Rainfall-Runoff Short-Term Forecasting Method Based on LSTM

Yuan Chen and Jian Xu

CHN Energy Dadu River Big Data Services Co., Ltd, Chengdu, China
Email: yuan.chen.ag@chnenergy.com.cn; xujian134@163.com

Abstract. A rainfall-runoff forecasting method based on Long Short Term Memory (LSTM) was proposed in this study, which can extract the trend characteristics of runoff time series data through introducing daily rainfall data collected at related upstream stations and making use of the advantages of LSTM in saving long-term sequence feature information and avoid vanishing gradient, and identify the nonlinear mapping relationship in between, thus establishing a short-term runoff forecasting model. In this study, 24-hour and 5-day short-term forecasting models were established based on the runoff data collected at Danba Hydrologic Station in Dadu River Basin and historical rainfall collected at three upstream stations (Xiaojin, Dawei, and Fubian River). The experimental results showed that the forecasting models performed well during the inspection period. In 24-hour forecasting, RMSE was 98.016 and MAE was 45.709, which were 73.993 and 32.699, respectively in 5-day forecasting, indicating better performance and increased forecasting accuracy than simple LSTM model.

Keywords. Rainfall-runoff forecasting; long short-term memory; short-term forecasting.

1. Introduction

In hydroelectric generation systems, precise forecasting of runoff is of great significance to planning and management activities of water resources such as flood control and economic dispatching. Runoff, the flow in a catchment area within a period, can reflect the influence of physiographic factors on the process of runoff formation in a basin. With the expansion of the scope of human economic activities and rise of population, the linearization and hardening of the river network would directly affect the runoff, and increased the randomness and nonlinearity, thus making it difficult to forecast the runoff [1].

In the past few decades, many researchers studied the forecasting of the trend of runoff changes and the related influencing factors. Runoff forecasting methods are generally divided into two categories, namely, the traditional method based on physical causes and the emerging method based on machine learning, the former of which is weak in reflecting random fluctuations in the mid- and long-term runoff based on physical phenomena, while the latter is a new and effective method for runoff forecasting that can explore the natural laws of hydrological activities based on historical data accumulated by various hydrologic stations, and avoid the weaknesses of the lack of mathematical models describing physical causes. The multiple linear regression (MLR) was adopted to predict the runoff of the Ogoki River located in Province of Ontario, Canada [2]. Autoregressive moving average (ARMA) based Gaussian function were used to get the prediction intervals of monthly river flow [3]. For overcoming the shortcomings of MLR and ARMA models in the process of runoff forecasting, such as the difficulties in identification of model parameters and collecting the non-linear features of runoff time sequence, Wang et al. (2006) forecasted the runoff-based time sequence with the artificial
neural network (ANN) [4]. And the support vector machine (SVM) was also applied to runoff forecasting and confirmed its effectiveness in small-sample learning [5].

The traditional statistical machine learning algorithm performed well in processing small-sample data, but with the increasing of hydrological statistical data and discovery of the related features, the traditional methods cannot meet the requirements of training time and forecasting effect. The big data and deep learning technologies can process and analyze time sequence problems with more data and measurement, thus providing a theoretical and technical basis for improving the accuracy of runoff forecasting.

In recent years, the LSTM network has been successfully applied to the identification of speed, handwriting and speech, and traffic flow forecasting by virtue of its advantages such as strong non-linear forecasting ability, high rate of convergence, and long-term correlation of time sequence. In this paper, a rainfall-runoff forecasting method based on LSTM was proposed, to scientifically analyze the mapping relationship between rainfall and runoff based on long-term memory characteristics of time sequence of runoff, and increase the accuracy of runoff forecasting based on the geographic information.

2. LSTM
Long Short Term Memory (LSTM), initially proposed by Hochreiter and Schmidhuber [6], is a recurrent neural network (RNN) making use of the characteristics of input sequence, with the input of text, voice, and time sequence. With special design architecture, LSTM can overcome the problem of vanishing gradient [7] in general machine learning algorithms such as RNN. LSTM is composed of self-circulating memory cells, which can convert the historical state and historical input into its internal state. There are three special gates in the modular construction of LSTM, namely the input gate, forget gate and output gate. Equations (1)-(6) show the working modes of these gates.

\[
\begin{align*}
    i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\
    f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \\
    o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \\
    g_t &= \sigma(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
    h_t &= o_t \odot \Phi(c_t)
\end{align*}
\]

where \(i, f, o\) respectively refer to input gate, forget gate and output gate. \(\sigma\) activation function compresses the output of these gates between 0 and 1. \(x_t\) is current time step input, \(h_{t-1}\) is the hidden state of the last time step. Forget gate defines how many previous states \((h_{t-1})\) will be passed; input gate defines how many current inputs \((x_t)\) will be passed; and output gate defines how many hidden states will be exposed to the next memory cell. \(g\) is calculated based on the \(x_t\) and \(h_{t-1}\). \(c_t\) represents the current cell state, and can be calculated with Equation (5) based on current cell state, forget gate, previous state, current input and \(g\).

As shown in figure 1, there is a dynamic change in self-loop weight due to the addition of input threshold, forget threshold and output threshold to LSTM. When the model parameters are fixed, the integration scale at different times can be dynamically changed, thus avoiding vanishing gradient or dilating gradient.
3. Data Pre-processing and Evaluation Indicators

3.1. Survey Region
In this paper, Danba Hydrologic Station in Dadu River Basin and its controlled area (52,738 km², accounting for 68% of total area of the Dadu River Basin) were taken as the subjects. It has been planned to construct 28 cascaded hydropower plants in the Dadu River Basin, 14 of which have been developed and put into operation, and all are distributed in lower reaches of Danba section. With the commissioning and operation of these cascaded hydropower plants, great changes have taken place in the hydraulic connection of the original river, and the regulation and storage capacity of the hydropower plants has significantly affected the runoff changes. Geographically, Danba Hydrologic Station, as an important upstream node of the Dadu River Basin, can still be regarded as a natural watercourse, whose runoff change can directly reflect the flow change in the source area to a large extent, and determine the base flow of the downstream cascaded hydropower plans; therefore, it is of guiding significance to the control of overall situation of the Dadu River.

3.2. Sample Data
The data set used in this study collected the 24-hour runoff at Danba Hydrologic Station and other 3 upstream stations (Xiaojin, Fubian River, and Dawei) from 2012 to 2016, involving date and time, runoff of Dadu River at Danba Hydrologic Station, and rainfall at Xiaojin, Fubian River, and Dawei Stations. The list of complete functions in the original data is shown in table 1:

| Number | Field     | Interpretation                  |
|--------|-----------|---------------------------------|
| 1      | APPLIEDTIME | Time                            |
| 2      | DATA      | Runoff                          |
| 3      | XIAOJIN   | Rainfall of Xiaojin             |
| 4      | FUHEBIAN  | Rainfall of Fubian River        |
| 5      | DAWEI     | Rainfall of Dawei Stations      |
There was a total of 43,848 articles of data, and each field of the data (part) was visualized with a visualization tool, as shown in figure 2.

![Figure 2. Time variation characteristics of original data.](image)

### 3.3. Data Standardization

Since data standardization is the basis of data mining, different evaluation indicators would have different measurement and measurement unit, thus affecting the results of data analysis. For eliminating the dimensional influence between indicators, all data should be standardized to resolve the problem of comparability between data indicators. After standardization of the raw data, the relevant indicators would be at the same order of magnitude, making it suitable for comprehensive comparative evaluation. In this paper, min-max is adopted for data standardization [8].

The min-max standardization method, as the linear conversion of raw data, can map the result values to the range of [0-1] through the following conversion function [9]:

\[
X^* = \frac{X - \text{MIN}}{\text{MAX} - \text{MIN}}
\]  

where, MAX refers to the maximum value and MIN refers to the minimum value of the sample data.

In this study, the training flow dataset of deep neural networks based on LSTM was used to forecast the flow data at the next time point or several time points with the historical data. The list of raw data should be converted into a data matrix with the following equation, so as to facilitate deep neural network training and learning, as well as its forecasting. The formula for converting a data list into a data matrix is as follows:

\[
\text{input}_t = [x_{t-d}, x_{t-d+1},..., x_t]
\]  

where, input refers to the input of an item from the dataset after conversion, d refers to the length of historical data required, t refers to a certain time point. In this paper, historical data were used for modeling, and 1*4 data points (the runoff of the day, and daily rainfall of the three upstream stations) were used as the inputs, to forecast the runoff of the next 5 days.
3.4. Evaluation Indicators

In this paper, the runoff forecasting method based on LSTM was adopted to compare the advantages and disadvantages of final forecasting. In order to better evaluate the forecasted results based on the LSTM model, RMSE, MAE and R2 were used to represent the difference between the forecasted and measured values, thus evaluating the quality of the forecasted values:

$$E_{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (R_{io} - R_{if})^2}$$  \hspace{1cm} (9)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|$$  \hspace{1cm} (10)

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$  \hspace{1cm} (11)

4. Modeling and Training

The 24-hour runoff at Danba Hydrologic Station and daily rainfall of the three upstream stations collected from 2012 to 2015 were used for modeling, and the data collected in 2016 were used for testing the model and evaluating the forecasting accuracy. Traditional runoff forecasting models based on historical data can only forecast the runoff of the next hour with the data of the previous hour, or forecast the total runoff of the following five days with the total runoff of the current day, which is known as runoff-based LSTM forecasting. In this paper, the historical data of rainfall of the three upstream stations were innovatively introduced to construct an LSTM model, as shown in figure 3, which can forecast the runoff of the next hour and the following five days based on the runoff of the current hour and rainfall at the three upstream stations.

![Figure 3. Model structure.](image)

Four LSTM models were constructed for performance comparison: 24-hour runoff LSTM model, 24-hour rainfall-runoff LSTM model, 5-day runoff LSTM model, and 5-day rainfall-runoff LSTM model.

Model 1: 24-hour runoff LSTM model. Based on the runoff of the previous hour (t-1), it can be used to forecast the runoff of the next hour (t);

Model 2: 24-hour rainfall-runoff LSTM model. Based on the runoff of the previous hour (t-1) and weather conditions (VAR1, VAR2, VAR3), it can be used to forecast the runoff of the next hour (t);

Model 3: 5-day runoff LSTM model. Based on the total runoff of the previous day (T-1), it can be used to forecast the total runoff of the next 5 days (T, T+1, T+2, T+3, T+4);
Model 4: 5-day rainfall-runoff LSTM model. Based on the total runoff of the previous day (T-1) and weather conditions (VAR1, VAR2, VAR3), it can be used to forecast the total runoff of the next 5 days (T, T+1, T+2, T+3, T+4).

Four models were constructed and trained with the optimizer of Adam [10] based on historical data from 2012 to 2015. In figure 4, the training losses of the four models are displayed. The number of iterations of the two 24-hour models above is 60, and the MAE curves of training are shown in figures 4a and 4b; while the number of iterations of the two 5-day models below is 250, and the MAE curves of training are shown in figures 4c and 4d. It can be concluded that the convergence speed of the rainfall-runoff models was accelerated, with smaller forecasting errors.

![Figure 4](image)

**Figure 4.** Experimental results of model prediction.

5. Experimental Analysis

5.1. Comparative Analysis of the Results

5.1.1. Experiment Results of the 24-Hour Runoff LSTM Model. In the first LSTM-based experiment, the total runoff of the previous hour (t-1) was used to forecast the total runoff (t) of the next hour. As shown in figure 5, in order to display the difference between the forecasted values and measured values more clearly and intuitively, four days of forecasted and measured values were selected for displaying; the RMSE of the forecasted daily runoff in 2016 was 115.397.

5.1.2. Experiment Results of the 24-Hour Rainfall-Runoff LSTM Model. In the second LSTM-based experiment, the total runoff of the previous hour (t-1) and daily rainfall of the three upstream stations (VAR1, VAR2, VAR3) were used to forecast the total runoff (t) of the next hour. As shown in figure 6, in order to display the difference between the forecasted values and measured values more clearly and intuitively, four days of data were selected for displaying; the RMSE of the forecasted daily runoff in 2016 was 98.016.
Figure 5. Experimental results of model 1 (four days).

Figure 6. Experimental results of model 2 (four days).

5.1.3. Experiment Results of the 5-Day Runoff LSTM Model. In the third LSTM-based experiment, the total runoff of the previous day (t-1) and the total runoff (t, t+1, t+2, t+3, t+4) of the next five days were used for forecasting. The overall forecasting results of the model (figure 7) showed the five-day forecasted values and measured values in 2016. In order to display the difference between the forecasted values and measured values more clearly and intuitively, four sections with the length of 20 were randomly selected from the annual forecasted data in figure 7 and zoomed in, as shown in figure 8, which indicated the good performance of the runoff model in forecasting the 5-day runoff; the RMSE of the forecasted daily runoff in 2016 was 97.531.
5.1.4. Experiment Results of the 5-Day Rainfall-Runoff LSTM Model. In the fourth LSTM-based experiment, the total runoff of the previous day (T-1) and daily rainfall of the three upstream stations (VAR1, VAR2, VAR3) were used to forecast the total runoff (T) of the next five days. The overall forecasting results of the model (figure 9) showed the five-day forecasted values and measured values in 2016. In order to display the difference between the forecasted values and measured values more clearly and intuitively, four sections with the length of 20 were randomly selected from the annual forecasted data in figure 9 and zoomed in, as shown in figure 10, which indicated the good performance of the rainfall-runoff model in forecasting the 5-day runoff; the RMSE of the forecasted daily runoff in 2016 was 73.993.
5.2. Analysis of Experimental Results

The error precision of the four models was evaluated with RMSE, MAE and R-square, as shown in table 2:

| Model | RMSE   | MAE   | R-square |
|-------|--------|-------|----------|
| 1     | 115.397| 52.984| 0.968    |
| 2     | 97.531 | 43.287| 0.978    |
| 3     | 98.016 | 45.709| 0.977    |
| 4     | 73.993 | 32.699| 0.987    |
As shown in table 2, the rainfall-runoff LSTM model introduced in this paper had smaller RMSE and MAE of 24-hour and 5-day forecasting results than the runoff LSTM model, which were decreased by 15.06% and 13.3% respectively. Compared with the runoff model, the errors in RMSE and MAE of the rainfall-runoff model in 5-day forecasting were respectively reduced by 24.13% and 24.46%. As for R-square coefficient (the larger the better), the rainfall-runoff model was superior to the runoff model in terms of 24-hour and 5-day forecasting results. Table 2 showed that the 5-day forecasting was superior to 24-hour forecasting, the reason must be that the dataset used in this paper was more regular in multi-day distribution, and the model was more suitable for 5-day forecasting.

6. Conclusion
In this paper, the rainfall-runoff forecasting method was used to construct forecasting models based on LSTM by making full use of the long-term memory of the runoff time sequence. The feature extraction based on the mapping relationship between rainfall and runoff improved the accuracy of runoff forecasting. Finally, based on the historical data of runoff and rainfall collected at Danba Hydrologic Station and three upstream stations, the 24-hour runoff and 5-day runoff were forecasted with the LSTM model; and the results verified the effectiveness of the method. In view of the effects of various factors on runoff, such as geomorphological characteristics of the basin, land cover classification, and soil physical properties, it is necessary to construct more accurate forecasting models.

References
[1] Xu J, Luo W and Huang Y 2019 Dadu River Runoff Forecasting via Seq2Seq Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science (AICS 2019) pp 513-517.
[2] Seidou O and Ouarda T B M J 2007 Recursion-based multiple changepoint detection in multiple linear regression and application to river streamflows Water Resources Research 43 (7) w07404.
[3] Anderson P L, Meerschaert M M and Zhang K 2013 Forecasting with prediction intervals for periodic autoregressive moving average models Journal of Time Series Analysis 34 (2) 187-193.
[4] Wang W, Van Gelder P, Vrijling J K, et al. 2006 Forecasting daily streamflow using hybrid ANN models Journal of Hydrology 324 (1-4) 383-399.
[5] Sharifi A, Dinpashoh Y and Mirabbasi R 2017 Daily runoff prediction using the linear and nonlinear models Water Science and Technology wst2017234.
[6] Hochreiter S and Schmidhuber J Long 1997 Short-term memory Neural Computation 9 (8) 1735-1780.
[7] Hochreiter S 2015 Vanishing gradient problem.
[8] Grannis S J, Xu H, Vest R, Kasthuriratnhe S, Bo N, Moscivitch B, Torkzadeh R and Rising J 2019 Evaluating the effect of data standardization and validation on patient matching accuracy Journal of the American Medical Informatics Association 5 (2019) 447-456.
[9] Prasad S and Prasad R 2020 Data standardization and scaling technique for the implementation of human bond communication Wireless Personal Communications: An International Journal 110 959-972.
[10] Kingma D and Ba J 2014 Adam: A method for stochastic optimization arXiv preprint arXiv:1412.6980.