# The LSTM Based Trend Prediction Model: A Case Study of Beijing COVID-19 Epidemic Data

Ruiling Zhao[1] and Linchao Yang[2](✉)

[1] Discipline Inspection and Audit Division, Aviation General Hospital, Beijing 100012, China
[2] School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
yanglinchao@buaa.edu.cn

**Abstract.** The paper tries to build a COVID-19 trend prediction model with data mining methods and deep learning algorithm by taking Beijing COVID-19 epidemic data as an example. Currently, the Corona Virus Disease 2019 (COVID-19) is spreading around the world, posing a huge threat to the human society. By studying the epidemic curve, the overseas imported cases and incidence rate in different age groups, we draw the conclusion that the overseas imported cases are the main reason for the second rise of the cumulative confirmed cases curve and the incidence rate of the older people over 60 in Beijing is the highest. To support the prevention and control of the COVID-19, we extract the daily cumulative confirmed cases in Beijing from January 20th, 2020 to April 26th, 2020 to establish a trend prediction model based on LSTM method (Long Short-Term Memory). And we compared the proposed LSTM based prediction model with the Support Vector Regression (SVR) based prediction model and the Autoregressive Integrated Moving Average (ARIMA) based prediction model. The result shows the effectiveness of the proposed model.

**Keywords:** COVID-19 · Data Analysis · LSTM · Prediction Model

## 1 Introduction

COVID-19 is a highly infectious and hazardous coronavirus [1]. The COVID-19 firstly broke out in Wuhan, China at the end of 2019 [2]. Soon after, the COVID-19 rapidly spread all over the world and was difficult to control. The World Health Organization (WHO) declared the COVID-19 a public health emergency of international concern on January 30th, 2020 [3]. Globally, as of 17the May, 2021, there have been more than one hundred and sixty-two million confirmed cases of COVID-19, including about three million deaths. The sudden outbreak of COVID-19 has severely affected the world economy and people's livelihood. Although the COVID-19 epidemic situation is becoming better under the efforts of all countries in the world, we still need to watch out for the resurgence of the COVID-19 epidemic and sum up experience from the history of COVID-19 development. Among them, the most critical two points are analyzing COVID-19 data to

G. Ali et al. (Eds.): ISEMSS 2022, ASSEHR 687, pp. 672–681, 2022.
https://doi.org/10.2991/978-2-494069-31-2_82

mine information, and establishing prediction model based on data to guide COVID-19 prevention and control.

Data mining methods and deep learning algorithms play an important role in epidemic analysis and forecasting [4]. Since the outbreak of the COVID-19, researchers have extensively analyzed the COVID-19 data. Chen et al. [5] analyzed the data of all COVID-19 confirmed cases in Wuhan Jinyintan Hospital from January 1st, 2020 to January 20th, 2020 and found these COVID-19 cases were mostly observed among elderly people. Ziyad et al. [6] analyzed the data from the national healthcare databases of the America Department of Veterans Affairs to study the possible sequelae caused by COVID-19. Furthermore, researchers have set up many prediction models to predict the development COVID-19 trend. At present, the existing prediction models are mostly based on machine learning methods [7]. Machine learning methods such as SVR [8] and ARIMA [9] are used to do COVID-19 trend prediction. Punn et al. [10] retrieved the day to day prevalence data of COVID-19 from January 22th, 2020, to April 1st, 2020, from the official repository of Johns Hopkins University and developed a prediction model based on SVR. Gupta et al. [11] analyzed the COVID-19 data in India and built a prediction model of the COVID-19 trend in India based on ARIMA. Although these models have achieved good results, SVR is not very suitable for time series and ARIMA is a simple linear time series model. Fortunately, LSTM networks proposed by Hochreiter and Schmidhuber [12] can solve the problems of SVR and ARIMA well and is widely used in time series modeling. Kratzert et al. [13] used LSTM to do rainfall-runoff modelling and effectively improved the accuracy of the prediction. Therefore, aiming at improving the accuracy of prediction model and providing support for COVID-19 epidemic prevention and control, this study intends to take Beijing as an example and apply LSTM to establish a trend prediction model on the basis of full analysis and mining of epidemic data.

The paper analyses the Beijing COVID-19 epidemic data details in Sect. 2. Then the Sect. 3 is the process of prediction modelling and results analysis. The result shows the proposed prediction model based on LSTM is better than the SVR and ARIMA prediction model.

## 2 Data Source and Statistical Analysis

The data we collected is extracted from verified sources of the official repository of Beijing Municipal Health Commission, which including the day to day prevalence information of COVID-19 from January 20th, 2020, to April 26th, 2020. And we extracted and integrated the information to form a data table which covers the daily cumulative confirmed cases, the daily existing confirmed cases, the daily cumulative cured cases, the daily cumulative dead cases, the daily new confirmed cases, the daily local new confirmed cases, the daily new overseas imported cases and age distribution of confirmed cases. On this basis, we analyzed the Beijing COVID-19 epidemic data in detail.

To study the basic situation of Beijing COVID-19 epidemic development, we draw the epidemic time curve as shown in Fig. 1 according to the daily cumulative confirmed cases, the daily existing confirmed cases, the daily cumulative cured cases and the daily cumulative dead cases of Beijing from January 20th, 2020, to April 26th, 2020. As can

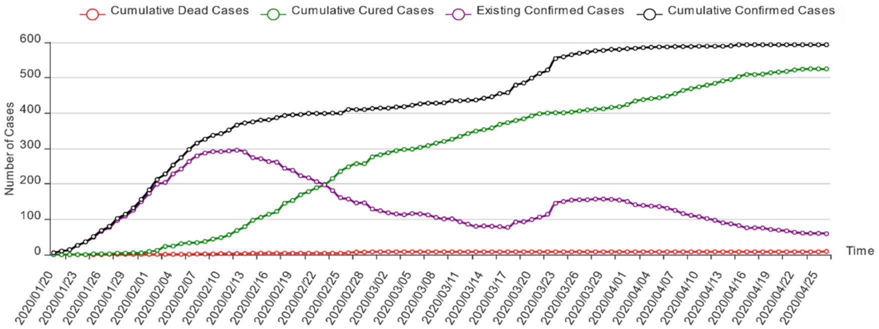

**Fig. 1.** Beijing COVID-19 Epidemic Time Curve

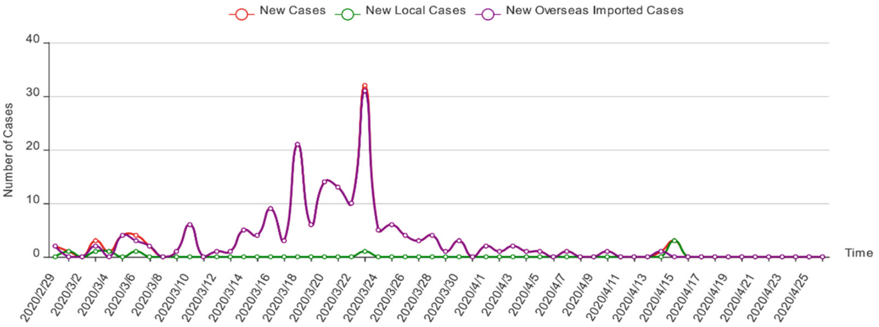

**Fig. 2.** Beijing COVID-19 Daily New Cases Curve

be seen from Fig. 1, as of April 26, the epidemic situation in Beijing has been basically controlled. The cumulative confirmed cases curve growth rate has basically reached 0, the existing confirmed cases has fallen into a lower value, the cumulative cure cases curve has been rising, and the cumulative dead cases has been always kept at a low level.

Furthermore, we also can find the cumulative confirmed cases curve has first reach platform period on February 19th but quickly rose again after February 28th. To find out the reason, we draw the daily new cases curve according to the daily new confirmed cases, the daily new local confirmed cases and the daily new overseas imported cases of Beijing from February 29th, 2020 to April 26th, 2020. It can be found from Fig. 2 that the new confirmed cases curve and the new overseas imported cases curve almost coincide, while the new local confirmed cases curve has been always kept at 0. Therefore, we can see clearly that the overseas imported cases are the main reason for the second rise of the cumulative confirmed cases curve.

To further study the overseas imported cases of Beijing, we counted the cumulative overseas imported cases before April 26th, 2020 by cases exported countries and continents, and draw a flow chart of cumulative overseas imported cases based on the exported countries with more than three cumulative overseas imported cases. The flow chart is shown in Fig. 3. Based on Fig. 3, it can be seen that the most of Beijing cumulative overseas imported cases are from Europe and North America, and the imported cases from

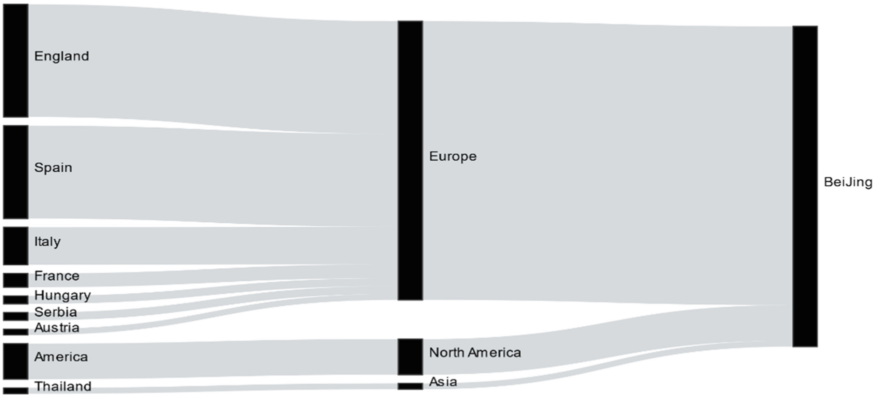

**Fig. 3.** Flow Chart of Cumulative Overseas Imported Cases

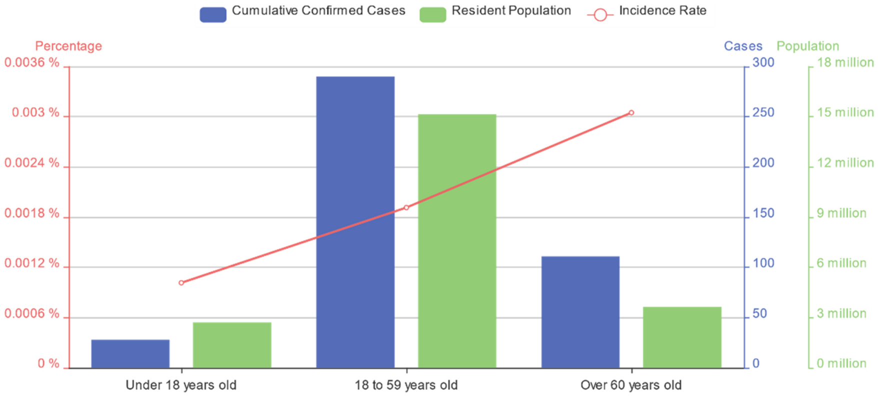

**Fig. 4.** Cumulative Confirmed Cases and Incidence Rate in Different Age Groups

Europe account for more than 80%. The imported cases from Europe are distributed in many European countries, mainly in England, Spain and Italy. The imported cases from England are the most, while the imported cases from North America are mainly from America.

In addition, we also counted the cumulative confirmed cases before April 26th, 2020 in different age groups and collected Beijing resident population data from Beijing Municipal Bureau of Statistics to analyze incidence rate. And the specific results are shown in Fig. 4. As shown in Fig. 4, the cumulative confirmed cases in Beijing are mainly distributed in the age of 18–59 and over 60 as of April 26th, and the number of the cumulative confirmed cases in the age of 18–59 is the largest for the reason of the largest resident population. And the incidence rate of the older people over 60 is the highest, which is the same as Chen's research results [5].

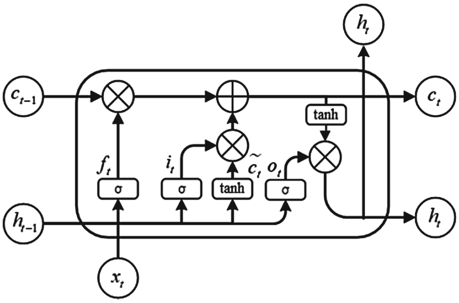

**Fig. 5.** LSTM Memory Cell Unit Structure

# 3 LSTM Based Prediction Model and Results Analysis

The study builds a more accurate time series prediction model for the cumulative confirmed cases in this section. The proposed prediction model uses LSTM deep learning method instead of traditional ARIMA and SVR methods, which can effectively improve the performance of the prediction model. At the same time, to verify the superiority of the proposed model, a comparative experiment is designed.

## 3.1 LSTM Prediction Modeling

The LSTM network consists of an input layer, a hidden layer, and an output layer like the RNN, with a state unit $c$ added on top of it, and the information is allowed to selectively influence the state of the LSTM network at each moment through forgetting gates, input gates, and output gates. The internal structure of the LSTM is shown in Fig. 5. The LSTM unit has 3 inputs at time $t$: the input value of the network at the current time $x_t$; the output value of the LSTM hidden layer at the previous time $h_{t-1}$; and the state of the unit at the previous time $c_{t-1}$. The LSTM unit has 2 outputs at time $t$: the output value of the hidden layer at the current time $h_t$ and the state of the unit $c_t$.

The specific operation steps of LSTM are as follows:

First, the forgetting door:

$$f_t = \sigma(W_f * [x_t, h_{t-1}] + b_f) \tag{1}$$

where$f_t$ is the scale factor of the control forgetting information at the current moment; $\sigma$ is the *sigmoid* activation function, which serves to smoothly map the data obtained after the operation to the (0, 1) interval, which exactly corresponds to the degree of switching of the control gate; $W_f$ is the weight matrix of the forgetting gate; $b_f$ is the bias term of the forgetting gate.

$$[W_f]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} = \begin{bmatrix} W_{fh}, W_{fx} \end{bmatrix}\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{2}$$

where the weight matrix $W_f$ consists of two weight matrices $W_{fh}$ and $W_{fx}$, corresponding to the input terms $h_{t-1}$ and $x_t$, respectively. The function of the forgetting gate is to let the LSTM forget the previously useless information, assuming that the input time series of the LSTM at moment $t$ $x_t = [x_t^1, x_t^2, \cdots, x_t^n]$ and the output of the implicit layer at moment $th_{t-1} = [h_{t-1}^1, h_{t-1}^2, \cdots, h_{t-1}^n]$, the coefficients that determine the proportion of forgotten information are obtained by using both as inputs through Eq. (2), information on dimensions where $f_t$ is close to 0 will be completely forgotten, while information on dimensions where $f_t$ is close to 1 will be completely retained.

Second, the input door:

$$i_t = \sigma(W_i * [x_t, h_{t-1}] + b_i) \tag{3}$$

$$[W_i]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} = [W_{ih}, W_{ix}]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{4}$$

where $i_t$ is the scale factor of the control inputting information at the current moment; the weight matrix $W_i$ consists of two weight matrices $W_{ih}$ and $W_{ix}$, corresponding to the input terms $h_{t-1}$ and $x_t$, respectively.

$$\underset{t}{\overset{\sim}{c}} = \tanh(W_c * [h_{t-1}, x_t] + b_c) \tag{5}$$

$$[W_c]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} = [W_{ch}, W_{cx}]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{6}$$

where $c_t$ is the candidate cell state at the current moment; tanh is the activation function, which differs from the *sigmoid* activation function in that the tanh activation function maps the data smoothly to the (-1,1) interval; $W_c$ is the weight matrix of the candidate cell state, consisting of the $W_{ch}$ and $W_{cx}$ weight matrices corresponding to $h_{t-1}$ and $x_t$; and $b_c$ is the bias term of the candidate cell state.

$$c_t = f_t * c_{t-1} + i_t * c_t \tag{7}$$

The LSTM network needs to be supplemented with the latest information from the current input, and this process is done through input gates. First, the input gate gets the scale factor $i_t$ controlling the input information through Eq. (4) to decide which information in the candidate cell state is added to the previous cell state $c_{t-1}$ to generate the new state $c_t$. Second, the candidate cell state $c_t$ is obtained by using $h_{t-1}$ and $x_t$ as input through Eq. (6). Finally, the current cell state $c_t$ can be obtained through Eq. (8).

Finally, the output door:

$$o_t = \sigma(W_o * [x_t, h_{t-1}] + b_o) \tag{8}$$

$$[W_o]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} = [W_{oh}, W_{ox}]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{9}$$

where $o_t$ is the scale factor of the control outputting information at the current moment; the weight matrix $W_o$ consists of two weight matrices $W_{oh}$ and $W_{ox}$, corresponding to the input terms $h_{t-1}$ and $x_t$, respectively.

$$h_t = o_t * \tanh(c_t) \tag{10}$$

**Table 1.** LSTM Model Parameters Selection

| Parameters Name | Parameters Selection |
|---|---|
| Input Layer Units | 1 |
| Output Layer Units | 1 |
| Hide Layer | 1 |
| Hide Layer units | 100 |
| Timestep | 1 |
| Epochs | 60 |
| Batch Size | 11 |
| Activation Function | Relu |
| Loss Function | *Min(Mean Absolute Error)* |
| Optimizer | Adam |
| Dropout | 0.1 |

The LSTM generates the output at the current moment by means of an output door. The parameter $o_t$, which controls the degree of switching of the output door, is first obtained through Eq. (9), and then the latest cell state $c_t$ is multiplied with the corresponding parameter of $o_t$ through Eq. (10) to obtain the output $h_t$ at the current moment after the activation function tanh operation.

### 3.2 Result Analysis

In this study, the daily cumulative confirmed cases in Beijing from January 20th, 2020 to April 26th, 2020 is extracted to use for LSTM prediction modeling. According to the approximate proportion 4:1, the data is divided into training set and test set. Specifically, the data from January 20th, 2020 to April 5th, 2020 is used for model training, and the data from April 6, 2020 to April 26, 2020 is used for model testing. The parameters selection of LSTM neural network model is shown in Table 1.

To verify the effectiveness of the proposed model, we compare the LSTM based prediction model proposed in this study with the SVR based prediction model proposed by Punn [10] and the ARIMA based prediction model proposed by Gupta [11]. We first compare the prediction results on the test set of the three models, and the specific results are shown in Fig. 6. It can be seen from Fig. 6 that the LSTM based prediction model has the best fitting effect on the test set, followed by the SVR based prediction model, and the ARIMA based prediction model has the worst fitting effect.

In addition, in order to more objectively compare the performance of the three prediction models, the evaluation indexes in this study mainly include $\boldsymbol{R^2}$ score, root mean square error (***MAE***) and mean absolute error (***RMSE***), where $\boldsymbol{R^2}$ reflects the degree of fitting of the prediction results, with $\boldsymbol{R^2}$ score equal to 1 when fully fitted, and the larger the value, the better the fitting effect; ***MAE*** reflects the error between the predicted and true values, and the smaller the value, the smaller the error; and ***RMSE*** reflects the

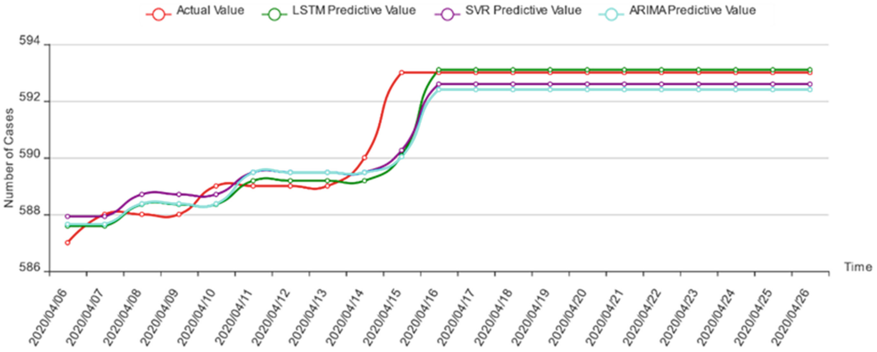

**Fig. 6.** Fitting Value of Three Prediction Models on Test Set

volatility of the model prediction, and again the smaller the value, the less volatility. The prediction model based on LSTM has the largest $\boldsymbol{R^2}$, so has the best fitting effect. In terms of ***MAE***, the LSTM based prediction model is the smallest, so has the smallest error, and it has a greater improvement than the other two models. In terms of ***RMSE***, the prediction model based on LSTM is also the smallest, so has the smallest volatility. Thus, it can be concluded that the prediction model based on LSTM proposed in this paper is better than the prediction model based on SVR and ARIMA.

## 4 Conclusion

The COVID-19 is continuing to spread throughout the world. To support the prevention and control of the COVID-19, this study takes Beijing as an example to analyze the development rules and characteristics of the epidemic situation in Beijing from the data level and multi angles and establishes a trend prediction model based on LSTM.

In terms of Beijing COVID-19 epidemic data analysis, we analyze the epidemic curve, the overseas imported cases and incidence rate in different age groups, and finally reached the following results:

1) The overseas imported cases are the main reason for the second rise of the cumulative confirmed cases curve.
2) The overseas imported cases in Beijing mainly come from Europe and North America, mainly concentrate in England, Spain, Italy and America.
3) The incidence rate of the older people over 60 in Beijing is the highest.

In terms of trend prediction modeling, we first extract the daily cumulative confirmed cases in Beijing from January 20th, 2020 to April 26th, 2020 for data modeling, and then select LSTM method for modeling. Finally, we compare the proposed LSTM based prediction model with the SVR based prediction model [10] and the ARIMA based prediction model [11]. The following conclusions can be drawn from the experiment:

1) For the data set used in this study, the LSTM based prediction model has the best fitting effect on the test set.

2) The proposed prediction model based on LSTM is better than the other two models in $\boldsymbol{R^2}$ , ***MAE*** and ***RMSE***. So it has higher goodness of fit, smaller error and smaller volatility.

**Authors' Contributions**

Data curation, formal analysis, Ruiling Zhao; methodology, supervision, Linchao Yang; writing, review and editing, Ruiling Zhao & Linchao Yang.

**Acknowledgments.** This research was funded by the National Natural Science Foundation of China (Grant No.71971013). The study is also sponsored by the Fundamental Research Funds for the Central Universities (YWF-20-BJ-J-943).

## References

1. Yuki K, Fujiogi M and Koutsogiannaki S. Covid-19 Pathophysiology: A Review. Clinical Immunology, 2020, 215: 108427.
2. Li Q and Guan X et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine, 2020, 13(382): 1199~1207.
3. Velavan T P and Meyer C G. The Covid-19 Epidemic. Tropical Medicine and International Health, 2020, 3(25): 278~280.
4. Kalipe G, Gautham V, Behera R K. Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis[C]//2018 International Conference on Information Technology (ICIT). IEEE, 2018: 33–38.
5. Chen N and Zhou M et al. Epidemiological and Clinical Characteristics of 99 Cases of 2019 Novel Coronavirus Pneumonia in Wuhan, China: A Descriptive Study. The Lancet, 2020, 10223(395): 507~513.
6. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19[J]. Nature, 2021, 594(7862): 259-264.
7. Gupta R and Pandey G et al. Machine Learning Models for Government to Predict Covid-19 Outbreak. Digital Government: Research and Practice, 2020, 4(1): 1~6.
8. Chen Y and Xu P et al. Short-Term Electrical Load Forecasting Using the Support Vector Regression (SVR) Model to Calculate the Demand Response Baseline for Office Buildings, Applied Energy, 2017, 195: 659~670.
9. Aasim Singh S N and Mohapatra A. Repeated Wavelet Transform Based Arima Model for Very Short-Term Wind Speed Forecasting. Renewable Energy, 2017, 136: 758~768.
10. Punn N S Sonbhadra S K and Agarwal S. Covid-19 Epidemic Analysis Using Machine Learning and Deep Learning Algorithms, MedRxiv, 2020.
11. Gupta R and Pal S K. Trend Analysis and Forecasting of Covid-19 Outbreak in India, MedRxiv, 2020.
12. Hochreiter S and Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 8(9): 1735~1780.
13. Kratzert F and Klotz D et al. Rainfall-Runoff Modelling Using Long Short-Term Memory (Lstm) Networks. Hydrology and Earth System Sciences, 2018, 11(22): 6005~6022.