Research on Water Demand Prediction in the Yangtze Valley Based on the STL-LSTM Model

Sheng-Lin Ke [4,5,6,a,*], Jun Zhang [4,5,6,b], Dang-Gen Guan [4,5,6,c], De-Jing Zeng [4,5,6,d], Hong-Ya Qiu [1,e], Zhi-Zhou Feng [1,2,f], Shu-Fei Li [3,g], Yu Hui [3,h]

[1]China Three Gorges Corporation, Wuhan, China
[2]China Yangtze Power Co., Ltd., Beijing, China
[3]Changjiang Survey, Planning, Design and Research Co., Ltd., Wuhan, China
[4]Network and Information Center, Changjiang Water Resources Commission, Wuhan, China
[5]Smart Yangtze River Innovation Team, Changjiang Water Resources Commission, Wuhan, China
[6]Technology Innovation Center of Digital Enablement for River Basin Management, Changjiang Water Resources Commission, Wuhan, China

[a]310109040@qq.com, [b]94029897@qq.com, [c]114622232@qq.com, [d]381651638@qq.com, [e]343387289@qq.com, [f]feng_zhizhou@ctg.com.cn, [g]lishufei@cjwsjy.com.cn, [h]455373313@qq.com

Abstract. Hydraulic engineering involves the utilization of water for various purposes, such as water supply, irrigation and power generation. The efficient utilization and scientific management of water resources constitute an essential research subject. Within the realm of water resources management, accurate water demand prediction is of paramount significance for guaranteeing the sustainable utilization of water resources and fulfilling the water requirements of human society. In order to accurately predict the short-term water demand of the Yangtze valley and support for water resource management and allocation, this paper combines the Long Short Term Memory (LSTM) model and Time Series Decomposition (STL) to construct the STL-LSTM water demand prediction model. The study takes the large-scale water users in the main stream of the Yangtze River as an example to predict water demand. LSTM and ARIMA are used as comparative models, and Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the accuracy of the model prediction. The results show that compared with LSTM and ARIMA models, STL-LSTM model can better predict water demand trends, and the MAPE of prediction is reduced by 0.94% and 0.85% respectively, with higher prediction accuracy. STL-LSTM model can be applied to water demand prediction in the Yangtze valley.

Keywords: water demand prediction; LSTM; STL time series decomposition; Yangtze valley

Y. Zhang et al. (eds.), *Proceedings of the 2024 10th International Conference on Architectural, Civil and Hydraulic Engineering (ICACHE 2024)*, Advances in Engineering Research 259,
https://doi.org/10.2991/978-94-6463-658-1_56

1 Introduction

The Yangtze River, recognized as one of the most significant rivers in China, plays a vital role in supplying essential water resources for agricultural production, industrial activities, and residential life within its basin. However, these water resources are finite; with ongoing economic development and demographic changes, the demand for water is continuously evolving. Accurately predicting water demand can provide critical data support for the effective management and allocation of water resources within the basin.

In recent years, a multitude of researchers have conducted extensive studies on water demand forecasting. Some have utilized quota analysis methods that intuitively account for various key factors influencing water demand [1][2]; while this approach is straightforward and accessible, it is inherently subjective, which may result in considerable deviations in the predictive outcomes. Additionally, regression analysis techniques based on the characteristics of water use structures have also been developed[3], which are easy to operate but require accurate identification of factors related to water demand. Furthermore, system dynamics[4], NSGAII-FORAGM[5], and composite approaches[6] are employed to predict the future annual water demand of various basins or regions. The results can serve as the basis for water allocation within the year but cannot be regarded as a reference for short-term water resource allocation. Regarding short-term water demand forecasting, Long Short-Term Memory[7], machine learning[8], linear regression model combined with evolutionary strategies[9] are beneficial for predicting the water demand of urban areas and water utilities. However, these methods do not directly take into account the seasonal characteristics of water supply in the basin and need to be further enhanced to predict the short-term mixed water demand of the basin.

Currently, numerous research outcomes have emerged regarding the forecasting of water demand in various industries and regions. However, research on the short-term prediction of the overall water supply in the Yangtze River Basin remains insufficient. As key hydroelectric projects in the Yangtze River Basin continue to be developed and put into operation, transitioning from single-reservoir scheduling to cascade scheduling, accurate short-term water demand predictions become critically important for effective water resource allocation, especially during drought periods or low-water seasons.

This study concentrates on the pertinent water abstraction units located in the river section downstream of the Xiangjiaba Dam within the mainstream of the Yangtze River Basin, employing long short-term memory (LSTM) network algorithms to investigate daily water demand forecasting.

2 Overview of the Study Area and Data Sources

The Yangtze River Basin encompasses an area of approximately 1.8 million km^2 and spans 19 provinces (or autonomous regions or municipalities). It is subdivided into 12 secondary water resource zones: above Shigu in the Jinsha River, below Shigu in the

Jinsha River, Min-Tuo River, Jialing River, Wu River, Yibin to Yichang, Dongting Lake water system, Han River, Poyang Lake water system, Yichang to Hukou, mainstream below Hukou, and Taihu Lake water system.

According to the Yangtze River Basin's water resources bulletins from 2008 to 2022, the average annual water supply amounts to 203.048 billion cubic meters. Of this total volume: agricultural usage constitutes approximately 49.27%, industrial use accounts for about 34.92%, while domestic consumption represents around 14.33%. The trends in water utilization are depicted in Figure 1.

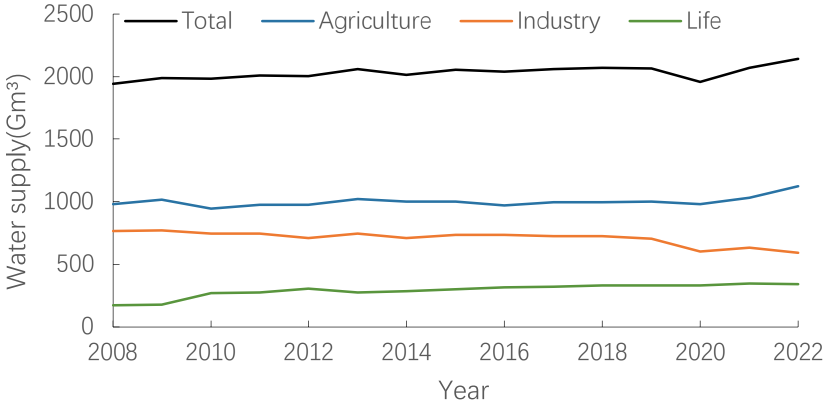

Fig. 1. Trend chart of water supply changes in the Yangtze valley from 2008 to 2022

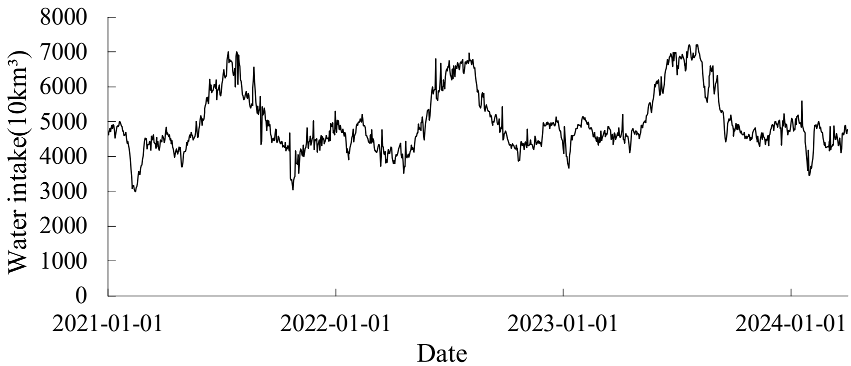

Fig. 2. Water intake from the main steam of the Yangtze River below Gezhouba

This study identifies a total of 789 monitoring points for water resource measurement along the mainstream of the Yangtze River downstream from Xiangjiaba Dam. The specific distribution across various zoning sections is as follows: 52 points between Xiangjiaba and Zhutuo; 46 points from Zhutuo to Cuntan; 42 points spanning Cuntan to Yichang; 75 points between Yichang and Chenglingji; 50 points from Chenglingji to

Hankou; 126 points extending from Hankou to Hukou; 47 points between Hukou and Datong; and finally, 351 points located below Datong.

In terms of industry distribution, there are 112 points for agriculture, 307 points for domestic use, and 370 points for industry. Daily water abstraction monitoring data for each monitoring point from January 1, 2021, to March 31, 2024, were obtained from the water resource monitoring management platform of the Yangtze River Basin, as shown in Figure 2.

3 Research Methodology

Water demand is influenced by numerous factors including temperature, precipitation, and economy. Moreover, Long Short-Term Memory possesses the capability of remembering and forgetting, enabling it to handle and predict such complex time series. Based on the monitoring of water consumption in the Yangtze River Basin, agricultural water utilization constitutes approximately half, and agricultural water use exhibits certain seasonal characteristics. Hence, the combination of LSTM and STL Time Series Decomposition can better predict water demand.

3.1 Long Short-Term Memory Model

Long Short-Term Memory (LSTM) is a specific architecture of Recurrent Neural Network (RNN), designed to address the challenges of vanishing and exploding gradients commonly encountered in traditional RNNs. Unlike feedforward neural networks, RNNs have recurrent connections, which make them very powerful in sequence modeling. The structure of an individual LSTM neuron[10] is shown in Figure 3.

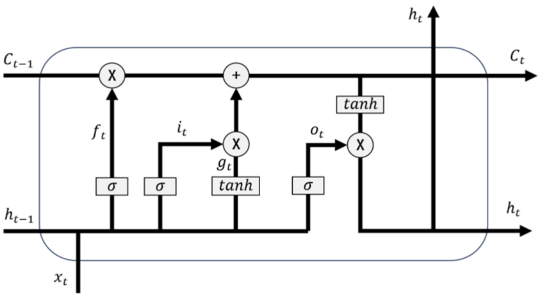

Fig. 3. LSTM nerual unit structure diagram

In the LSTM model structure, the key is the cell state, which runs horizontally across the top of the diagram and is regulated by gate structures. A gate is a way to selectively allow information to pass through, consisting of a sigmoid neural network layer and a point-wise multiplication operation. Among them, the sigmoid layer output ranges from 0 to 1, describing how much of the component should be allowed to pass. LSTM has three such gates to protect and control the unit state, namely, the forget gate, the input gate, and the output gate. In each LSTM layer, information first passes through the forget gate layer, which determines what information to discard from the state. The formula is as follows:

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{1}$$

In the formula, σ is the sigmoid function, x_t is the input at time t, and h_{t-1} is the state value of the hidden layer at time t-1 or the initial state at time 0. Next is the input gate layer, which determines what new information to store in the state. The formula is as follows:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{2}$$

$$g_t = tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \tag{3}$$

In the formula, the sigmoid function determines which values to update, while the tanh function (hyperbolic tangent activation function) creates a vector of new candidate values. These are combined with the input gate layer and the forget gate layer to update the cell state. The formula for updating the cell state is as follows:

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{4}$$

In the formula, $\odot$ denotes element-wise multiplication. c_t is the cell state at time t. This equation represents the combination of the information retained from the previous cell state controlled by the forget gate and the candidate cell state controlled by the input gate, resulting in the current cell state. Finally, the output gate operates a sigmoid layer to determine which parts of the cell state to output, and multiplies this by the tanh-transformed cell state to obtain the hidden state output. The formula is as follows:

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

In the formulas, $W_{if}, W_{hf}, W_{ii}, W_{hi}, W_{ig}, W_{hg}, W_{io}, W_{ho}, b_{if}, b_{hf}, b_{ii}, b_{hi}, b_{ig}, b_{hg}, b_{io}$ and b_{ho} are all weight and bias parameters that need to be learned.

3.2 STL Time Series Decomposition

STL is a classical time series decomposition method that is suitable for decomposing a time series into three components: trend, seasonality, and residuals, featuring excellent smoothness and flexibility. The additive model expression of STL time series is as follows:

$$Y_t = T_t + S_t + R_t \tag{7}$$

In this equation, Y_t is the sequence value at time t, T_t denotes the trend component at time t, S_t indicates the seasonal component at time t, and R_t stands for the residual value at time t.

The core idea of STL time series decomposition is to smooth the sequence via locally weighted regression (Loess), and then separately estimate the trend, seasonality, and residuals at each time point. It mainly encompasses the following three fundamental steps:

(1) Locally weighted regression (Loess) smoothing: Apply Loess smoothing to the time series to estimate the overall trend and seasonal components.

(2) Extraction of seasonal components: Subtract the trend component from the sequence after Loess smoothing to obtain the seasonal component.

(3) Calculation of residuals: Subtract the trend obtained from Loess smoothing from the original time series to obtain the residual sequence, which is the portion of the original sequence excluding the trend and seasonality.

3.3 STL-LSTM Water Demand Forecasting Model

Water demand data is influenced by factors such as seasonality, trend, and randomness, which have a significant impact on the predictive model. To enhance the prediction accuracy, this paper employs the STL-LSTM model for water demand forecasting. The prediction process is depicted in Figure 4.

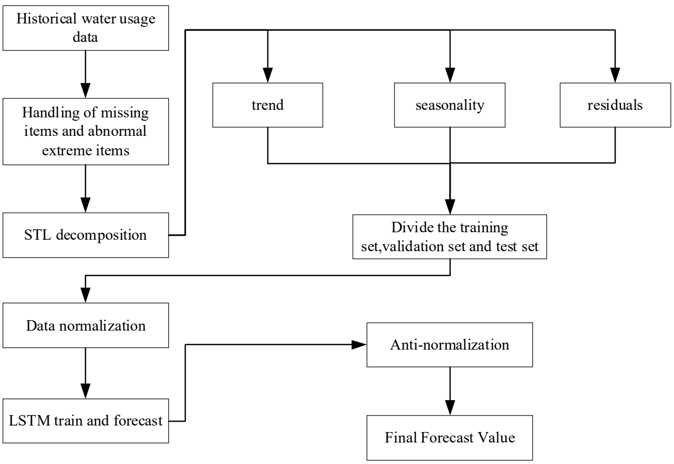

Fig. 4. Flow chart of water demand prediction based on STL-LSTM model

Firstly, the missing and abnormal extreme values in the historical water consumption data are corrected using the average values before and after the sequence time. Then, the processed water consumption time series data is decomposed using STL to obtain the trend component, seasonal component, and residual component. Next, the decomposed trend and seasonal components are separately divided into training sets, validation sets, and test sets, and normalized before LSTM model training. After model training, predictions are made on the test set, and finally, the predicted water consumption is denormalized to obtain the final predicted water demand.

4 Results and Analysis

To verify the feasibility of the STL-LSTM model for water demand forecasting in the Yangtze River Basin, this paper conducts a case study on water demand forecasting for water abstraction entities above a certain scale along the mainstream of the Yangtze River, and compares the results with those from an LSTM model without time series decomposition and a conventional ARIMA model. The LSTM model is constructed using the PyTorch open-source library, with mean squared error set as the loss function, Adam as the optimizer with a learning rate of 0.01, and the model is trained by adjusting three key parameters: the number of hidden layer states, the number of recurrent layers, and the length of input historical data. Before model training, the daily water consumption data is normalized, and over 570,000 daily water consumption data points from 2021 and 2022 are divided into a training dataset, while data from 2023 is used as a validation dataset. After model training, a predictive analysis is performed on the daily water consumption data for January to March 2024.

By adjusting the key parameters to train the model, the optimal prediction results are obtained when the LSTM model is set with 32 hidden layer states, 5 recurrent layers, and an input data length of 7. The predicted water demand results are presented in Figure 5, the model evaluation metrics are shown in Table 1, and the relative absolute prediction errors at various time points are illustrated in Figure 6.

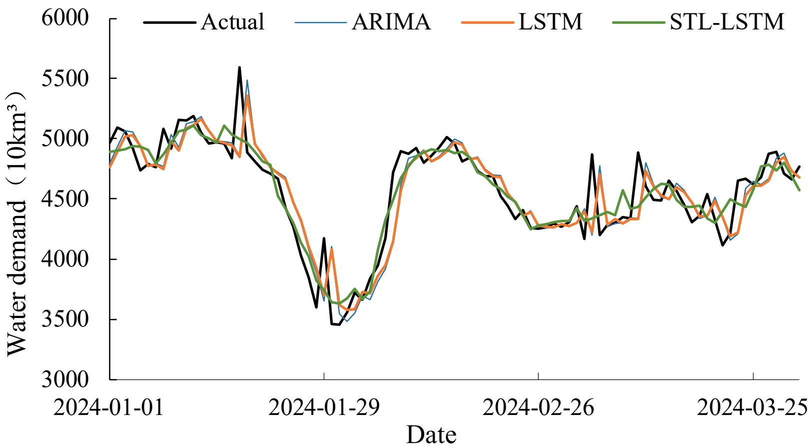

Fig. 5. Actual and predicted water demand

As demonstrated in Figure 5, the STL-LSTM, LSTM, and ARIMA models can all effectively predict the water demand along the mainstream of the Yangtze River. Among them, the prediction results of the LSTM model and the ARIMA model are relatively close, both of which can well predict the continuous fluctuating data trends, but there is a certain lag in the prediction results, indicating that the models lack sufficient learning of irregular fluctuations with large water withdrawals and tend to use the previous water withdrawal amount as the predicted value. In contrast, the STL-LSTM model offers smoother predictions at water volume mutation points with no significant lag. This is because after time series decomposition, the model can accurately predict the smooth trend and seasonal components, but there is a certain bottleneck in learning the larger and irregular fluctuations of the residual component, resulting in smoother predictions. When there are mutations, the predicted value changes relatively little. Additionally, based on practical experience, the actual fluctuations in water withdrawals tend to decrease after the year-end water withdrawal summary and verification. Therefore, the prediction results of the STL-LSTM model are more in line with actual conditions.

Table 1. Comparison of evaluation indicators

model	RMSE/(10km^3)	MAPE/%
ARIMA	156.85	3.49
LSTM	153.02	3.39
STL-LSTM	115.78	2.55

As shown in Table 1, the STL-LSTM model has the lowest RMSE and MAPE values compared to the LSTM and ARIMA models, indicating that the STL-LSTM model has higher prediction accuracy and better prediction performance.

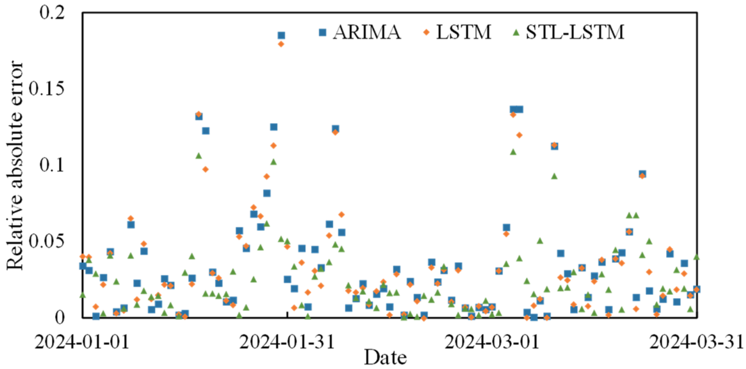

Fig. 6. Relative absolute error of prediction results

As shown in Figure 6, the maximum relative absolute error of the STL-LSTM model's prediction results is around 10%, which is less than the 18% of the LSTM and ARIMA models. Moreover, over 90% of the model's prediction results have a relative absolute error within 5%, indicating that the prediction results are more stable.

5 Conclusion and Outlook

In this paper, the STL-LSTM model is employed to predict the water demand of water abstraction entities above a certain scale along the mainstream of the Yangtze River. In contrast to the LSTM and ARIMA models, the STL-LSTM model attains the lowest

RMSE and MAPE values, signifying higher prediction accuracy and more stable prediction outcomes. Hence, the STL-LSTM model can be utilized for water demand forecasting in the Yangtze River Basin, offering valuable reference for water resources management and water scheduling in the region.

The case study in this paper only utilizes historical water withdrawal time series data. For future research, additional relevant factors that have an impact on water demand, such as temperature and rainfall, can be collected to further enhance the water demand forecasting in the Yangtze River Basin.

Acknowledgment

This work was supported by the Scientific Research Project Fund of China Three Gorges Corporation (Grant No. 0704231).

References

1. LU D, HU C H, WU Z Y, Predicting Method for Demand of Agriculture Water Based on Quantitative Analysis[J]. Journal of Irrigation and Drainage, 2008, 27(06): 88-91.
2. HE G, WU Z L, HU C H. Forecast model of industrial water demand based on quota and quantitative analysis[J]. Journal of Water Resources and Water Engineering, 2008, (02): 60-63+67.
3. JING D M, RONG N. Study on the water demand prediction of Changchun City based on regression analysis[J]. Water Resources & Hydropower of Northeast China, 2018, 36(10): 19-21+29+71.
4. HUANG M T, LI X J, WANG J M. Research on regional water demand prediction of the upper and middle reaches of The Pearl River Basin based on system dynamics[J]. IOP Conference Series: Earth and Environmental Science, 2021, 784(1).
5. LI J, LIU C Y, TANG L. Forecast of regional water demand based on NSGAII-FORAGM[J]. Water Supply, 2022, 22(2): 1889-1904.
6. F Y ,N H ,Z Z N , et al. Analyzation of water demand projection for Johor River Basin[J]. IOP Conference Series: Earth and Environmental Science, 2023, 1143(1).
7. Mu L, Zheng F, Tao R, et al. Hourly and Daily Urban Water Demand Predictions Using a Long Short-Term Memory Based Model[J]. Journal of Water Resources Planning and Management, 2020, 146(9).
8. Antunes A, Andrade-Campos A, Sardinha-Lourenço A, et al. Short-term water demand forecasting using machine learning techniques[J]. Journal of Hydroinformatics, 2018, 20(6): 1343-1366.
9. Justyna S, Joanna S K ,Piotr L, et al. Improving short-term water demand forecasting using evolutionary algorithms[J]. Scientific Reports, 2022, 12(1): 13522-13522.
10. Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.