Hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM

Zheng Wang
College of Computer and Information
Hohai University
Nanjing, China
1275708168@qq.com

Yuansheng Lou
College of Computer and Information
Hohai University
Nanjing, China
Wise.lou@163.com

***Abstract*—Hydrological time series is affected by many factors and it is difficult to be forecasted accurately by traditional forecast models. In this paper, a hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM is proposed. The model first removes the interference factors in the hydrological time series by wavelet de-noising, and then uses ARIMA model to fit and forecast the de-noised data to obtain the fitting residuals and forecast results. Then we use the residuals to train LSTM network. Next, the forecast error of the ARIMA model is forecasted by LSTM network and used to correct the forecast result of ARIMA model. In this paper, we use the daily average water level time series of a hydrological station in Chuhe River Basin as the experimental data and compare this model with ARIMA model, LSTM network and BP-ANN-ARIMA model. Experiment shows that this model can be well adapted to the hydrological time series forecast and has the best forecast effect.**

Keywords—hydrological time series; forecast; wavelet de-noising; ARIMA-LSTM

I. INTRODUCTION

Hydrological time series forecast is to mine the potential law of hydrological process change according to known information and forecast future hydrological data. Accurate hydrological forecasting is of great value in flood prevention and flood control, water resources planning and so on. The hydrological data in this paper mainly refer to the water level data of rivers.

There are many methods for time series forecast. The autoregressive integrated moving average (ARIMA) model proposed by Box and Jenkins is one of the most widely used time series forecast models. Many scholars have applied this model to hydrological time series analysis[1-2], but ARIMA model has the following basic shortcomings: in ARIMA model, it is assumed that the future value and the past observation values of the time series satisfy the linear relationship[3]. In fact, most of the time series data contain nonlinear relationship, which limits the scope of the application of ARIMA model.

Felix Gers[4], Yang Juanli[5], Liu Li[6], Valunjkar[7] and others used different neural network models to predict hydrological time series. Among all the neural network models, only the recurrent neural network (RNN) introduces the concept of time series into the design of network architecture, which makes it more adaptable in the analysis of time series data[8]. Yang Wei[9] used deep recurrent neural network(DRNN) for hydrological time series forecast, but DRNN also has problems such as gradient disappearance, gradient explosion, and long-term memory deficiency, which limits the use of RNN. Until the emergence of Long Short-Term Memory[10], the above problems were solved.

ARIMA model can well fit the linear relationship in the sequence, and LSTM network can mine the nonlinear relationship in the sequence. By the combination of the two models, advantages of both them can be achieved, and the complex time series can be forecasted effectively. Luo Long[11] forecasted the equivalent salt density of insulators on a tower in Guangzhou by ARIMA-LSTM model and achieved good results. However, in Luo Long's experiment, the amount of experimental data is small, and the experimental data are accurate. In the real situation, due to measurement errors, the hydrological time series contain a lot of noise, which may mask the true variation of the hydrological time series and even change the autocorrelation structure of the sequence. Therefore, it is necessary to deal with the noise of hydrological time series to improve the forecast accuracy of hydrological time series. To solve the above problems, this paper proposes a hydrological time series prediction model based on wavelet de-noising and ARIMA-LSTM. The model first eliminates the interference factors in the hydrological time series by the improved wavelet threshold de-noising method, then it uses ARIMA model to forecast the linear part of the de-noised hydrological time series and uses LSTM network to forecast the nonlinear part of the data. Finally the two forecast results are integrated to obtain the final forecast value.

II. HYDROLOGICAL TIME SERIES FORECAST MODEL

A. Wavelet threshold de-noising

Wavelet threshold de-noising is proposed by Donoho and Johnstone[12-14]. It is the most widely used method in engineering. In this theory, a one-dimensional signal containing noise can be expressed as the formula: $f(t) = s(t) + n(t)$. In the formula, $s(t)$ is the useful signal and $n(t)$ is the noise. Because the wavelet transform is a linear transform, the wavelet coefficients obtained by the discrete wavelet transform

978-1-5386-6243-4/19/$31.00 ©2019 IEEE

of $f(t)$ are still composed of two parts: one is the wavelet coefficients corresponding to the useful signal, the other are the wavelet coefficients corresponding to the noise. According to the statistical characteristics, the useful signal corresponds to the wavelet coefficient with the larger amplitude, and the noise corresponds to the wavelet coefficient with the smaller amplitude. Therefore, a suitable threshold can be found. If the wavelet coefficient is larger than the threshold, it will be considered to be caused by the useful signal, and the reservation will be given. If the wavelet coefficient is smaller than the threshold, it will be considered to be caused by the noise and discarded. Finally, we can remove the noise from the signal. The current common threshold selection rules are rigrsure, sqtwolog, minimaxi and heursure.

B. ARIMA

ARIMA model is called the autoregressive integral moving average model. It is a famous time series forecast method proposed by Box and Jenkins on the basis of autoregressive model, moving average model and autoregressive moving average model in the 1970s. It considers time series as a random sequence and approximates it by a mathematical model. When the sequence becomes stationary after the differencing process, the formula of the ARIMA model can be expressed as follows:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} \cdots + \varphi_p y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \cdots - \theta_q e_{t-q} \quad (1)$$

In the formula, p is the order of the autoregressive model, q is the order of the moving average model, e is white noise sequence, φ and θ are model parameters, and y_t is the observation value at time t.

C. LSTM

Long Short-Term Memory network is a variant of the recurrent neural network. It is composed of LSTM units. By introducing the gating mechanism and the memory cell, LSTM network can learn the long-term dependencies, and alleviate the problems caused by gradient disappearance and gradient explosion. A common LSTM unit is composed of a memory cell, an input gate, an output gate and a forget gate. The cell can remember values over arbitrary time intervals and three gates can adjust the information flowing into and out of the cell. The architecture of LSTM unit is shown in Figure 1.

The compact forms of the equations for the forward pass of an LSTM unit are:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (5)$$

$$h_t = o_t \circ \tanh(c_t) \quad (6)$$

In these equations, x_t is the input vector of the LSTM unit, h_t is the output vector of the LSTM unit, f_t is the forget gate's activation vector, i_t is the input gate's activation vector, o_t is output gate's activation vector, c_t is the cell state vector, σ is the sigmoid function, and the subscript t indexes the time step. W, U and b are weight matrices and bias vector parameters which need to be learned during training.

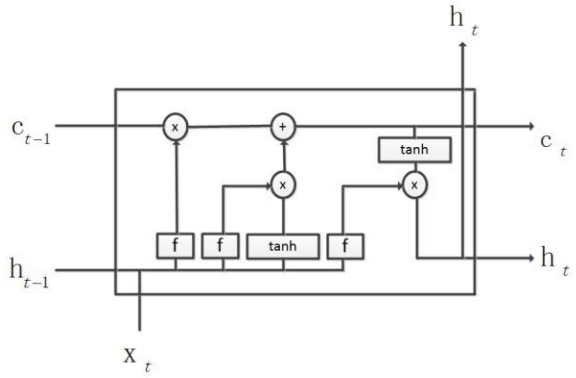

Fig. 1. LSTM unit

D. Model concept

Because of the measurement error, the hydrological time series contains a lot of noise. The existence of noise may interfere with useful signal and cover up the variation law of the hydrological time series. Therefore, the effective de-noising method can improve the forecast accuracy of the model.

Utilizing the data characteristics of hydrological time series, an improved wavelet threshold de-noising method is proposed. The variation of hydrological time series is relatively stable in the dry season and the frequency of the sequence is relatively low, contrary to that in the wet season. The noise distribution of the hydrological time series also has certain regularities. When the frequency of hydrological time series is high, the signal-to-noise ratio is relatively low. By contrast, when the frequency is low, the signal-to-noise ratio is relatively high. Therefore, after the wavelet decomposition of the hydrological time series, the thresholds of high-frequency wavelet coefficients and the low-frequency wavelet coefficients are determined by the rigrsure criterion and the heursure criterion respectively. The rigrsure criterion is conservative. It can effectively extract the weak useful signal in the high-frequency wavelet coefficients. The heursure criterion can completely remove the noise in the low-frequency wavelet coefficients. This improved wavelet threshold de-noising method is closely related to the characteristics of hydrological time series and improves the effect of de-noising.

Hydrological data is easily affected by natural conditions such as the season. Therefore, the law of hydrological data is complex and changeful, and it is difficult to describe this law with a single linear or nonlinear relationship. ARIMA model is a high precision linear time series forecast model, and LSTM network is an excellent nonlinear time series forecast method. In this paper, the combination of the two models makes them complement each other and improve the forecast accuracy of hydrological time series. The de-noised time series can be expressed as a formula: $y_t = L_t + N_t$. In the formula, y_t is the de-noised time series, L_t is the linear part of the time series and

N_t is the nonlinear residual. Firstly, the ARIMA model is used to fit and forecast the linear correlation part of the de-noised hydrological time series. Then the residuals are obtained by comparing the real data with the fitting data. The nonlinear relationship of the hydrological data is contained in the residuals. Next, we use the residuals to train LSTM network. LSTM network learns the nonlinear relationship of the data and forecasts the residual. Finally, the residual forecasted by LSTM network is used to correct the ARIMA model's forecast result to obtain the final result.

E. Modeling process

The specific modeling steps for the model are as follows:

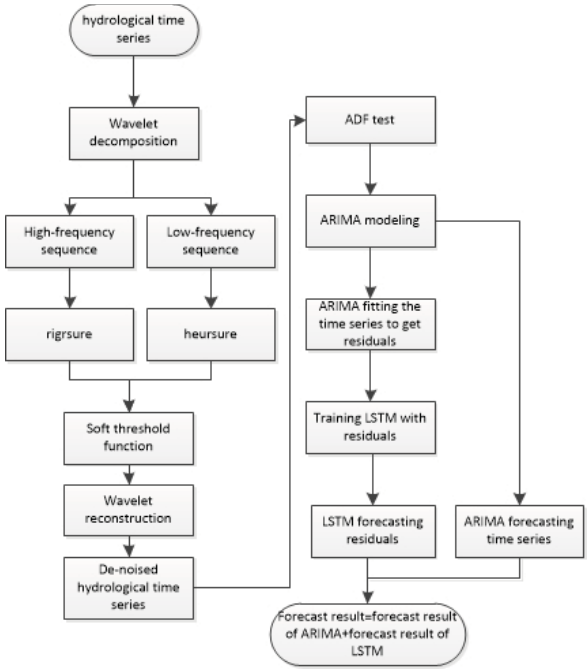

Fig. 2. Modeling flowchart

Step 1: Determine the wavelet basis function and the number of decomposition layers, and then perform wavelet decomposition on the hydrological time series to obtain the high-frequency wavelet coefficients and the low-frequency wavelet coefficients.

Step 2: Determine the thresholds of the high-frequency wavelet coefficients and the low-frequency wavelet coefficients with the rigrsure criterion and the heursure criterion respectively, and use the soft threshold function to perform threshold processing on the wavelet coefficients.

Step 3: Obtain de-noised time series y_t by wavelet reconstruction.

Step 4: Perform the stationarity test on time series y_t. If it passes, go to the next step. If it does not pass, continue to perform differencing operation on the series until it can pass the stationarity test. The frequency of differencing operations is d.

Step 5: Define the range of parameters: p and q, find the values of p and q, which make the value of Akaike Information Criterion (AIC) smallest.

Step 6: According to the determined parameters: p, q, d, establish ARIMA(p,d,q) model.

Step 7: Use ARIMA(p,d,q) to model L_t which is the linear part of y_t, consider the forecast result $\hat{L}_t$ to be the forecasted value at time t, obtain the residual e_t by comparing the original sequence with $\hat{L}_t$, which is:

$$e_t = y_t - \hat{L}_t \tag{7}$$

The nonlinear relationship in the original sequence is included in the sequence e_t.

Step 8: Use LSTM network to learn this nonlinear relationship. The nonlinear relationship can be expressed as:

$$e_t = f(e_{t-1}, e_{t-2}, \ldots e_{t-n}) + \epsilon_t \tag{8}$$

Where f is the nonlinear relationship learned by LSTM network, ϵ_t is a random error, we use LSTM network to forecast the residual e_t, the forecast result is $\hat{N}_t$;

Step 9: Obtain forecast result of ARIMA-LSTM model $\hat{y}_t$ by adding ARIMA model's forecast results and LSTM network's forecast results together, the formula is expressed as:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{9}$$

The flow chart of the model is shown in Figure 2.

III. Experiment And Analysis

A. Experiment Purpose

In order to verify the forecast accuracy of the hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM in real data, this section applies the model to the hydrological data of a site in Chuhe River Basin to test whether the model can accurately forecast. The model will be evaluated by the mean square error (MSE) and the mean absolute percent error (MAPE).

B. Data and Environment

The experimental data in this paper is the daily average water level data of a hydrological station in Chuhe River Basin from January 1, 2010 to December 31, 2015, and the actual number of the data is 2,129. The first 2000 days' data is used as training data for the model modeling, and the last 129 days' data is the test data, which is used to test the accuracy of model's forecast result. The experimental data is shown in Figure 3.

The experimental environment is Win8 64 bit operating system, the hardware platform is Intel Core i7 3612QM processor, 8G memory notebook computer, and the development tool is matlab2016a.

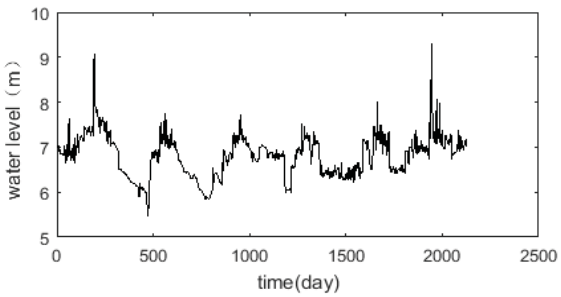

Fig. 3. Experimental data

C. Process and Analysis

The first step is to de-noise the training set. In this paper, Daubechies10 wavelet function is used as the wavelet basis function of the wavelet decomposition, and the original hydrological time series is decomposed into 6 layers. The wavelet coefficients obtained from the wavelet decomposition are shown in Figure 4. d_1, d_2 and d_3 are high-frequency wavelet coefficients, and their thresholds are determined by the rigrsure criterion. d_4, d_5 and d_6 are low-frequency wavelet coefficients, and we determine their thresholds by the heursure criterion. After calculation, the thresholds of d_1 to d_6 are 0.004, 0.012, 0.035, 0.206, 0.1 and 0.1. We make the soft threshold function as the threshold function. After threshold processing and wavelet reconstruction, the effect of de-noising is shown in Fig. 5.

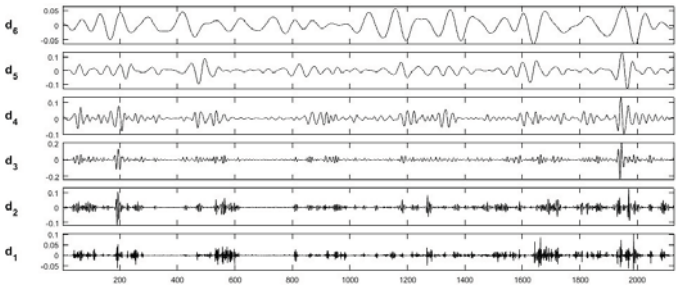

Fig. 4. Wavelet coefficients

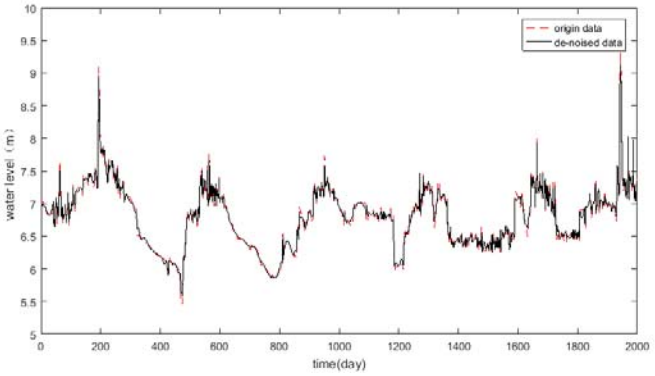

Fig. 5. The result of de-noising

It can be seen that the improved threshold de-noising method removes the glitch noise in the hydrological time series well. It eliminates the noise interference on the basis of the integrity of the sequence's useful signal, while maintaining the de-noised sequence smooth.

The ADF test is performed on the de-noised hydrological time series. The p-value is 0.007, which is much smaller than 0.05, so the sequence is stable. Since the sequence is stable, d is set to 0. We set the maximum values of p and q to 5, and find that the AIC value is the smallest when p = 1 and q = 1 by iteration calculation. So we establish ARIMA (1,0,1) model. We use ARIMA (1,0,1) to fit the training set and compares the fitted values to the real values to get the residuals of the training set.

Since ARIMA model is not suitable for long-term forecast, we forecast the test set by one-step forecast method [15] to achieve forecast results of the test set. The residuals of the test set are obtained by comparing the forecast results with the real values. The one-step forecast method means that when forecasting the data of the Nth day, we use the data of the previous days as the training sample for model modeling, and the subsequent data cannot participate in the modeling. Then, when forecasting the data of the N+1th day, the real data of the Nth day will be added to the training sample for model modeling. The forecast accuracy of the one-step forecast method is quiet good, and the mean square error of the forecast results is 0.0078.

After many tests, LSTM uses five nodes in the input layer, 15 nodes in the hidden layer and one node in the output layer. The residuals from ARIMA model fitting the training set are used as the training data of LSTM network and the residuals from ARIMA model forecasting the test set are taken as the test data of LSTM model. After 1500 rounds of training, the mean square error of forecast results of the test data tends to be stable, the mean square error is 0.0083.

Finally, the forecast results of LSTM network and the forecast results of ARIMA model are added one by one to get the final forecast results of the last 129 days. The forecast results of the model are shown in Figure 6. As we can see from the figure, the forecast accuracy is high enough.

Du Yi, Ma Rongyong[16] also forecast hydrological time series by combined models such as EMD-ARIMA model, WA-ARIMA model and BP-ANN-ARIMA model. The forecast results show that BP-ANN-ARIMA model has the best forecast effect among them. In order to verify the effect of the model proposed in this paper, the hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM is compared with LSTM network, ARIMA model and BP-ANN-ARIMA model. The forecast effect of each model on test set is shown in Figure 7.

It can be clearly seen from the figure that the forecast results of LSTM network and the hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM are very close to the true values, while, there is a certain deviation between the forecast values of ARIMA model and the real values. The mean square error and the mean absolute percentage error of each model's forecast results are calculated,

and the results are shown in Table 1. The table shows that MSE and MAPE of the hydrological time series forecast model are smaller than those of LSTM network, ARIMA model and BP-ANN-ARIMA model.

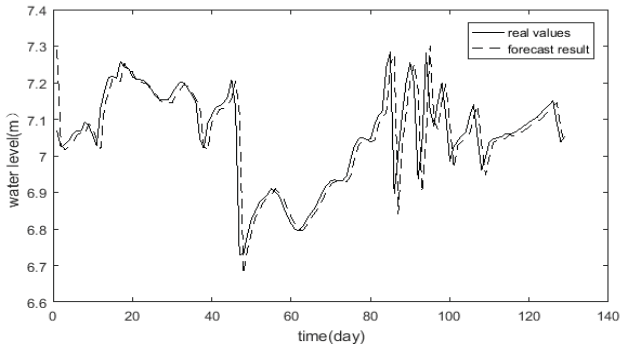

Fig. 6. Forecast result

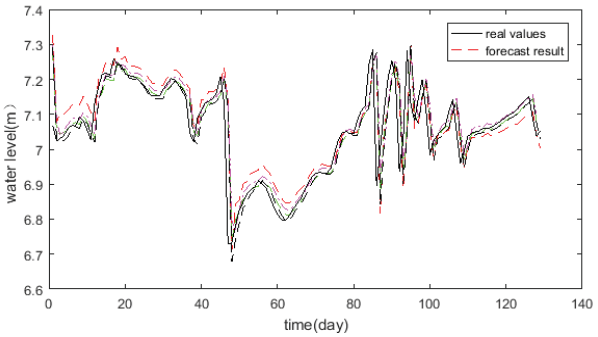

Fig. 7. Comparision diagram

TABLE I. COMPARISION

Model	*MSE*	*MAPE*
Arima	0.0178	0.72%
LSTM	0.0049	0.56%
BP-ANN-ARIMA	0.0055	0.59%
de-noising-ARIMA-LSTM	0.0044	0.51%

IV. CONCLUSION

In this paper, a hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM is proposed. The daily average water level of a hydrological station in Chuhe River basin is used as experimental data. The model is compared with LSTM network, ARIMA model and BP-ANN-ARIMA model. Experiments show that the model has a good forecast effect, and the forecast accuracy is higher than that of other models. The experiment of the model is successful.

The hydrological time series forecast model proposed in this paper can effectively remove the interference factors in the water level time series. Through the complementary advantages of ARIMA model and LSTM network, the complex law of water level time series is well studied and the forecast accuracy is improved.

REFERENCES

[1] Katimon A, Shahid S, Mohsenipour M. Modeling water quality and hydrological variables using ARIMA: a case study of Johor River, Malaysia[J]. Sustainable Water Resources Management, 2017(3):1-8.

[2] Dastorani M, Mirzavand M, Dastorani M T. Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition[J]. Natural Hazards, 2016, 81(3):1811-1827.

[3] Khashei M, Bijari M, Ardali G A R. Improvement of Auto-Regressive Integrated Moving Average models using Fuzzy logic and Artificial Neural Networks (ANNs)[J]. Neurocomputing, 2009, 72(4):956-967.

[4] Gers F A, Eck D, Schmidhuber J. Applying LSTM to Time Series Predictable through Time-Window Approaches[C]. International Conference on Artificial Neural Networks. Springer-Verlag, 2001:669-676.

[5] J. L. Yang, M. Xu, F. L Wang. Research Based on BP Neural Network Time Series Forecasting[J]. Mathematics in Practices and Theory, 2013, 43(4):158-164.

[6] L. Liu, W. Ye. Precipitation prediction of timeseries model based on BP artificial neural network [J]. Journal of Water Resources & Water Engineering, 2010, 21(5):156-159.

[7] Valunjkar S S, Patil S. Prediction of Daily Runoff Using Time Series Forecasting and ANN Models,[C] Proceeding of International Conference on Science and Technology 2k. 2014.

[8] X. Wang, J. Wu, C. Liu. Exploring LSTM based recurrent neural network for failure time series prediction [J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(4):772-784.

[9] Y. Y. Yue, Q. Fu, D. S. Wang. A Prediction Model for Time Series Based on Deep Recurrent Neural Network [J]. Computer Technology And Development, 2017, 27(3):35-38.

[10] GRAVES A. Long short-term memory[M]. Berlin:Springer, 2012:1735-1780

[11] L. Luo, L. H. Li,C. Y. Wang, P. D. Lu, P. S. Yang. Insulator status data mining method based on ARIMA-LSTM[J]. Journal of Electric Power Science & Technology, 2017.

[12] Donoho D. De-noising by soft-thresholding[M]. IEEE Press, 1995.

[13] Donoho D, Johnstone I M. Ideal spatial adaptation via wavelet shrinkage[J] Biometrika. 1994.

[14] Donoho D, Johnstone I. Adapting to Unknown Smoothness via Wavelet Shrinkage[J]. Publications of the American Statistical Association, 1995, 90(432):1200-1224.

[15] M. W. Liu, D. Y. Wang, S. L. Zhou. Forecasting Methods for Port Throughout Capacity [J].Port & Waterway Engineering, 2005(3):53-56.