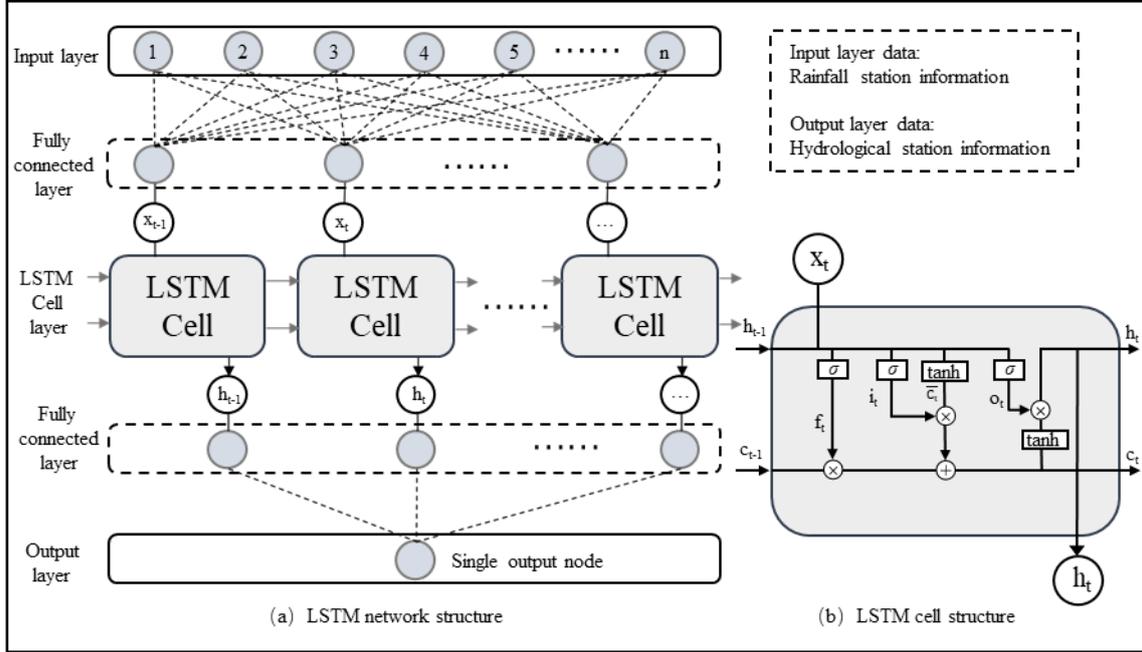


Fig. 2 LSTM network structure and cell structure

LSTM cell structure is shown in Fig. 2 (b). There are two key states in calculation of LSTM cell, that is, Cell State (S_{t-1}) and Hidden State (the output time step $t-1$, $y-1$). In the hidden state, information can be added or deleted from the cell state, which is controlled by “Forget Gate”, “Input Gate” and “Output Gate”. LSTM cell uses all of these gates to control memory progress so as to avoid long-term dependence (Yu et al. 2019). The details of these gates are shown below.

The “Forget Gate” f_t decides what cell states are forgotten, and the forgotten extent is determined by sigmoid network output whose value ranges $[0, 1]$, where 1 represents “Keeping all” and 0 “Forgetting all”. The “Forget Gate” makes it possible to recall past information in current time step, see the Formula (1); “Input Gate” i_t determines the update of cell state,

consisting of sigmoid network and a tanh network. The sigmoid network here determines how much “Hidden State” information is involved in updating, see the Formula (2); the output of the tanh network is one-dimensional vector, which determines how much “Hidden State” information needs updating, see the Formula (3); based on the output from “Forget Gate” and “Input Gate”, the cell state can be updated, see the Formula (4); in the “Output Gate” o_t , the sigmoid network is used to determine which information of the hidden state is output, see the Formula (5); finally, the output of LSTM cells is as shown in Formula (6).

$$f_t = \sigma(W_f \times [y_{t-1}, x_t] + b_f) \quad (1)$$

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

$$\overline{C}_t = \tanh(W_c \times [y_{t-1}, x_t] + b_c) \quad (3)$$

$$S_t = f_t \times S_{t-1} + i_t \times \overline{C}_t \quad (4)$$

$$o_t = \sigma(W_o \times [y_{t-1}, x_t] + b_o) \quad (5)$$

$$A \times y_t = o_t \times \tanh(S_t) \quad (6)$$

Where, σ represents s -network; W_f and b_f represent weight matrix and bias of sigmoid network in “Forget Gate”, respectively; W_i and b_i represent weight matrix and bias of sigmoid network in “Input Gate”, respectively; \tanh is one-dimensional matrix, with value range [0,1]; W_c and b_c represent weight matrix and bias of tanh network, respectively; $f_t \times S_{t-1}$ indicates the information transmitted from “Forget Gate”, and $i_t \times \overline{C}_t$ indicates the updated information from “Input Gate”; W_o and b_o represents weight matrix and bias of “Output Gate”, with o_t being the output of LSTM cell.

The LSTM network has two categories of parameters: neuron-parameters and super-parameters. The neuron-parameters such as neuron weight and bias are updated in the

model iterations, while super-parameters namely time-step, batch-size, and cell-size vary in their sensitivity and need adjustment based on experience. In particular, the parameter time-step reflects the length of data series. It can extract and learn the periodic relationship within time-step by learning data series. The parameter batch-size represents the number of data series in the training data packet extracted at one time; the product of the parameters time-step and batch-size represents the training data amount extracted for a single training. And parameter cell-size represents the number of neurons in a single hidden layer.

In rainfall-runoff simulation application, the rainfall data is used as the input data of the LSTM network, the number of input nodes equals the number of rainfall stations in study basin. The runoff data is the output data of the LSTM network, and the number of output nodes is equal to the number of hydrological stations in study basin.

2.3 Performance Metrics

Nash-Sutcliffe Efficiency Coefficient (NSE) and Correlation coefficient R^2 are used to evaluate the simulation effect of the LSTM network. The calculation formula is given below:

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_o^t - Q_m^t)^2}{\sum_{t=1}^T (Q_o^t - \overline{Q_o})^2} \quad (7)$$

Where, Q_o represents the measured value, Q_m the simulation value, Q^t the value at time t , and $\overline{Q_o}$ mean value of the measured values. The closer the NSE value is to 1, the higher the reliability of the model.

$$R^2 = \frac{[\sum_{i=1} (Q_{m,i} - \overline{Q_m})(Q_{s,i} - \overline{Q_s})]^2}{\sum_{i=1} (Q_{m,i} - \overline{Q_m})^2 \sum_{i=1} (Q_{s,i} - \overline{Q_s})^2} \quad (8)$$

Where, $Q_{m,i}$ represents measured data, $Q_{s,i}$ simulation data, \bar{Q}_m mean value of measured data, and \bar{Q}_s mean value of simulation data. The value range of R^2 is [0, 1]. The closer the R^2 is to 1, the better the simulation effect.

3 Study area and model building

3.1 Overview of the study area

Three watersheds including Minjiang River, Jialing River and Yalong River, are selected as study area. The study details of the three watersheds are shown below.

The Minjiang River, an important tributary of the upper Yangtze River, covers a total drainage area of 135,400 km². Except for a very small amount of snow water replenishment in Zhenjiangguan Town and beyond of its trunk stream in Spring, the seasonal variation of Minjiang River's runoff in a year corresponds to its rainfall season. The annual average flow at the Minjiang Estuary is 2850m³/s, with the annual runoff of 90 billion m³. The rainfall data from 8 meteorological stations and runoff data from Xiangjiaba hydrological station in the watershed are used in the paper.

Jialing River is the largest tributary of the Yangtze River, with a length of 1345 km and a drainage area of 160,000km². The main runoff sources is rainfall and groundwater. The flood period is June to September each year. The rainfall data from 12 meteorological stations and runoff data from Shapingba hydrological station in the watershed are used in the paper.

Yalong River covers a total drainage area of approximately 136,000 km², with a length of 1571 km. The average annual flow is 1890m³/s and the annual runoff is 59.6 billion m³. Half of

the runoff comes from rainfall, with the other from underground water and snow melt water. Its annual runoff is abundant and stable, with little annual variability. The rainfall data from 7 meteorological stations and runoff data from Panzhihua hydrological station in the watershed are used in the paper.

The location of the three river basins and the distribution of their meteorological stations are shown in Fig. 3 below.

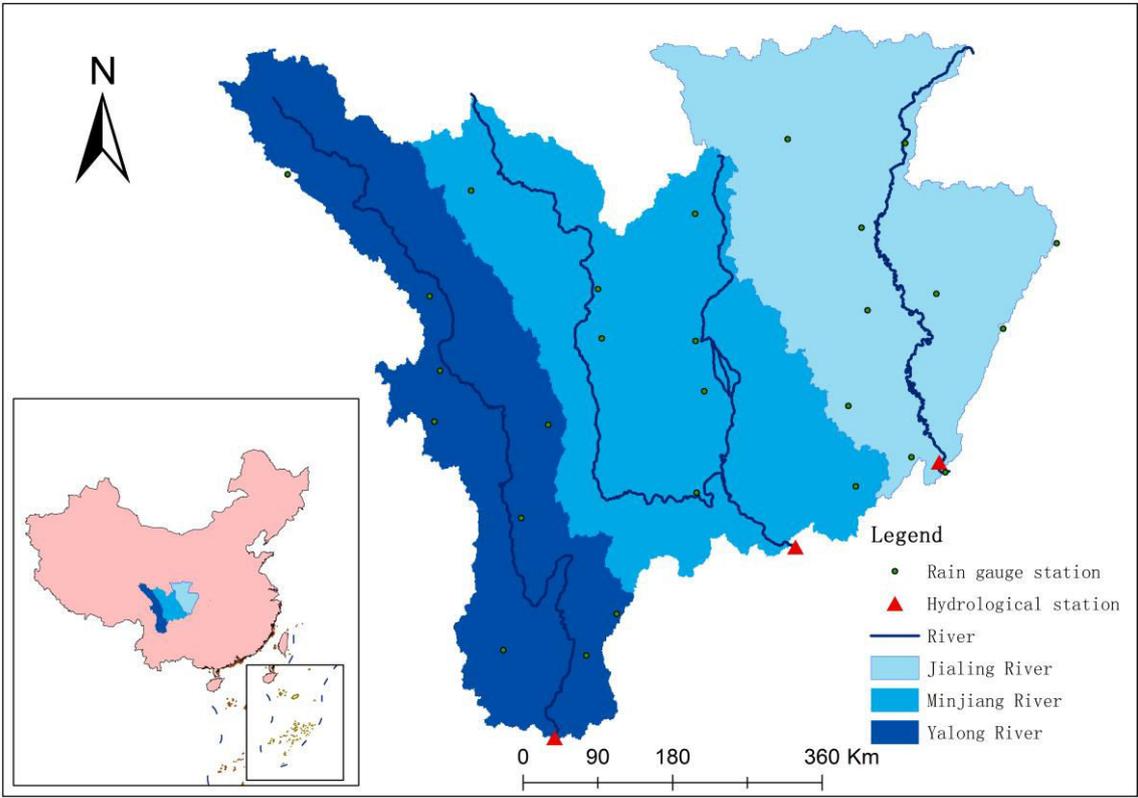


Fig. 3 The location of three river basins and distribution of their meteorological stations.

3.2 Generation of training data

The study was based on the daily meteorological observation data (temperature, wind speed and humidity, etc.) in Minjiang River, Jialing River and Yalong River basins observed during 1991

to 2016. First, the daily meteorological data and runoff observation data observed during 1991 to 2016 were used to calibrate SWAT model. Then, KNN algorithm, featured by being non-parametric and inert, was used to derive 130 years of daily rainfall data. Finally, the calibrated SWAT model was used to simulate 130 years of daily runoff process in the three river basins.

The procedure to derive prediction data using KNN model included the following steps: (1) prepare data (i.e., daily meteorological observation data during 1991 to 2016); (2) select parameter K value to predict number of neighboring days; (3) traverse the data set to extract the meteorological data of the neighbor days; (4) calculate the distance between the extracted meteorological data and their mean value; (5) select a probability depending on the distance to generate random meteorological data.

As a distributed watershed hydrological model, SWAT is often used to simulate a long-term hydrological process. In this study, the input of the SWAT model included such data as land use, rainfall, temperature, wind velocity and humidity, and the output was runoff data. The model was calibrated by using Sufi2 given in SWAT-CUP software.

Besides, the 130 years of rainfall and runoff data were divided into two parts: the first 80 years of data and the remaining 50 years of data. The former was used for model training and the latter for model verification. Among the training data, the first 5, 10, 20, 40 and 80 years of rainfall data (input factor) and runoff data (output factor) were taken respectively to train the LSTM network, so as to analyze the effect of training data amount on the runoff simulation performance of the LSTM network.

3.3 Selection of training parameters

Drawing lessons from the parameter selection experience in Huang (2020) and Xu et al. (2020) in this study, the parameters time-step, batch-size and cell-size are discretized into multiple values, while the parameters Num_layer, Learning_rate and epoch are set as a fixed value. Therefore, there are 36 different parameter combinations in total. The detailed parameter value ranges and meanings are shown in Table 1. Based on these 36 combinations, the learning performance was analyzed according to the schemes of 5, 10, 20, 40 and 80 years of training data.

Table 1 Table of LSTM Network Parameters

Parameter name	Parameter meaning	Value range
time-step	Maximum time series stage length	80, 120, 160
batch-size	Size of training data packet	10, 20, 30
cell-size	Number of neurons in single hidden layer	10, 20, 30, 40
Num_layer	Number of hidden layers in LSTM cell layer	1
Learning_rate	Learning rate	0.01
epoch	Iterations	600

4 Results and analysis

4.1 Analysis of KNN simulation results

To evaluate the rationality of KNN simulation results, the CV and CS of the simulated meteorological data series of 130 years are compared to those of measured series. The results of

monthly rainfall and daily maximum temperature at Leshan rainfall station in Minjiang basin are shown in Fig. 4. It can be seen that the hydrological characteristics of simulated data was similar to that of measured data, which indicates that the simulated data can be used for the building of the LSTM network.

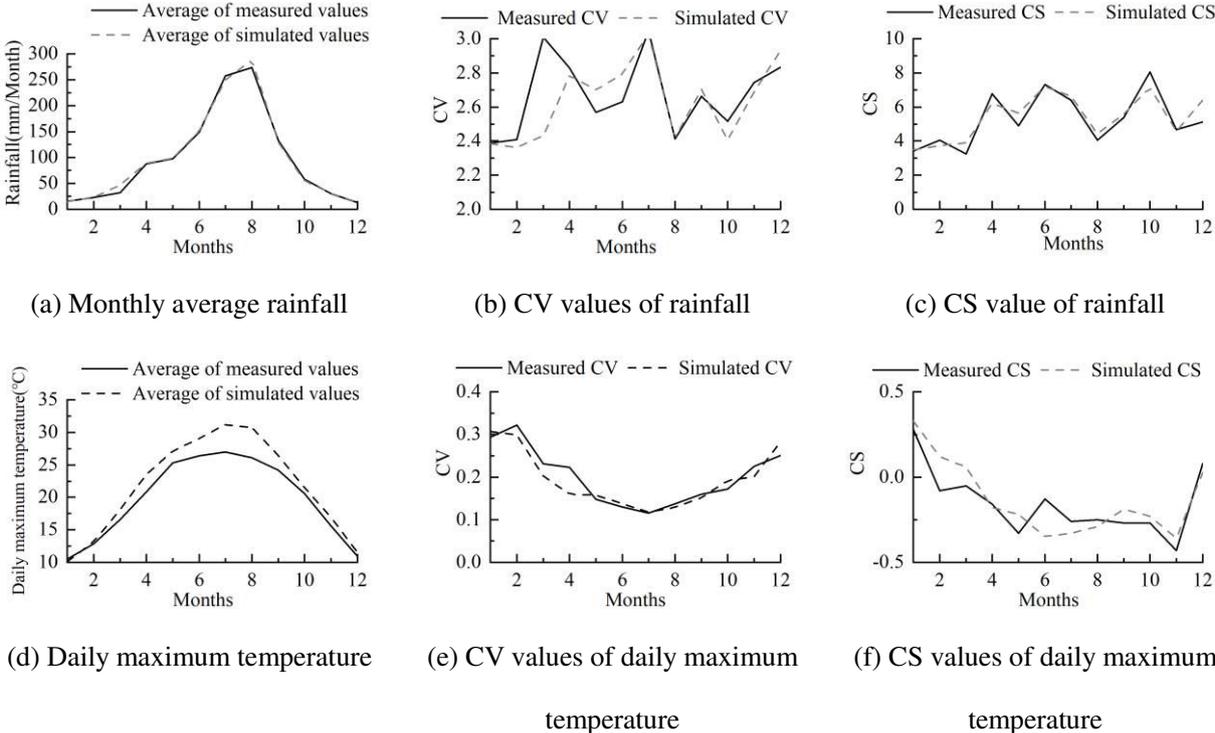
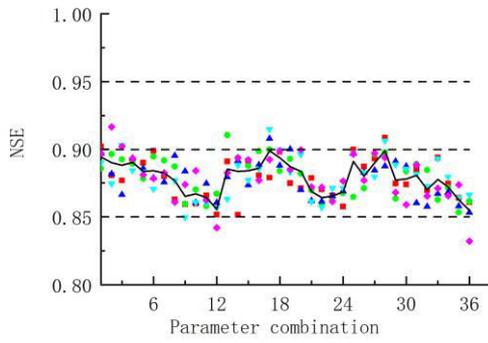


Fig.4 Verification of KNN simulation results in Yushu rainfall station

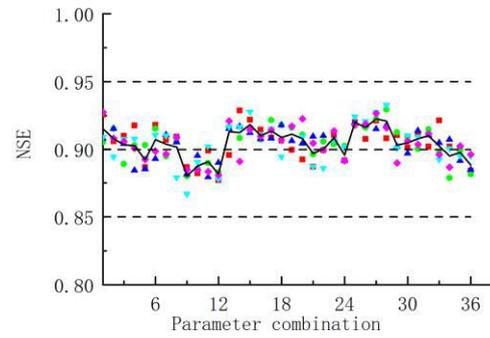
4.2 Effect of parameter combinations

To test the effect of the parameter combinations, 900 simulations, consisted of 36 available parameter combinations, 5 training data sets and 5 simulating times, were conducted for each of the three basins. Fig. 5 shows the NSEs of the verification phase in the Yalong River Basin. It shows that the accuracy of the LSTM network at the verification phase was relatively high, and the settings of time-step, batch-size and cell-size caused obvious and regular changes of the

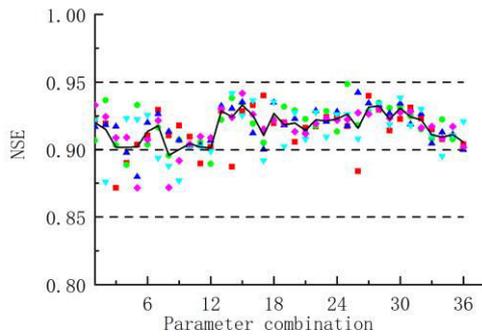
accuracy.



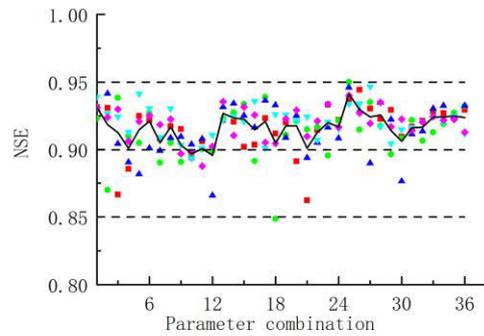
(a) 5 years



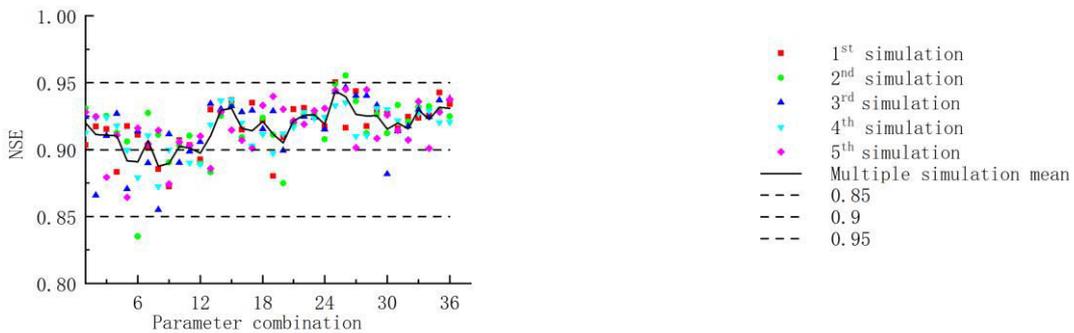
(b) 10 years



(c) 20 years



(d) 40 years



(e) 80 years

Fig. 5 Analysis of NSE of Yalong River Basin in the verification phase

The NSEs at the verification phase for the Minjiang River basin is shown in Fig. 6. It can be seen that, as the parameter combination changed from 1 to 4, time-step and batch-size were

80 and 10 respectively, and cell-size increased from 10 to 40, the NSE showed a downward trend, indicating that increasing the cell-size value could reduce the performance of the LSTM network.

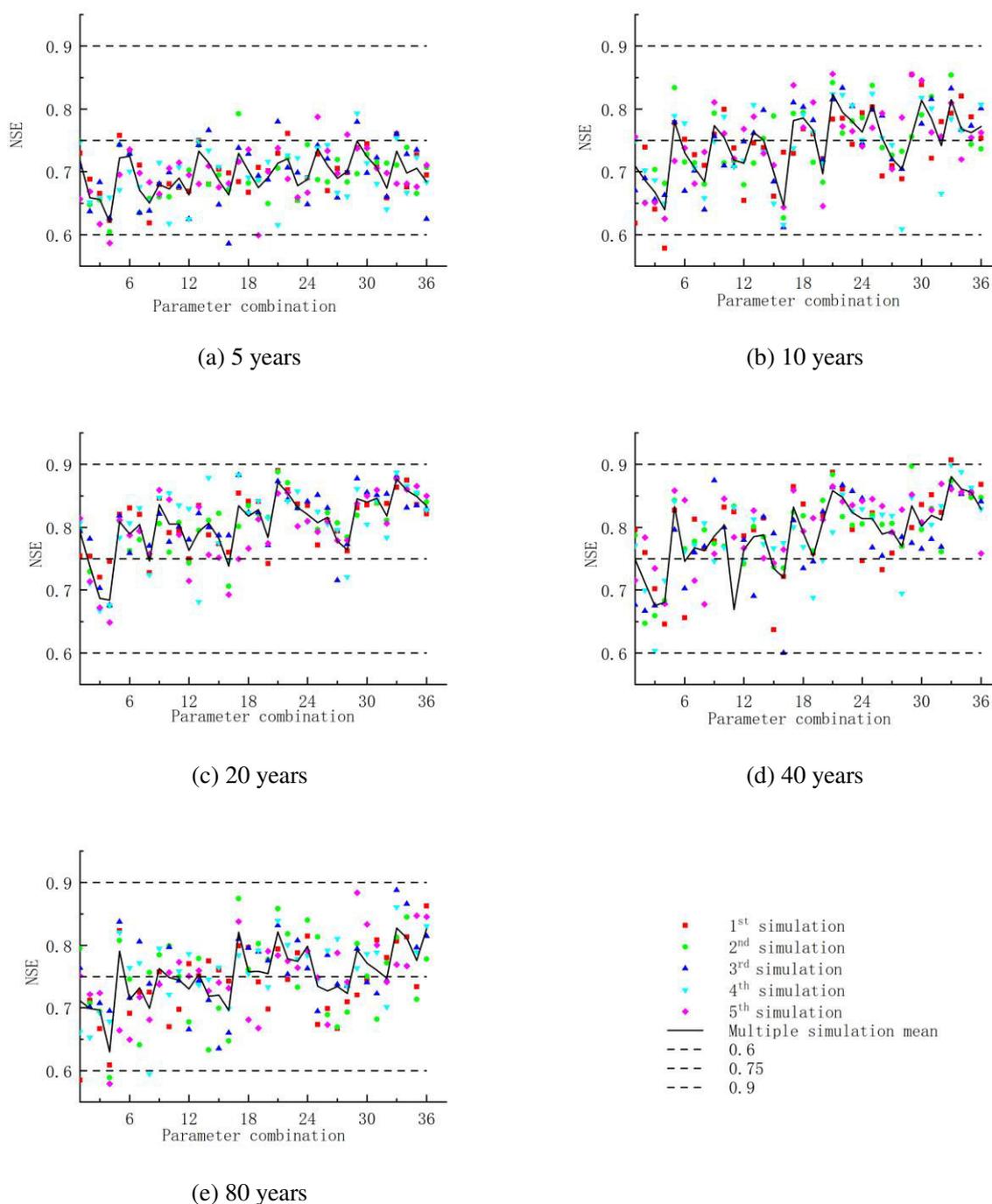


Fig.6 Analysis of NSE in the verification phase of Minjiang River Basin

The simulation results of Jialing River, shown in Fig. 7, are similar to that of Yalong River and Minjiang River. It also can be observed that parameter changes may lead to the obvious change in the performance of LSTM network.

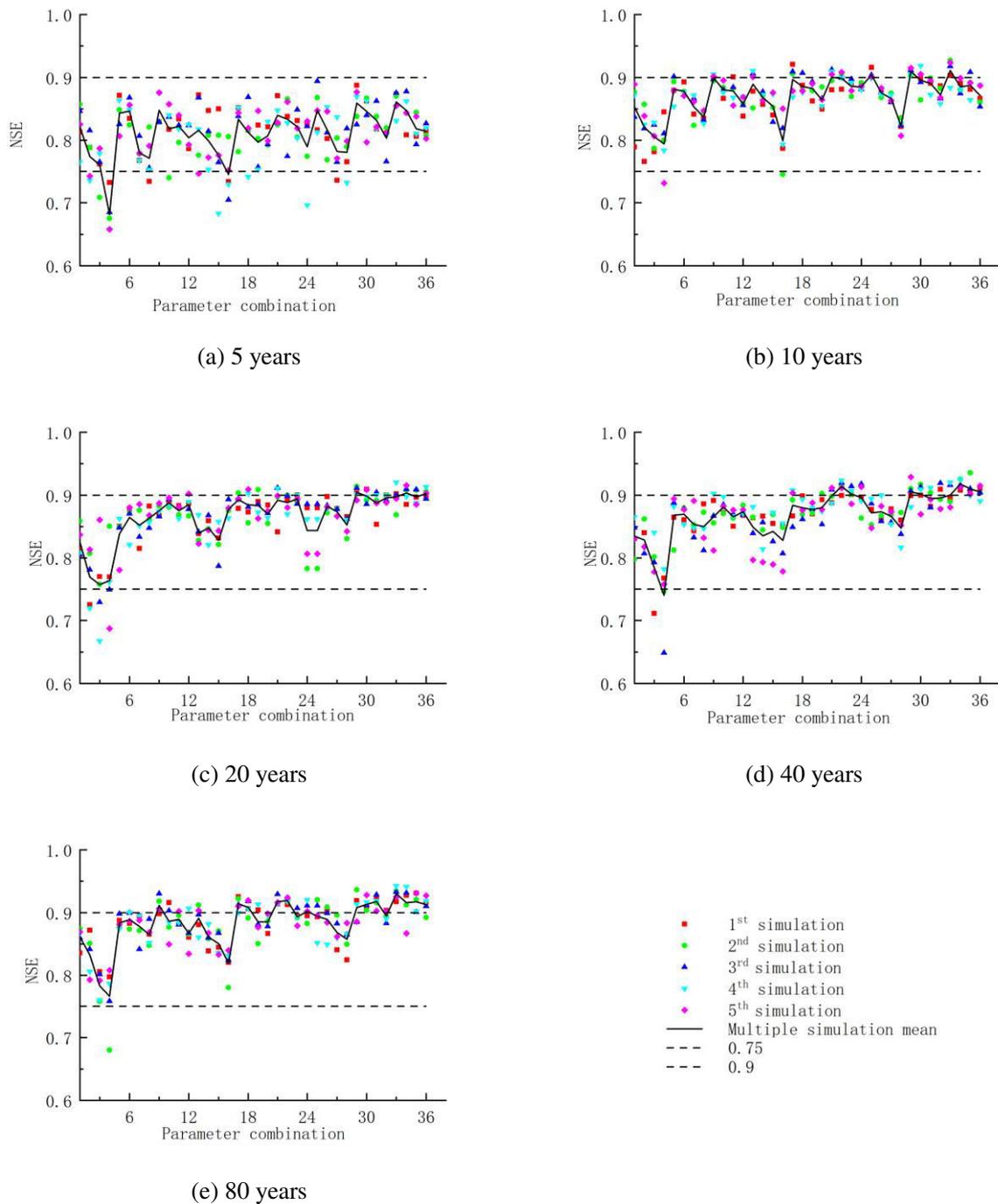


Fig.7 Analysis of NSE at the verification phase of Jialing River Basin

In summary, the prediction accuracy of the LSTM network fluctuated as the parameters changed. In order to further study the effect of each parameter on the prediction accuracy, Fig. 5(a) was enlarged and divided by crest and trough, as shown in Fig. 8. It can be seen that, Zone 1, Zone 2 and Zone 3 were located at the crest, and the corresponding LSTM network had a higher prediction accuracy; Zone 4, Zone 5, and Zone 6 were located at the trough, and the corresponding LSTM network had a lower prediction accuracy.

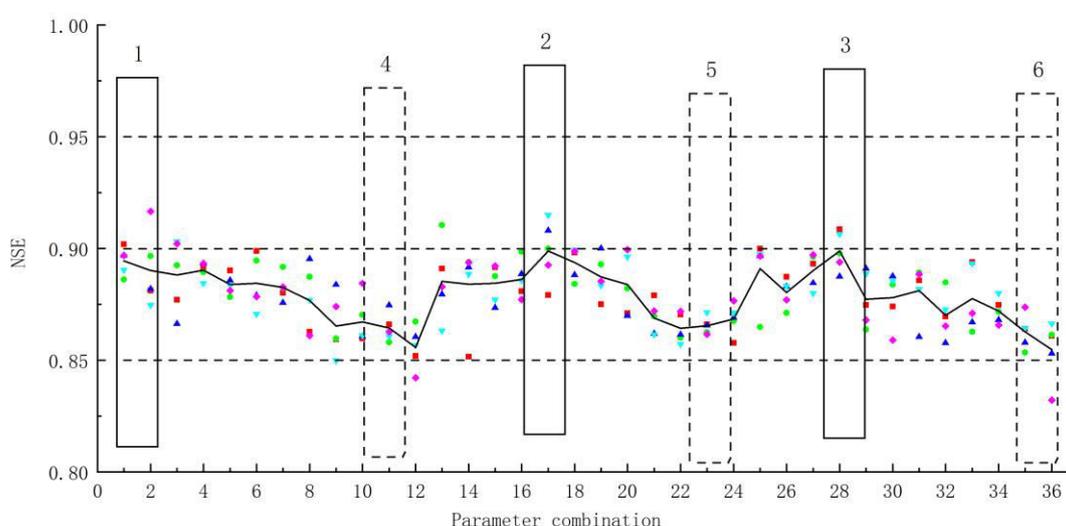


Fig. 8 Parameter analysis under different data amount conditions in the Yalong River Basin

Further statistics was conducted to get the information of each parameter corresponding to the parameter combination in each zone (see Table 2). For the parameter combinations at the crest, both the batch-size and cell-size values are mostly 10 and 20. This conclusion was similar to the results of the Huang's (2020) research which is that when the batch-size is small, the prediction of the LSTM network is accurate. For parameter combinations in the trough area, both batch-size and cell-size had larger values, 30 and 40 respectively. Therefore, it can be said that, the LSTM network could provide higher prediction accuracy when the parameter

batch-size and cell-size were smaller.

Table 2 Parameter combination values of peak and trough zone

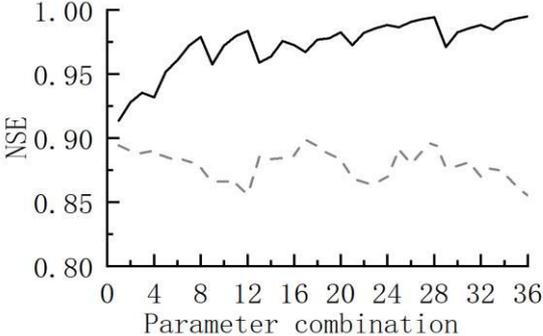
Zone	Parameter combination	1	17	18	28	29
Peak	time-step	80	120	120	160	160
	batch-size	10	20	20	10	20
	cell-size	10	10	20	40	10
Zone	Parameter combination	11	12	23	24	36
Trough	time-step	80	80	120	120	160
	batch-size	30	30	30	30	30
	cell-size	30	40	30	40	40

4.3 Effect of training data amount

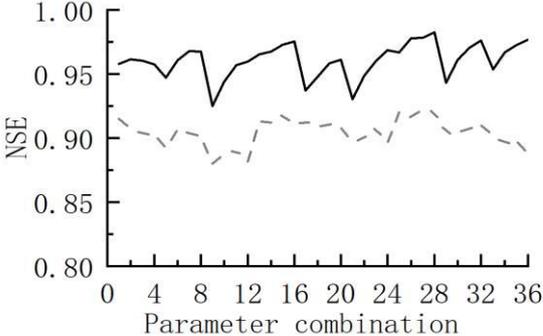
To evaluate the effect of training data amount on the performance of LSTM network, for the three basins, the mean NSE of 5-time simulation for the training phase and the verification phase with the same parameters and the same training data amount is calculated.

The results of Yalong River Basin is shown in Fig. 9. It can be seen that, in Fig. 9(a), the NSE at the training phase was much larger than that at the verification phase. The reason was that, the over-fitting, caused by the inadequate hydrological data at the training phase, leads to very poor extension capabilities of the LSTM network at the verification phase. Compared to Fig. 9(a), Fig. 9(b) showed that increasing the training data amount can lead to the decrease of the NSE at the training phase, and the raising of the NSE at the verification phase, as well as the reduction of the over-fitting phenomenon. In Fig. 9(c), the NSE at the 20-year training phase

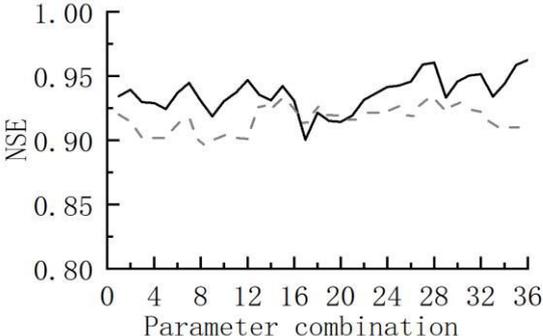
declined considerably than that at 5-year and 10-year training phases, with the NSE at the verification phase significantly improved; the gap between the NSE at the training phase and that at the verification phase became smaller, which proved that the trained network was highly stable. Fig. 9(d) and Fig. 9(e) showed that the increase of the data amount could further weaken the over-fitting phenomenon of the LSTM network, but the improvement of the NSE at the verification phase is limited. The results show that, as the data amount at the training phase increases, the LSTM network tends to become more stable.



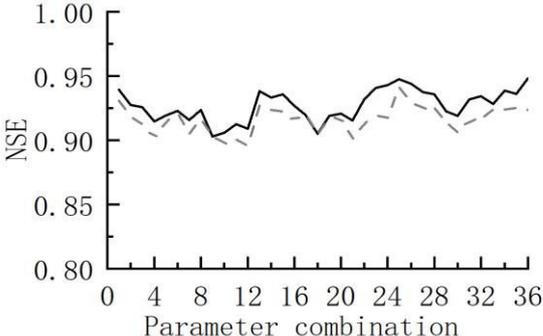
(a) 5 years



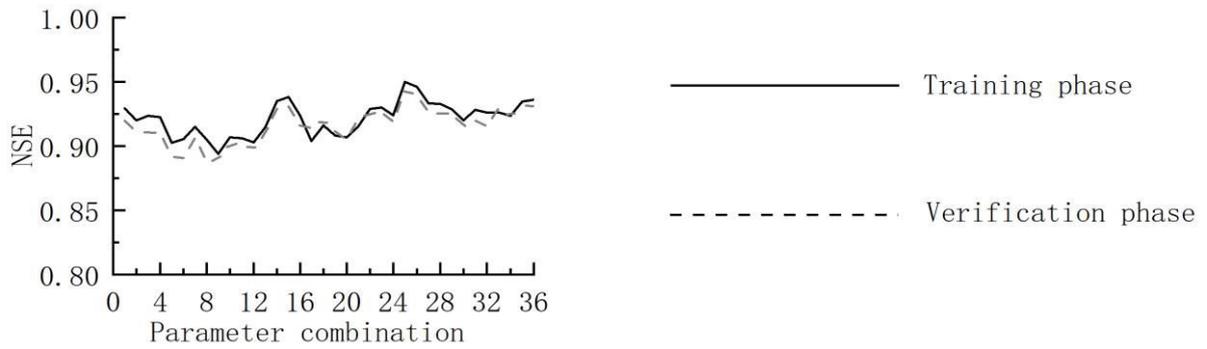
(b) 10 years



(c) 20 years



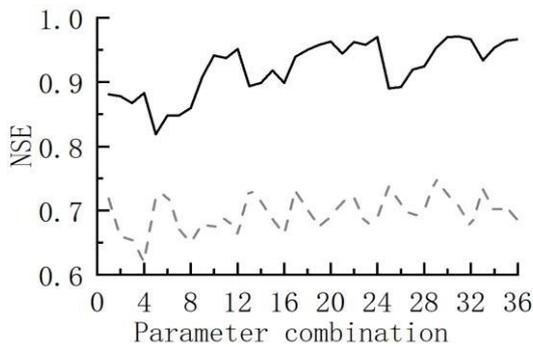
(d) 40 years



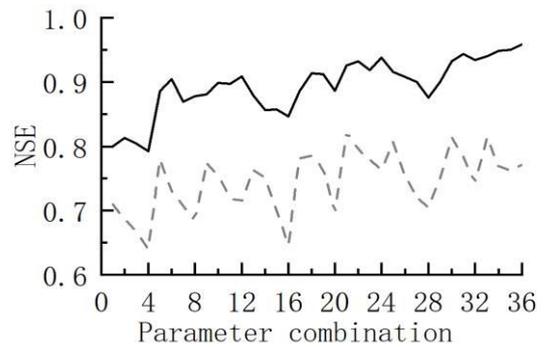
(e) 80 years

Fig. 9 NSE of Yalong River at the training phase and verification phase

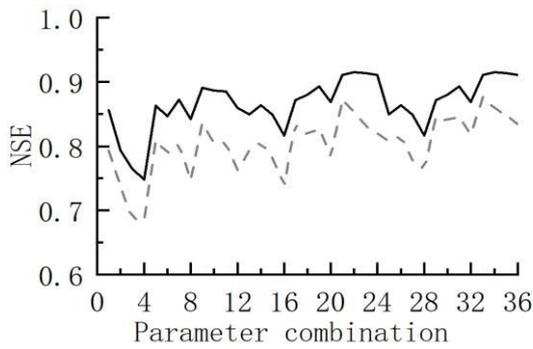
The simulation results of Minjiang River basin are shown in Fig. 10. In Fig. 10 (a), the NSE in the training phase remains around 0.9, while it is only around 0.7 in the verification phase; in Fig. 10 (b), 10(c), 10(d) and 10(e), the NSE at the training phase declines obviously while it increases at the verification phase. It is also corroborated that the increase of training data weakened the over-fitting of LSTM network.



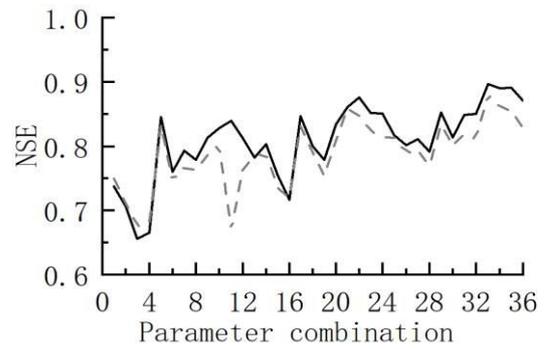
(a) 5 years



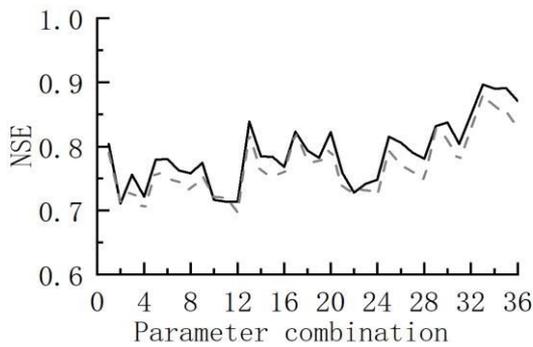
(b) 10 years



(c) 20 years



(d) 40 years



(e) 80 years

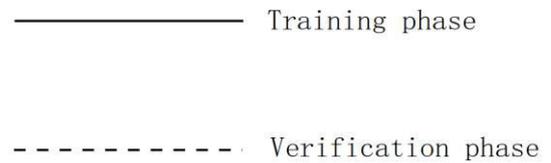
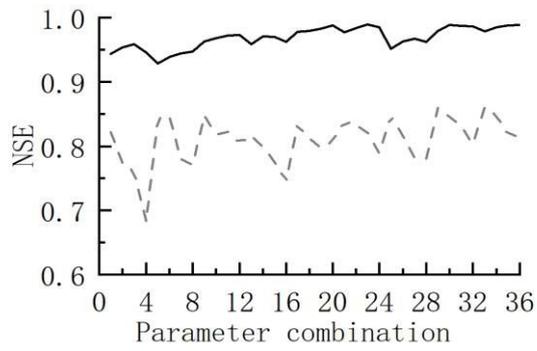
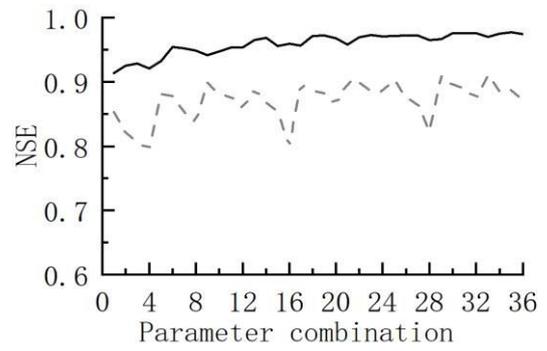


Fig. 10 NSE of Yalong River at the training phase and verification phase

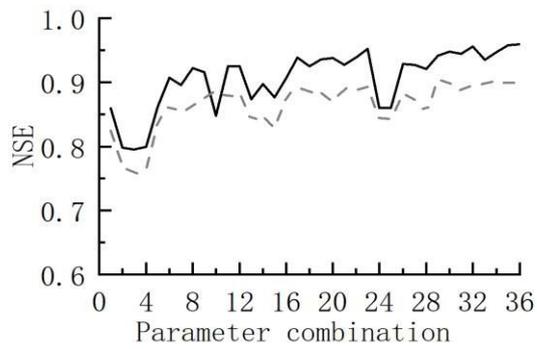
The simulation results of Jialing River basin is shown in Fig. 11. It can be seen that, when the training data amount increases from 5 years of data to 10, 20 and 40 years of data, the NSE at the training phase and verification phase gradually decreases. But when the training data amount increases to 80 years of data, the NSE at the training phase and verification phase stands at the same level, indicating that the simulation performance of LSTM network kept stable.



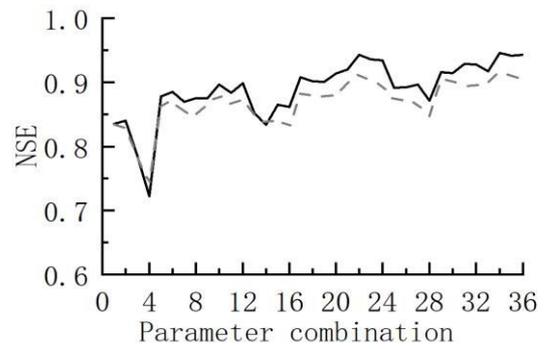
(a) 5 years



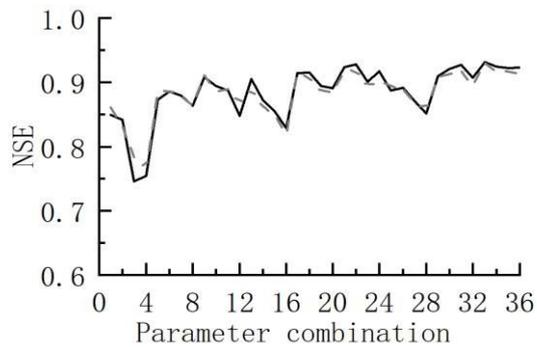
(b) 10 years



(c) 20 years



(d) 40 years



(e) 80 years



Fig. 11 NSE of Jialing River at the training phase and verification phase

In order to further analyze the effect of training data amount on performance of the LSTM network, the NSE and R^2 of 180 simulations for the verification phase with the same training

data amount for the three basins is averaged. The mean NSE and R^2 are shown in Fig. 12. It can be seen that, as the training data increases, the value of NSE and R^2 increases too, which indicates that the increase of training data amount can considerably reduce the over-fittings in the LSTM network and improve its prediction accuracy. But it also can be observed that when the training data amount exceeded a certain value or value range, the performance of LSTM network can't be improved remarkably, and the critical value varied for different basins. That is, increasing training data amount can improve the forecast performance, but over-increase of training data amount cannot necessarily result in better results. In other words, choosing an appropriate amount of training data can better exert the learning ability of the LSTM network.

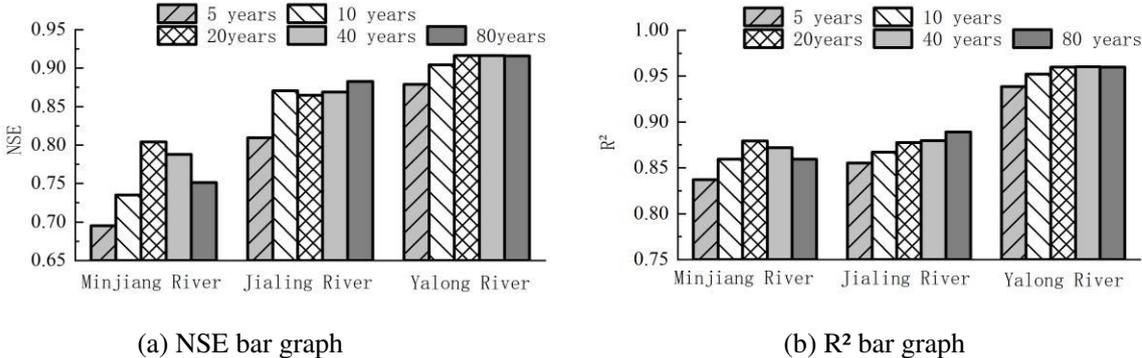


Fig. 12 NSE and R^2 average value at the verification phase

4.4 Analysis of runoff process

In Section 4.2 and Section 4.3, it is illustrated that LSTM network can achieve better performance when the parameters time-step, batch-size, and cell-size are set as 120, 10, and 20, respectively, and the amount of training data set as 20 years. The results of the corresponding LSTM network for the three basins are analyzed in this section to observe the learning ability of the LSTM network.

For the 50-year verification phase, the simulated runoff of the three basins are compared with the observed runoff in Fig. 13. It can be seen that the NSEs for Minjiang River, Jialing River and Yalong River at the verification phase are up to 0.91, 0.94 and 0.96 respectively, which indicates that the LSTM network has a strong learning ability and thus can be used for runoff simulation.

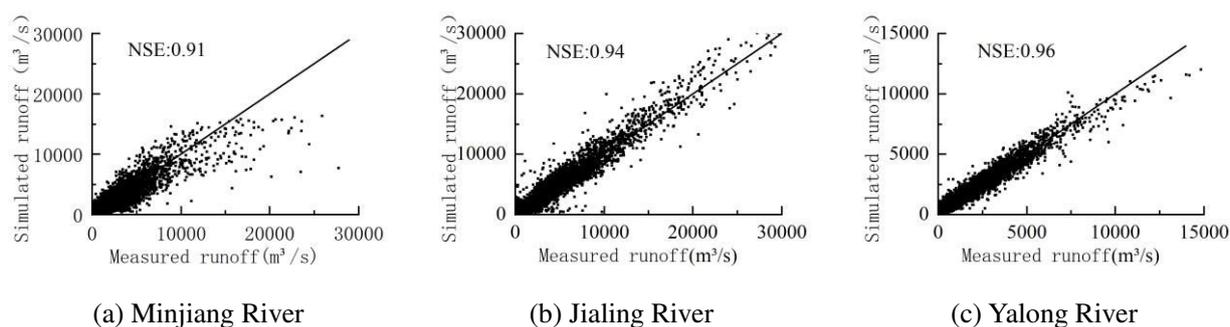


Fig. 13 Simulated runoff of the three basins in the verification phase

5 Conclusion

In order to analyze the influence of training data amount on the learning performance of the LSTM network in runoff simulation, the LSTM network is used for the runoff simulation of Minjiang River, Jialing River and Yalong River, and its performance under different parameters and training data amount schemes are evaluated in this study. The following conclusions can be drawn from the study:

- (1) When the parameters change, the prediction accuracy of the LSTM network changes too. And the results shows that the LSTM network could provide higher prediction accuracy when the parameter batch-size and cell-size are smaller.
- (2) As the training data amount increases, the over-fitting phenomena can be reduced and

the learning performance of the LSTM network improves significantly. But when the training data amount increases to a certain extent, the prediction accuracy of the LSTM network no longer improves much. The LSTM network in this study achieves better prediction accuracy and stability when more than 20-year training data amount are used.

(3) When the parameters of time-step, batch-size, and cell-size are set as 120, 10, and 20, respectively, and the amount of training data set as 20 years, the NSEs for Minjiang River, Jialing River and Yalong River at the verification phase are up to 0.91, 0.94 and 0.96 respectively, which shows that the LSTM network can be used for runoff simulation.

Statements and Declarations

Ethical Approval:

Not applicable. There are no ethical issues involved in this study.

Consent to Participate:

Not applicable.

Consent to Publish:

Not applicable.

Funding:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests:

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions:

All authors contributed to the study conception and design. Material preparation, data collection were performed by XW , model built by PAB, data analysis were performed by ZXL, figure drawn by TYY. The first draft of the manuscript was written by PAB and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Availability of data and materials:

The datasets used or analysed during the current study are available from the corresponding author on reasonable request.

References

- Ad A, Bpabc D, Aha E (2021) An improved SPEI drought forecasting approach using the long short-term memory neural network. *Journal of Environmental Management*, 283.
- Dami S, Esterabi M (2021) Predicting stock returns of Tehran exchange using LSTM neural network and feature engineering technique. *Multimedia Tools and Applications*, 1-24.
- Hinton GE, Salakhutdinov RR (2006) Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504–507.
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation*, 9(8): 1735-1780.

- Hu C, Wu Q, Li H, Jian SQ, Li N, et al (2018) Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water*, 10(11).
- Huang WF (2020) The effect of super parameter adjustment of long and short-term memory model in runoff prediction. *Journal of Xihua University (Natural Science Edition)*, 39(01): 23-29. (in Chinese).
- Huang Z, Hasan A, Shin K, Li RH, et al (2020) Long-term pedestrian trajectory prediction using mutable intention filter and warp LSTM. *IEEE Robotics and Automation Letters*, (99): 1-1.
- Huang Z, Wang J, Pi L, et al (2021) LSTM based trajectory prediction model for cyclist utilizing multiple interactions with environment. *Pattern Recognition*, 112: 107800.
- Jiang Y, Liu X (2021) Image/video aesthetic quality management based on fusing 3D CNN and shallow features. *Future Generation Computer Systems*, 118 (7626).
- Lall U, Sharma A (1996) A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour Res* 32: 679-693.
- Lin ZH, Chen AY, Hsieh SH (2021) Temporal image analytics for abnormal construction activity identification. *Automation in Construction*, 124: 103572.
- Mao G, Wang M, Liu J, et al (2021) Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall-runoff simulation. *Physics and Chemistry of the Earth Parts A/B/C* (1): 103026.
- Wang XJ, Wang YP, Yuan PX, et al (2021) An adaptive daily runoff forecast model using VMD-LSTM-PSO hybrid approach. *HYDROLOG SCI J*, 66(9): 1488-1502.
- Xu W, Jiang Y, Zhang X, et al (2020) Using long short-term memory networks for river flow prediction. *Hydrology Research*, 51(2).
- Yang W, Liu C, Jiang D (2018) An unsupervised spatio-temporal graphical modeling approach for wind

turbine condition monitoring. *Renewable Energy*, 127: 230-241.

Yin, HL, Zhang XW, Wang FD et al (2021) Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model. *Journal of Hydrology*, 598, 126378.

Yu Y, Si X, Hu C, et al (2019) A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 31(7): 1235-1270.