2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

Spatio-Temporal Attention LSTM Model for Flood Forecasting

1st Yukai Ding
College of Computer and Information
Hohai University
Nanjing, China
dingyukai@hhu.edu.cn

2nd Yuelong Zhu
College of Computer and Information
Hohai University
Nanjing, China
ylzhu@hhu.edu.cn

3rd Yirui Wu
College of Computer and Information
Hohai University
Nanjing, China
wuyirui@hhu.edu.cn

4th Jun Feng
College of Computer and Information
Hohai University
Nanjing, China
fengjun@hhu.edu.cn

5th Zirun Cheng
School of Foreign Languages
Tongji University
Shanghai, China
1833052@tongji.edu.cn

***Abstract*—In order to reduce the loss caused by flood, a large number of researches based on data, algorithms, machine learning and other technical means are used to realize flood forecasting. It will be a kind of flexible research method to realize the flood prediction of small and medium-sized rivers through intelligent models such as neural network. The area of small and medium-sized river basins is relatively small. Precipitation, soil moisture, evaporation and other factors can affect the timely total runoff prediction. However, not all the hydrological features is always valuable for flood forecasting, even at some time, noise of the factors will have larger interference on forecast process. Therefore, dynamic extraction of key feature vectors from various hydrological information plays an important role in flood forecasting. This paper proposed a flood forecasting model (STA-LSTM model) by using long short-term memory model (LSTM) and attention mechanism. We take the Lech river basin in Europe as the experimental basin and the results show that STA-LSTM performs well and has high research value with comparison of support vector machine (SVM),fully connected network (FCN) and original LSTM.**

***Index Terms*—flood forecasting, neural network, LSTM, spatio-temporal model, attention mechanism**

I. Introduction

As one of the most common and widespread hydrological phenomena, floods can occur in all types of rivers. Flood often cause huge economic losses to human society due to their sudden and devastating nature. Therefore, it is very important to forecast hydrological disasters, especially flood disasters. It is of great significance to explore innovative hydrological forecasting methods.

In order to minimize the impact of flood on us, researchers have made great efforts. We usually divide the designed flood forecasting models into two types: physical model [1]–[6] and data-driven model [7], [8]. Physical models usually represent complex water conservancy processes through mathematical physical functions and output the forecast results according to the input and output. However, the traditional physical model is very sensitive to its internal parameters, which requires a great deal of attention from researchers. The data-driven model is an end-to-end model that directly explores the relationship between various historical hydrological features and runoff. Since 2000, with the rapid development of computer technology and related disciplines, many new forecasting methods have been created and practiced. Among them, the model structure based on big data has made remarkable achievements. These technologies include not only Bayesian network [9], [10], SVM model [11], [12], deep learning and neural network [7], [13]–[15], but also the variation of the above technologies and the research methods of mutual integration. Under the existing conditions, the results of these studies are satisfactory. And with the further development of technology, many other fields research results are of great significance for the development of hydrological prediction.

We take the LSTM as the main part of the neural network model structure. LSTM network is a variant of the recurrent neural network (RNN), which is suitable for processing long time series information. We apply it to explore the relationship between flood influencing factors and final runoff. However, without the help of hydrologists in building physical models, it would be difficult to obtain satisfactory results if the obtained hydrological feature data were directly applied to the model construction. This is because there are some features that are useless or even interfere with the final flood forecasting in the collected hydrological data. Take soil water content as an example: soil water content, especially in the early stage of rainfall, is an important factor affecting runoff, which has been revealed by many watershed hydrological simulation studies [16]. However, soil water content has different effects on final runoff in different regions. In humid areas, soil water content has a great influence on runoff in the early stage of flood, while in dry areas, its correlation is not so strong. And once the soil water content exceeds the maximum capacity of the soil, the soil water content will no longer significantly affect runoff. Therefore, it can be seen that the influence of the

978-1-7281-2980-8/19/$31.00 ©2019 IEEE
DOI 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00095

same hydrological characteristics on the same flood may also change at different times and in different locations.In order to achieve better flood prediction, a dynamic feature selection model should be established so that different feature quantities can be selected according to the actual situation.

In view of the above problems, we find that attention mechanism can achieve the goal of dynamic feature extraction. As a relatively new research method, attention mechanism has made great achievements in various structural prediction tasks, such as image/video labeling and visual problem response. The highlight of the attention mechanism is that it borrows from the human visual attention feature: instead of looking at the entire image (or data environment) at once, people usually view parts of the the entire data environment (or picture) as needed (for example, according to time). Inspired by attention mechanism, we study basin by means of gathering since the related hydrological data (such as precipitation, precipitation, air temperature, soil moisture content, surface runoff), of generalization of from different locations in different time hydrologic hydrologic characteristics were dynamically select (distribution), to inherent law, to deal with small and medium-sized rivers complex hydrological environment of various kinds of noise and uncertainty.

This design is inspired by the visual attention mechanism. Through the study on the right bank of the Danube in Europe tributary Lech river basin, the proposed data driven flood forecast model is established by taking advantage of the LSTM network to process sequence information for a long time. The LSTM weight distribution enables the model to the scheduled within the scope of the dynamic focus on important hydrological characteristics at any spatial location, and eventually the model gives the weighted calculated runoff values then realizes the flood forecasting.

II. Related Work

This part will introduce the relevant research results that inspired us to design the STA-LSTM model, mainly including data-driven hydrological prediction model, LSTM network and attention mechanism.

A. Data Driven Model for Flood Forecasting

Data driven model for flood forecasting has been developed for a long time accompany with cloud-edge computing [17], big data technology [18] and other technologies, in order to offer a desirable computing service. Li et al. [11] analyzed support vector machine (SVM) model through experiments, and obtained the conclusion that the support vector machine model could achieve a relatively better performance. Srivastava et al. [19]successfully predicted the inflow of Tarbela reservoir by using the regression and neural network fusion model. Redmon et al. [20] built an RNN (recurrent neural network) model to predict time series, so as to realize flow prediction by processing multi-source precipitation information. Ma et al. [21] constructed four models including SVM, BP and other models. After the analysis and optimization of the model results, the hybrid SVM and BP neural network models applied in the Changhua river basin were proposed. Cheng et al. [22] proposed an artificial neural network model based on the quantum particle swarm optimization algorithm to predict the daily flow of reservoirs. Recently, Wu et al. [8] proposed an context-aware LSTM network model, and achieved a more accurate sequence flow prediction target based on the collected characteristics of various flood data. The above network model uses the attention module in each step of LSTM to achieve a high prediction accuracy.

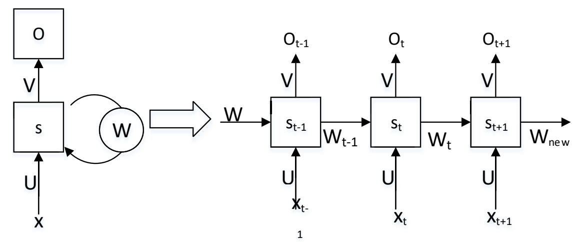

Fig. 1. Illustration of RNN.

B. LSTM Structure

Long short-term memory network is a kind of time recurrent neural network, which was proposed by Hochreiter and Schmidhuber [23] in 1997. It is suitable for dealing with and predicting important events with relatively long intervals and delays in time series. The typical recurrent neural network(RNN) is shown in Fig.1. After receiving the input x_t at time t, the output value of RNN unit is o_t after processing, and the hidden layer value is s_t. Among them:

$$o_t = g(Vs_t) \tag{1}$$

$$s_t = f(U_{x_t} + W_{s_{t-1}}) \tag{2}$$

Equation (1) is the calculation formula of the output layer. Nodes of the output layer are directly or indirectly connected to each node of the hidden layer. Equation (2) is the calculation formula of the hidden layer, and it is the part of realizing ”loop” in the network.

Where V is the weight matrix of the output layer, U is the weight matrix of the input x, g is a activation function, and f is also a activation function. W is the weight matrix that measures the importance of s_{t-1} at the last moment for current moment.

However, the basic version of RNN trained by gradient descent is difficult to deal with long-distance-dependence task, and usually it can only retain the information in the last three moments. Therefore, in order to preserve the influence of input in previous moments on the system, the long-short memory network adds a special unit to the original RNN network to store long-term state. The newly added state c is called the cell state.

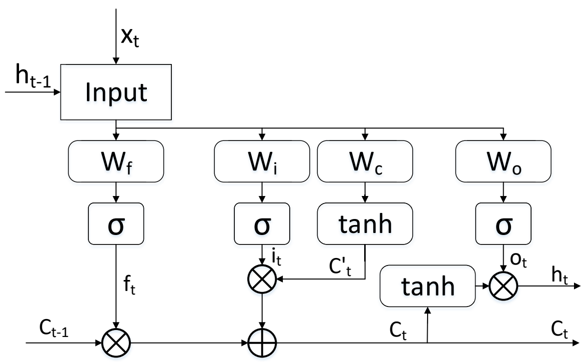

Fig. 2. Illustration of LSTM structure.

AS shown in Fig.2, main structure of LSTM network includes :

- forget gate. This structure determines how much state of the element c_{t-1} to c_t is maintained at the previous time, depending on the parameters.
- input gate. Determines how many input entries x_t to c_t are left in the network at time t.
- output gate. Controls the h_t output of the network at the current time.

In general, the specific state of the network at a certain moment is controlled by the forget gate and input gate, that is, the state of the unit is affected not only by the input at the current moment, but also by the previous network state. In this way, the output of the current moment can be associated with all the previous inputs to solve the problem of memorizing long time series information. And there is no doubt that LSTM training can use the backpropagation algorithm (BPTT).

C. Attention Mechanism

The information processing mechanism of human visual system is a highly complex process, among which the most important is the visual attention feature. Human vision can always quickly locate important target areas for detailed analysis, while other areas are only roughly analyzed or even ignored. This kind of active and selective psychological activity is called the visual attention mechanism. This is a method that human beings use limited attention resources to quickly pick out high-value information from a large amount of information. It is a survival mechanism formed in the long-term evolution of human beings. The human visual attention mechanism greatly improves the efficiency and accuracy of visual information processing.The attention mechanism in deep learning is essentially similar to the human selective visual attention mechanism, and the core goal is also to select more critical information from numerous information for the current task goal.

This mechanism of visual attention has been applied in many fields, such as machine translation, image tagging and video motion recognition. The research and application of visual attention model in related fields are mainly divided into two types: rigid attention and flexible attention. Hard focus refers to the mechanical selection of the input data area, reflected in the mathematical level is the input of the system is directly multiplied with the numerical value of 0 or 1 (the concern weight). The flexible attention refers to that the weight w is the value between 0 and 1, and the weight selection range is more flexible.

For example, Mnih et al. [24] proposed a model of RNN. This model can dynamically extract the information in an image by giving it a high resolution in a specific region. Song et al. [25] proposed an end-to-end spatio-temporal attention model to realize the recognition and prediction of human actions in video. Chen et al. [26] proposed a model of spatial and channel focus and image labeling combined with convolutional neural network, which performed well in the experimental data set. Liu et al. [27] proposed an LSTM network model integrating global situational awareness and attention to realize human 3D motion recognition. The model is capable of selectively focusing on different joints of the human body at different times, so as to accurately recognize 3D movements. These research results make full use of the visual characteristics of human beings, combine the characteristics of computer and traditional network to select the original image, video and other data with multi-dimensional weighted information, so as to achieve the goal of the system.

Inspired by the above models, on the basis of the traditional LSTM model, we adjusted the attention model to distribute the attention weight from the time and space dimensions, and solved the problem of runoff forecasting to some degree.

III. The Proposed Method

A. Overview

Our attention mechanism based LSTM model takes the Lech river basin as research object. Fig.3 shows the Lech river basin. Lech river is the right bank tributary of the Danube, Europe's second longest river. It originates in northwest slope of Lysitar mountain in Tirol, Austria,and flows into the Danube at the 40 km north of Augsburg in Germany. The climate throughout the region is warm and humid with scattered rainfall. The total length of Lech is 263 km;the basin area is 4126 km^2; and estuary average annual flow is 120 m^3/s. Our work is to realize the forecasting of the surface runoff(in the red circle at Fig.3) through the attention-based model proposed. The input time length is k (back from the the time interval is 3 hours), and the prediction period is 9 hours. STA-LSTM model uses data from the European Centre for medium-range Weather Forecasts (ECMWF). The input system's feature data including precipitation, evaporation, soil moisture content and so on. The output of the system is the surface runoff at the downstream observation points. The time accuracy of the data is 3 hours, and the spatial accuracy is 0.01*0.01 radian (longitude and latitude).

The structure of the model is shown in Fig.4. The spatio-temporal attention module will assign different corresponding weights to the input and output of LSTM network unit, so that

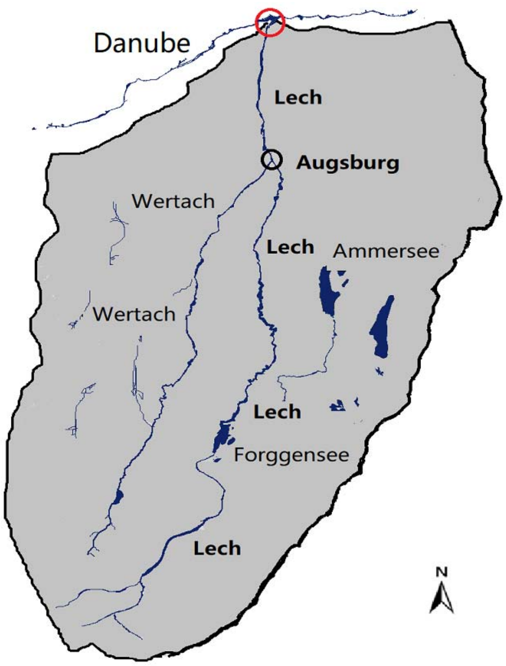

Fig. 3. Illustration of Lech river basin.

the model can more effectively discriminate the feature vectors of system input. Because LSTM network is designed to deal with the gradient-disappear problem that RNN comes acrossed and the information will be kept for a long time. However, that cause LSTM can not decide whether the information retained is useful or not. Our model uses the spatial and temporal attention module to selectively focus the information, thus the model can disposal the useless and retain the useful then achieve more accurate output forecast results. And the workflow can be described as follow:

$$I = D(rawdata) \tag{3}$$

$$\alpha_t = \varphi_S(\tilde{h}_{t-1}) \tag{4}$$

$$\beta_t = \varphi_T(I_{t-1}, I_t) \tag{5}$$

$$I_t^{\alpha} = I_t \otimes \alpha_t \tag{6}$$

$$h_t = LSTM(\tilde{h}_{t-1}, I_t^{\alpha}) \tag{7}$$

$$\tilde{h}_t = h_t \odot \beta_t \tag{8}$$

Raw data become the normalized input after the relevant data processing $D(\cdot)$. All processed data are collectively referred to as $I = I_1, I_2...I_t$, $I \in R^{(H \times W \times n \times k)}$, in which HxW for a single feature dimension of the amount of data, n is the number of eigenvectors, t is the time span of the network.

At time t,there are n features(such as rainfall,soil moisture content), then $I_t$and α_t are going to take $\otimes$ operation. $\otimes$ denotes element-wise operation. The result of that isI_t^{α}. α_t is the spatial weight. h_t is the hidden state of LSTM cell. When the module gets temporal weight β_t, the h_t will take the $\odot$ operation with β_t. $\odot$ operation is the first operator(in 3D tensor form) is multiplied by the second operator(in numerical form), and the dimension of the result is the same as that of the first operator. We want the length of the second operator to be the same as the number of independent 2D matrices of the first operator.

φ_S and φ_T denote the operation that can get the α_t and β_t(Details will be show in next two subsections.). Data feature F is the result of the $\oplus$ operation. That means we get F from all hidden states of LSTM cells. Then we can get the forecasting regression results.

B. Spatial Attention Module

According to the characters of category, all data can be divided into multiple channels to be sent into the model, but not all the data will have a positive contribution to the results. A slightly worse case is giving large weight to the less important features, causing the model part of the phase. The worst situation is the introduction of the noise, leading to training the model in a wrong way and the failing research . In order to improve the intelligence and extendibility of the system, this model allows a variety of input data features, compared with the traditional hydrological models rely on expert knowledge and experience as well as ordinary data-driven intelligent model can simply select several obvious characteristics for reference. This model is more flexible for different application environment. But there is no denying that compared with the traditional intelligence model, this model consuming more computing resources, but its cost performance is better.

In order to further improve the utilization rate of the data and the accuracy of prediction, we add a spatial module into our model. This part uses the space information carried by the original data for reference and focus on the hydrologic characteristics of different space position with different weight. That is shown in Fig.5

In Fig.6, we can obviously get that α_t is from LSTM cell and fully connected network. The operation in 2 is defined as in 9.

$$\alpha_t = sigmoid\left(\left(W_S\tilde{h}_{t-1} + b_s\right)\right) \tag{9}$$

where b_S is a bias with the same dimension as the variables added. W_S is the weight inside the full-connected network structure. Sigmoid function is an activation function. One of the important purposes of sigmoid function is to change the input value into the value between 0 and 1. The mathematical expression is in 10.

$$sigmoid\,(x) = 1/\left(1 + exp\,(-x)\right) \tag{10}$$

C. Temporal Attention Module

In order to achieve efficient extraction of data characteristics at different moments, we analyze data at different moments due to time differences which is quite similar to the process in last subsection. Therefore, we can see in Fig.6 that k moments

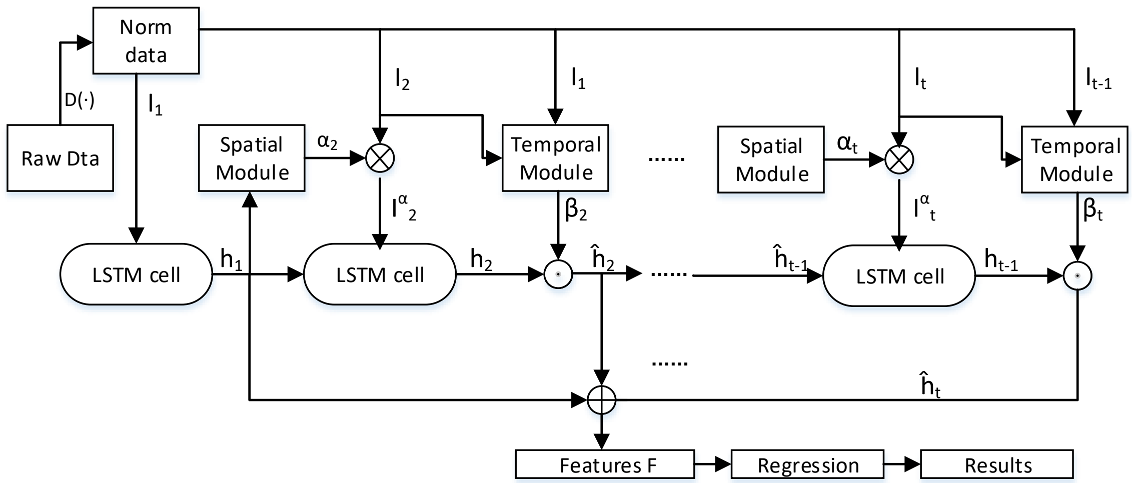

Fig. 4. Illustration of Lech river basin.

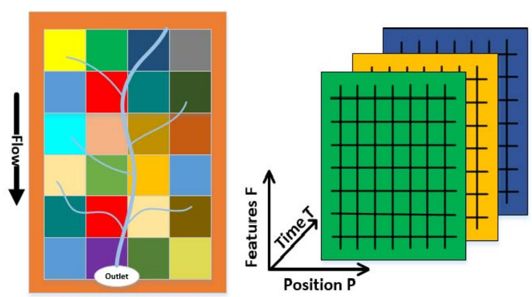

Fig. 5. Illustration of data structure.

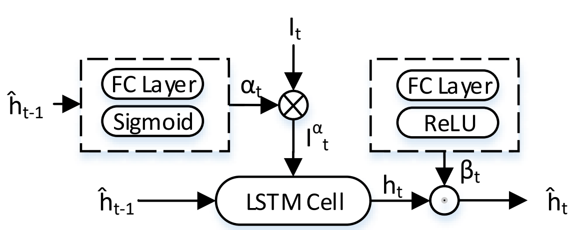

Fig. 6. Illustration of spatial and temporal module of STA-LSTM.

before the prediction received different attention. And the φ_T operation is defined in 11.

$$\varphi_t = ReLU\left(\left(W_{T,t-1}I_{T-1} + W_{T,t}I_t + b_T\right)\right) \tag{11}$$

where ReLU is defied in 12.

$$ReLU\left(x\right) = max\left(0, x\right) \tag{12}$$

IV. Experiments

We have four models, STA-LSTM, FCN, SVM and original LSTM, to compare and analyze in our experiments. This part will introduce the details of experiments, which includes dataset, implementation details and results.

A. Dataset

A total of 7360 hydrological data were collected from the selected Lech river basin from May 2002 to January 2018 for a total of 7360 moments during the period of significant increase in the downstream runoff volume. Among them, a total of 174 hydrological data at 174 moments during the period of significant increase of the downstream runoff (multiple floods) in the basin in May 2002, August 2013 and January 2018 were taken as the test data collection. A total of 58 consecutive sets of data were randomly selected from the test set for display. Table I is the basic characteristic analysis of data from training set and test set.

Holt-Winters double exponential smoothing filter are used to decrease the noise of raw data and to keep the original trend of the time series data.

B. Implementation

Python language is used as the actual coding language in the design of this system. The data were collected from the European Centre for medium-range Weather Forecasts(ECWMF). The spatial accuracy of the data was 0.01×0.01 radians, and the latitude and longitude range was (10.68E, 47.65N) to (10.94E, 48.73N). The time accuracy is 3 hours. The length of data input time k is 6 (the actual length is 18 hours), and the prediction period is 9 hours.

All experiments are carried out on a Linux server equipped with 2.4GHz 6-core Xeon CPU, 60GB RAM and Nvidia

TABLE I
DATASET CHARACTERISTICS

Data characteristics	Mean	Variance	Standard deviation	Median	Kurtosis	Skewness
Training dataset	430.81	18056.23	134.37	36.96	2.89	0.58
Test dataset	540.15	22562.36	150.21	536.95	2.26	-0.0651

GeForce GTX 1080 Ti. For LSTM and STA-LSTM model, the hidden state dimension of LSTM network is set to 128, and the number of cycles is set to 32. And for FCN, LSTM and STA-LSTM, the learning rates 0.01(the reason is shown in Fig.7), the weight attenuation is 10^{-6}, the total training epoches is set to 500 (because the RMSE will decrease very slowly even stop after 300 epochs according to our experiment results), the learning rate is updated(decline) every 100 epochs with the rate is 0.1. The parameters of SVM model are set to the default values which means C(Penalty coefficient) is set to 1.0 and gamma (Kernel function coefficient) is $1/n$, where n is the number of features.

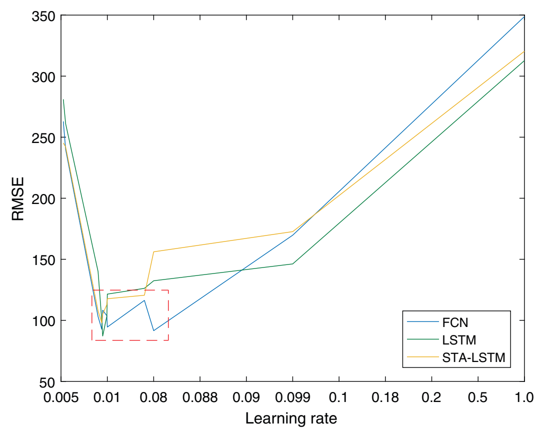

Fig. 7. This figure shows the relation between the learning rate and prediction error(RMSE).It is obvious that a proper learning rate should not be too large or small. The red dotted box marks the appropriate range of learning rate for our models.

In the test process, RMSE and DC are used to evaluate the performance of the model comprehensively. Among them, deterministic coefficients is usually used in hydrological forecasting.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - q_i)^2} \tag{13}$$

$$DC = 1 - \frac{\sum_{i=1}^{n}(y_i - q_i)^2}{\sum_{i=1}^{n}(q_i - \overline{q})^2} \tag{14}$$

In 13 and 14, n is the number of test samples while y_i means prediction result, q_i is groundtruth and $\overline{q}$ denotes the mean of all groundtruths. And if DC is large then the model is good for prediction.

C. Performance Analysis

This part will list the performances of FCN, SVM,LSTM network and STA-LSTM network model. The test data are selected from the test dataset at the time of $61 - 118$ in the test data set . The runoff at T+3, T+6 and T+9 (the adjacent time interval is 3 hours and the forecast period is 9 hours) is predicted through the data at the first 6 moments and the data at the last 3 moments, that is, the runoff at T+3, T+6 and T+9 (the adjacent time interval is 3 hours and the forecast period is 9 hours) according to the data from T to T-5. In the figures of experiment results, the X-axis is order number of the data points in time sequence, and the Y-axis is surface runoff, unit is cubic meters per second (m^3/s).

- T+3
  At T+3, it can be seen from Fig.8 that all models perform well, among which FCN model performs best, STA-LSTM performs close to LSTM and SVM models. However, SVM model is based on mathematical regression, and the predicted curve obtained is relatively smooth, and other models are jumping to varying degrees.

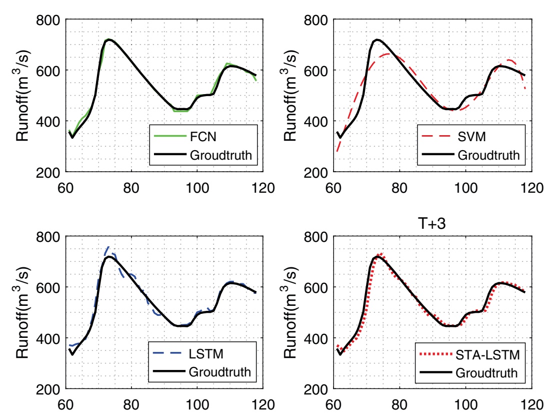

Fig. 8. Results at T+3. In this figure, the black solid line represents the original flow data, and the curve of other colors in the figure represents the predicted value of each model. At this time (T + 3), all models perform well, and the order of accuracy is as follows: FCN >STA-LSTM >SVM >LSTM.

- T+6
  The prediction curve of the whole model at time T+6 has started to be distorted compared with that at time T+3, but it is still relatively accurate. Since this model does not take the predicted flow rate as the input of the next prediction moment, the graphical distortion mainly comes from the fact that the model does not fully learn at different moments. At this moment, STA-LSTM

model performs relatively best, LSTM and SVM network perform slightly worse, and FCN performs the least.

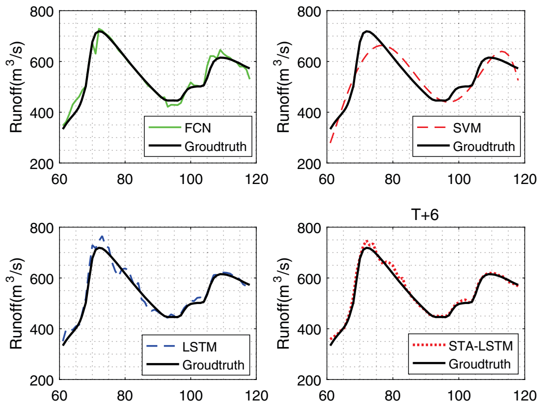

Fig. 9. Results at T+6. In this figure, the black solid line represents the original flow data, and the curve of other colors in the figure represents the predicted value of each model. At this time (T + 6), the prediction curve of the whole model starte to be distorted compared with that at time T+3, but it is still relatively accurate. And the order of accuracy is as follows: STA-LSTM >LSTM >SVM >FCN.

- T+9
  Similar to the characteristics at time T+6, the error of model prediction gradually increases with the passage of prediction time. At the time of T+9, the overall performance of STA-LSTM model was more stable compared with other models. There is some distortion but the curve fitting degree is the best. SVM and LSTM are relatively good, but there is a certain degree of distortion. However, the FCN model has serious distortion and large deviation in many data points.

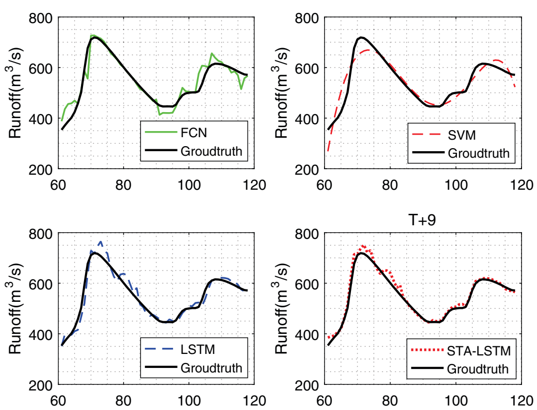

Fig. 10. Results at T+9. In this figure, the black solid line represents the original flow data, and the curve of other colors in the figure represents the predicted value of each model. At this time (T + 9),the error of model prediction gradually increases with the passage of prediction time while the order of accuracy is still: STA-LSTM >LSTM >SVM >FCN.

D. *Performance Comparison*

This part mainly compares and analyzes the performance of each model under the standard of RMSE and DC. Table II summarizes the root mean square error of the predicted results of each model at each moment while Table III describes the DC results. The results reflect our STA-LSTM model perform best among all four models at T+6 and T+9. Fig.11 is a graphical description of Table II. Fig.12 is a graphical description of Table III.

TABLE II
RMSE COMPARISION OF MODELS PERFORMANCE

Model	T+3	T+6	T+9	Average
FCN	51.27	82.54	123.41	85.74
SVM	63.99	79.98	92.49	78.82
LSTM	68.92	75.12	80.85	74.96
STA-LSTM	60.94	65.56	71.56	66.02

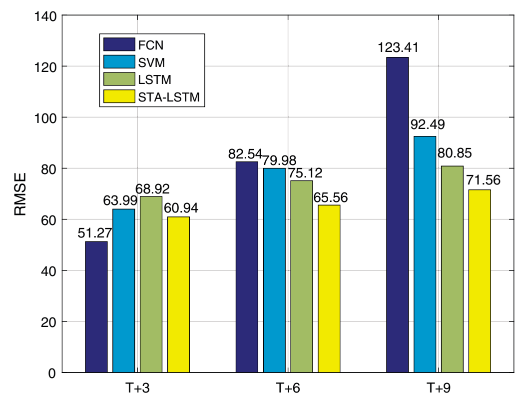

Fig. 11. RMSE comparison of models performance.

TABLE III
DC COMPARISION OF MODELS PERFORMANCE

Model	T+3	T+6	T+9	Average
FCN	0.88	0.70	0.32	0.633
SVM	0.82	0.72	0.62	0.720
LSTM	0.79	0.75	0.71	0.750
STA-LSTM	0.84	0.81	0.77	0.807

V. CONCLUSION

In this paper, we analyzed the attention mechanism developed rapidly in the field of data-driven intelligent model and video image, and proposed a STA-LSTM model, a hydrological prediction model based on attention mechanism. The basic structure and parameters of the model are introduced through analysis and derivation. The STA-LSTM model has shown good performance in comparison with FCN, SVM and traditional LSTM models. The researchers believe that the combination of attention mechanism and hydrological forecasting model of small and medium-sized rivers has higher research significance. In the future research work, we will

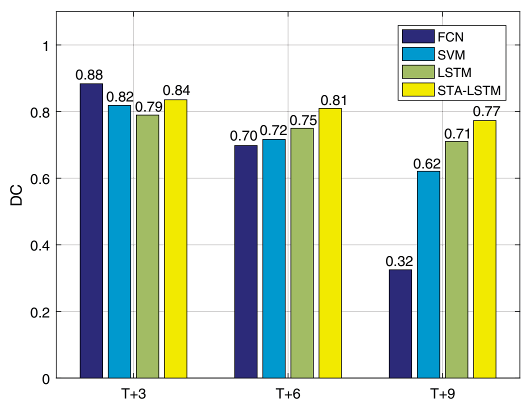

Fig. 12. DC comparison of models performance.

continue to conduct more model comparison experiments in combination with the research of attention mechanism and hydrological prediction, hoping to further optimize and adjust the structure and parameters of the model.

ACKNOWLEDGMENT

This work was supported by National Key R&D Program of China under Grant 2018YFC0407901, the Natural Science Foundation of China under Grant Grant 61702160, the Science Foundation of Jiangsu under Grant BK20170892, and the open Project of the National Key Lab for Novel Software Technology in NJU under Grant K-FKT2017B05.

REFERENCES

[1] E. Paquet, F. Garavaglia, R. Garçon, and J. Gailhard, "The schadex method: A semi-continuous rainfall–runoff simulation for extreme flood estimation," *Journal of Hydrology*, vol. 495, no. 15, pp. 23–37, 2013.

[2] R. Zhao, "The xinanjiang model applied in china," *Journal of Hydrology*, vol. 135, no. 1–4, pp. 371–381, 1992.

[3] N. A. Pierini, E. R. Vivoni, A. Robles-Morua, R. L. Scott, and M. A. Nearing, "Using observations and a distributed hydrologic model to explore runoff thresholds linked with mesquite encroachment in the sonoran desert," *Water Resources Research*, vol. 50, no. 10, pp. 8191–8215, 2015.

[4] D. H. Nam, T. M. Dang, K. Udo, and A. Mano, "Short-term flood inundation prediction using hydrologic-hydraulic models forced with downscaled rainfall from global nwp," *Hydrological Processes*, vol. 28, no. 24, pp. 5844–5859, 2015.

[5] M. A. Kabir, D. Dutta, and S. Hironaka, "Estimating sediment budget at a river basin scale using a process-based distributed modelling approach," *Water Resources Management*, vol. 28, no. 12, pp. 4143–4160, 2014.

[6] M. Rogger, A. Viglione, J. Derx, and G. Blöschl, "Quantifying effects of catchments storage thresholds on step changes in the flood frequency curve," *Water Resources Research*, vol. 49, no. 10, pp. 6946–6958, 2013.

[7] L. Fan, X. Feng, and S. Yang, "A flood forecasting model based on deep learning algorithm via integrating stacked autoencoders with bp neural network," in *Proceedings of IEEE Third International Conference on Multimedia Big Data*, 2017.

[8] W. Yirui, X. Weigang, F. Jun, S. P, and L. Tong, "Local and global bayesian network based model for flood prediction," in *Proceedings of ICPR*, 2018.

[9] K. Schröter, H. Kreibich, K. Vogel, C. Riggelsen, F. Scherbaum, and B. Merz, "How useful are complex flood damage models?" *Water Resources Research*, vol. 50, no. 4, pp. n/a–n/a, 2014.

[10] S. Han and P. Coulibaly, "Bayesian flood forecasting methods: A review," *Journal of Hydrology*, vol. 551, pp. 340–351, 2017.

[11] S. Li, K. Ma, Z. Jin, and Y. Zhu, "A new flood forecasting model based on SVM and boosting learning algorithms," in *IEEE Congress on Evolutionary Computation, CEC 2016, Vancouver, BC, Canada, July 24-29, 2016*, 2016, pp. 1343–1348. [Online]. Available: https://doi.org/10.1109/CEC.2016.7743944

[12] H. M. Azamathulla, A. A. Ghani, C. K. Chang, Z. A. Hasan, and N. A. Zakaria, "Machine learning approach to predict sediment load – a case study," *CLEAN - Soil, Air, Water*, vol. 38, no. 10, pp. 969–976, 2015.

[13] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "High-order graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting," *CoRR*, vol. abs/1802.07007, 2018.

[14] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, "Lstm-based traffic flow prediction with missing data," *Neurocomputing*, vol. 318, pp. 297–305, 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.08.067

[15] H. Li, Y. Shen, and Y. Zhu, "Stock price prediction using attention-based multi-input LSTM," in *Proceedings of The 10th Asian Conference on Machine Learning, ACML 2018, Beijing, China, November 14-16, 2018.*, 2018, pp. 454–469. [Online]. Available: http://proceedings.mlr.press/v95/li18c.html

[16] L. Jintao, L. Xiaopeng, C. Xi, and M. Can, "Soil moisture content distribution in intermittent rainfall and its effect on infiltration," *Journal of Soil and Water Conservation*, vol. 23, no. 5, pp. 96–100, 2009.

[17] X. Wang, L. T. Yang, X. Xie, J. Jin, and M. J. Deen, "A cloud-edge computing framework for cyber-physical-social services," *IEEE Communications Magazine*, vol. 55, no. 11, pp. 80–85, 2017.

[18] X. Wang, L. T. Yang, L. Kuang, X. Liu, Q. Zhang, and M. J. Deen, "A tensor-based big-data-driven routing recommendation approach for heterogeneous networks," *IEEE Network*, vol. 33, no. 1, pp. 64–69, 2018. [Online]. Available: https://doi.org/10.1109/MNET.2018.1800192

[19] P. K. Srivastava, D. Han, M. A. Rico-Ramirez, M. Bray, and T. Islam, "Selection of classification techniques for land use/land cover change investigation," *Advances in Space Research*, vol. 50, no. 9, pp. 1250–1265, 2012.

[20] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 779–788. [Online]. Available: https://doi.org/10.1109/CVPR.2016.91

[21] M. Kaikai, L. Shijin, L. Jimin, and Y. Yufeng, "Comparative study of data-driven intelligent flood forecasting methods for small and medium-sized rivers," *Journal of University of Science and Technology of China*, no. 9, pp. 774–779, 2016.

[22] C. T. Cheng, W. J. Niu, Z. K. Feng, J. J. Shen, and K. W. Chau, "Daily reservoir runoff forecasting method using artificial neural network based on quantum-behaved particle swarm optimization," *Water*, vol. 7, no. 8, pp. 4232–4246, 2015.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proceedings of Annual Conference on Neural Information Processing Systems*, 2014, pp. 2204–2212.

[25] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4263–4270.

[26] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6298–6306.

[27] J. Liu, G. Wang, P. Hu, L. Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3671–3680.