# IEICE TRANSACTIONS

# on Information and Systems

DOI:10.1587/transinf.2024EDL8047

Publicized:2024/08/14

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

# A clustering-based deep learning method for water level prediction

**Chih-Ping Wang**[†], ***student member and*** **Duen-Ren Liu**[††]

**SUMMARY** Accurate water level prediction systems improve safety and quality of life. This study introduces a method that uses clustering and deep learning of multisite data to enhance the water level prediction of the Three Gorges Dam. The results show that Cluster-GRU-based can provide accurate forecasts for up to seven days.

***key words:*** *water level prediction, clustering, Deep learning, Cluster-GRU-based, Three Gorges Dam.*

## 1. Introduction

The Yangtze River Basin (YZRB) is 6.387 kilometers long and flows through 11 provinces in mainland China. It has a population of 440 million along the coast and more than 200,000 square kilometers of arable land. Dams have water storage, power generation, irrigation, flood control, and navigation functions. Rainfall in this basin has significant seasonal changes, with snowmelt and early spring rainfall causing the water level to rise, summer rainfall with frequent heavy rains, and the water level reaching its highest point, autumn rainfall decreasing, and the water level gradually falling until winter, when rainfall is minimal. The water level reaches its lowest point. The upstream river area is located on a plateau, and its rainfall and snow melt speed directly affects the water level regulation of the Three Gorges Dam (TGD). In recent years, due to climate change, heavy rains and snowfall have occurred in the Yangtze River Basin, and the uncertainty in the rise and fall of water levels has increased [1][2]. Upgrading accurate modeling and water level predictions is critical to disaster prevention and control. This research applies deep learning and various time series forecasting methods to reduce losses and improve response systems [3].

Research has highlighted the effects of different similarity measures on amplitude, temporal, and spatial differences [4]. Using stepwise clustering analysis, a statistical hydrological model for the Yangtze River basin addresses basin complexities [5]. Rapid industrialization requires a balance of economic growth, energy sustainability, and environmental conservation. Meticulous planning, particularly for dam water level forecasts, is crucial for agriculture and aquatic ecosystems [6].

The study explores time series classification (TSC) and prediction techniques in hydrology, using modern deep learning algorithms for comparative analysis, including experimental group Cluster-GRU-based, control group LSTM-based, GRU-based, and Cluster-LSTM-based networks.

Studying how to observe water levels accurately raises two issues that must be discussed and resolved.

A. Water spatial feature extraction problem.

B. Water level time feature extraction and prediction problem.

In the study, three water level observation stations were taken as examples. Fuling, Wanzhou, and Zigui upstream of the TGD; and Yichang, Zhijiang, and Shashi downstream.

A. The clustering method for the TGD is upstream and downstream observation stations.

B. Compute water level forecasts using a clustering method based on temporal similarity of water level time series.

Figure 1 shows a clustering-based deep learning water level prediction flow chart method.

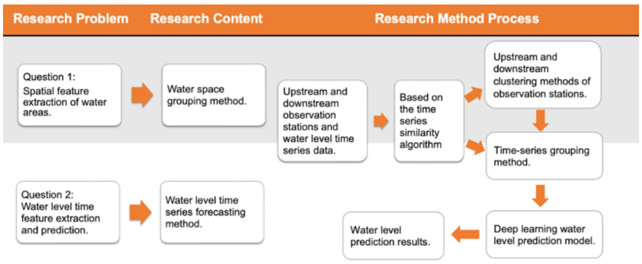
Research Problem
Research Content
Research Method Process
Question 1: Spatial feature extraction of water areas.
Water space grouping method.
Upstream and downstream observation stations and water level time series data.
Based on the time series similarity algorithm
Upstream and downstream clustering methods of observation stations.
Time-series grouping method.
Question 2: Water level time feature extraction and prediction.
Water level time series forecasting method.
Water level prediction results.
Deep learning water level prediction model.

**Fig. 1** A clustering-based deep learning water level prediction flow chart method.

## 2. Methodology

### 2.1 Study Area

There are 18 water level monitoring stations in the YZRB. Figure 2 is divided into the upper reaches, the middle reaches, the lower reaches, and the estuary, each with critical hydrological stations.

Six daily water level stations at the TGD were analyzed to provide detailed predictions and analysis of these changes. Figure 3 Diagram highlighting the TGD region and its six

[†] Chih-Ping Wang is with the Institute of Information Management, National Yang-Ming Chiao Tung University, No. 1001, Daxue Rd. East Dist., Hsinchu City 300093, Taiwan.

[††] Duen-Ren Liu is with the Institute of Information Management, National Yang-Ming Chiao Tung University, No. 1001, Daxue Rd. East Dist., Hsinchu City 300093, Taiwan.

observation hydrological stations.

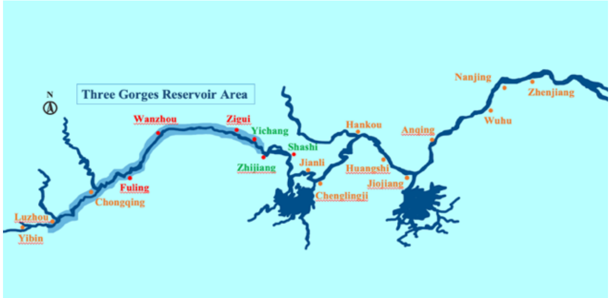

**Fig. 2** Distribution map of eighteen water level observation stations in the YZRB.

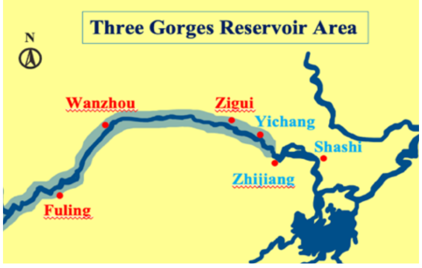

**Fig. 3** Diagram highlighting the TGD region and its six observation hydrological stations.

## 2.2 Modeling Strategy

### 2.2.1 Clustering Method based on water area data

The model is constructed using TGD daily water level station data, and the k-means clustering method is used for clustering based on water area data. The proper noun “K-Means” refers to the L1 norm in this study.

The study calculates the distance between each observation and the previous centroid point. The centroid point in each cluster is then updated after updating the observations belonging to that cluster. Iterate these steps until the center of mass point changes. The steps of k-means are as follows:

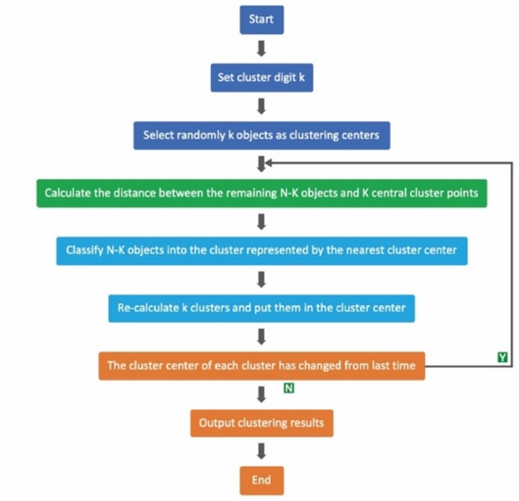
Start
Set cluster digit k
Select randomly k objects as clustering centers
Calculate the distance between the remaining N-K objects and K central cluster points
Classify N-K objects into the cluster represented by the nearest cluster center
Re-calculate k clusters and put them in the cluster center
The cluster center of each cluster has changed from last time
Y
N
Output clustering results
End

**Fig. 4** Flowchart of steps using k-means at six water level observation stations of the TGD.

This study applies two clusters and Euclidean distance to a k-means model, clustering of time series data using the R suite.

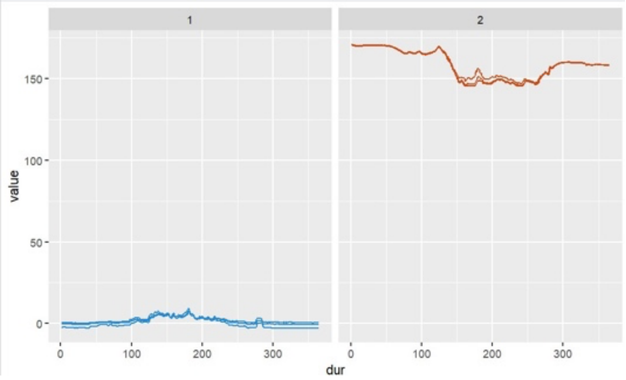

**Fig. 5** The TGD water level observation stations are clustered into two groups.

There are six observation stations upstream and downstream of the TGD, each with unique topography and flow characteristics. The upper reaches are hilly, and the water level is high; the downstream terrain is flat, and the water level is low. These characteristics significantly influence changes in water levels.

The results show that the k-means clustering method was used in the Three Gorges Dam. Fuling, Wanzhou, and Zigui in the upper reaches are clustered into one group and Yichang, Zhijiang, and Shashi in the lower reaches are clustered into one group. As shown in Figure 5.

Cluster analysis was carried out upstream and downstream of the TGD and the following phenomenon was obtained: The data trends at the three observation points upstream of Fuling, Wanzhou, and Zigui are similar, showing a concave shape. In contrast, the data trends in Yichang, Zhijiang, Shashi, and other places downstream of the TGD are precisely the opposite, showing a convex shape. This is an exciting phenomenon in the forecasting of water levels.

### 2.2.2 Forecasting method based on water level time series

In the GRU-plus-based method, time series-based similarity is used to calculate clustered water levels upstream and downstream of the TGD. Fuling, Wanzhou, and Zigui are located upstream of the TGD, and the water level changes at each observation station are similar. Fuling is situated in the central area of Wanzhou, and Zigui is the first county in the TGD Reservoir Area. According to observations, when the water level in Fuling reaches its peak, the water level in Wanzhou, Zigui, and other places will reach its peak later or the next day. For example, the modeling uses water level data in Fuling, Wanzhou, and Zigui as input, and the Zigui water level observation station is used to establish a prediction model.

The downstream location of the Three Gorges Dam is lower than the upstream location. Observation of the water levels in Yichang, Zhijiang, and Shashi downstream found that after the water level changed, the water level trends in Yichang, Zhijiang, and Shashi differed from those upstream. For example, in the modeling, the water level data of Yichang, Zhijiang, and Shashi were used as input, and the Shashi water level observation station was used to establish a prediction model, which achieved good results.

Research shows that using the k-means clustering method for time series prediction on GRU-PLUS-based time series data from upstream and downstream water level observation stations of the TGD, we can expect differences in water level observations to understand the accuracy of predictions.

## 3. Experimental Results and Discussions

The experiment aims to predict water levels upstream and downstream of the TGD and provide accurate predictions for up to seven days. The experiment was conducted using predictions from six water level observation stations of the TGD. To evaluate the water level prediction, this study collected water level information from January 1 to December 31, 2022, and analyzed and assessed the water level prediction.

This research method utilizes the gated recurrent unit (GRU), a recurrent neural network (RNN) architecture proposed by Cho et al. in 2014 [7]. GRU combines input and forgets gates into a single update gate, simplifying the model structure and improving computational efficiency.

In the GRU model, this study uses two types of gates: update gate ($z$) and reset gate ($r$). Update gates control the flow of messages to hidden states, determining how much past messages should be passed into the future. The reset gate determines how much past information should be forgotten. By using these gates, GRUs can capture long-term dependencies in sequential data using fewer parameters and faster training times than long-short-term memory (LSTM) networks while still maintaining similar performance. In the initial GRU model of this study, we used the applied row windowing method to obtain prior data for assumed window periods (7 days, 14 days, 28 days) at the target site as covariates in the GRU model.

However, in this research method the Cluster-GRU-based model, the research method first uses Pearson's correlation to cluster those water level stations whose data have the same trend. Then, the study applied the row-window method to obtain previous data for assumed window periods (7 days, 14 days, 28 days) in the target site and cluster sites as covariates in the GRU model.

The data come from January to September 2022 as the training data set. The data from October to December 2022 is used as the test data set.

To evaluate the accuracy and reliability of the proposed Cluster-GRU-based model, multiple performance metrics are used to analyze the prediction results, including MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), MSE (Mean-Square Error), and RMSE (Root Mean Square Error), as shown in Equations (1), (2), (3), (4).

MAE is the mean absolute difference between the predicted and actual values. The formula for MAE is as Eq. (1).

$$\text{MAE} = \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{n} \tag{1}$$

MAPE stands for mean absolute percentage error, which calculates the absolute percentage between the predicted and actual values. The formula for the MAPE is as Eq. (2).

$$\text{MAPE} = \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|/y_i}{n} \tag{2}$$

MSE stands for mean square error, which is close to MAE but uses the squared difference instead of the absolute value. The formula for MSE is as Eq. (3).

$$\text{MSE} = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n} \tag{3}$$

RMSE represents the root mean square error, which is the root of MSE. The formula for RMSE is as Eq. (4).

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{4}$$

In this study, the data from January to September are considered the training data set. Data from October to December were used as a test data set. All analyses were performed under R4.4.1 with the Karas3 package.

The deep learning method based on Cluster-GRU can predict the water level time series for seven days per week. First, clustering was performed to separate two groups, namely the upstream and downstream of the TGD. The historical data of the three water level observation stations in Fuling, Wanzhou, and Zigui upstream of the TGD were normalized to predict the water level heights in Fuling, Wanzhou, and Zigui, respectively. The historical data of the three water level observation stations in Yichang, Zhijiang, and Shashi downstream of the TGD were regularized to predict the water level heights in Yichang, Zhijiang, and Shashi respectively. Table 1 shows the water level height forecast for the six water level observation stations of the TGD from the 1st to the 7th day (October 1, 2022, to October 7, 2022), such as the predicted water level days P1, P2, P3, P4, P5, P6, and P7.

To compare the accuracy of the water level prediction, this study analyzed the water level information of the six water level stations of the TGD. The Cluster-GRU-based method of the experimental group was compared with the GRU-based, LSTM-based, and Cluster-LSTM-based methods of the control group. The experiment analyzed and compared different methods' water level prediction evaluation indicators. Data from January to September 2022 is considered the training data set. Data from October to December are used as the test data set. GRU-based, LSTM-based, Cluster-LSTM-based, and Cluster-GRU-based methods predict water levels. The results of the evaluation indicators using MAE, MAPE, MSE, and RMSE are shown in Table 2.

Although GRU-based, LSTM-based, and Cluster-LSTM-based methods can predict good water level heights in a shorter period, the water level prediction errors become more significant as time increases. Therefore, the expected water level height of the Cluster-GRU-based method is more optimized. Therefore, the Cluster-GRU-based process is more suitable for predicting the TGD water level.

**Table 1** Six water level observation stations at the TGD predict seven-day water level heights (m).

| Prediction Station | P1 | P2 | P3 | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| Fuling | 154.1 | 154.5 | 154.5 | 153.3 | 153.4 | 153.8 | 155.0 |
| Wanzhou | 154.0 | 154.3 | 154.3 | 153.6 | 153.1 | 153.1 | 154.0 |
| Zigui | 153.6 | 153.8 | 153.6 | 153.2 | 152.5 | 152.4 | 154.2 |
| Yichang | 0.5 | 0.6 | 1.4 | 2.6 | 3.0 | 3.0 | 3.1 |
| Zhijiang | -0.5 | -0.3 | -0.5 | 0.9 | 1.8 | 1.7 | 1.8 |
| Shashi | -2.6 | -2.5 | -2.3 | -1.4 | 0.3 | 0.3 | 0.3 |

**Table 2** Comparison of four deep learning models' MAE, MAPE, MSE, and RMSE evaluation indicators from the six water level observation stations.

| Prediction Station | Deep Learning | MAE | MAPE | MSE | RMSE |
|---|---|---|---|---|---|
| Fuling | Cluster-LSTM-based | 2.07 | 1.30 | 3.08 | 7.20 |
| | **Cluster-GRU-based** | **1.13** | **0.71** | **3.21** | **1.58** |
| | LSTM-based | 2.83 | 1.78 | 11.23 | 12.03 |
| | GRU-based | 2.89 | 1.82 | 12.75 | 13.03 |
| Wanzhou | Cluster-LSTM-based | 1.56 | 0.98 | 2.14 | 4.58 |
| | **Cluster-GRU-based** | **0.78** | **0.49** | **1.45** | **1.29** |
| | LSTM-based | 2.51 | 0.84 | 9.34 | 7.30 |
| | GRU-based | 2.92 | 0.89 | 14.20 | 14.26 |
| Zigui | Cluster-LSTM-based | 1.03 | 0.65 | 1.91 | 2.17 |
| | **Cluster-GRU-based** | **0.63** | **0.40** | **1.14** | **1.01** |
| | LSTM-based | 1.95 | 0.77 | 7.48 | 7.48 |
| | GRU-based | 1.91 | 0.83 | 6.50 | 6.48 |
| Yichang | Cluster-LSTM-based | 0.33 | 38.39 | 0.75 | 0.66 |
| | **Cluster-GRU-based** | **0.27** | **28.49** | **0.42** | **0.61** |
| | LSTM-based | 0.58 | 65.32 | 1.03 | 1.11 |
| | GRU-based | 0.65 | 62.36 | 1.21 | 1.21 |
| Zhijiang | Cluster-LSTM-based | 0.31 | 61.13 | 0.61 | 0.61 |
| | **Cluster-GRU-based** | **0.29** | **42.80** | **0.34** | **0.57** |
| | LSTM-based | 0.53 | 82.80 | 0.76 | 0.79 |
| | GRU-based | 0.50 | 88.10 | 0.71 | 0.70 |
| Shashi | Cluster-LSTM-based | 0.61 | 58.42 | 0.7 | 1.13 |
| | **Cluster-GRU-based** | **0.36** | **58.78** | **0.57** | **0.70** |
| | LSTM-based | 0.83 | 101.18 | 1.92 | 1.84 |
| | GRU-based | 0.78 | 95.16 | 1.50 | 1.73 |

## 4. Conclusions

This study proposes a method Cluster-GRU-based to predict the water levels upstream and downstream of the TGD. In experiments, the study compared data from the TGD water level observation station with water levels estimated using this method. Although the GRU-based and LSTM-based methods can provide good water level predictions, their errors increase over time. However, the Cluster-LSTM-based and Cluster-GRU-based methods were compared through clustering, and the Cluster-LSTM-based method did not achieve the optimal results. Experimental results show that the experimental group's method Cluster-GRU-based is better than other methods in water level prediction. Therefore, the deep learning method Cluster-GRU-based can predict the water levels upstream and downstream of the TGD and provide accurate forecasts for up to 7 days.

The scalability of this method is limited, but long-term historical data can be applied to rivers around the world, such as the Nile, Amazon, Mississippi, etc., to perform similar clustered water level predictions.

The prediction of the water level of the TGD may become the focus of research on cargo loading, transportation, and ship fuel in the future. It will help predict the berthing and navigation schedules of the ships to ensure safe navigation.

### References

[1] Z. Yuan, J. Liu, Y. Liu, Q. Zhang, Y. Li and Z. Li 'A two-stage modeling method for multi-station daily water level prediction', Environ. Modell. & Softw., vol.156, pp.105468, Oct. 2022.
DOI:10.1016/j.envsoft.2022.105468

[2] S. J. Birkinshaw, S. B. Guerreiro, A. Nicholson, Q. Liang, P. Quinn, L. Zhang, B. He, J. Yin, and H. J. Fowler, "Climate change impacts on Yangtze River discharge at the Three Gorges Dam," Hydrol. Earth Syst. Sci., vol.21, no.4, pp.1911-1927, Apr. 2017.
DOI:10.5194/hess-21-1911-2017

[3] A. N. Ahmed, T. V. Lam, N. D. Hung, N. V. Thieu, O. Kisi, and A. El-Shafie, "A comprehensive comparison of recent developed meta-heuristic algorithms for streamflow time series forecasting problem," Appl. Soft Comput., vol.105, pp.107282, Jul. 2021.
DOI:10.1016/j.asoc.2021.107282

[4] J. C. Magyar and M. Sambridge, "Hydrological objective functions and ensemble averaging with the Wasserstein distance," Hydrol. Earth Syst. Sci., vol.27, pp.991-1010, Mar. 2023.
DOI:10.5194/hess-27-991-2023

[5] F. Wang, G. Huang, Y. Li, J. Xu, G. Wang, J. Zhang, R. Duan, and J. Ren, "A Statistical Hydrological Model for Yangtze River Watershed Based on Stepwise Cluster Analysis," Front. Earth Sci., vol.9, pp.742331, Sep. 2021.
DOI:10.3389/feart.2021.742331

[6] D. K. Vishwakarma, R. Ali, S. A. Bhat, A. Elbeltagi, N. L. Kushwaha, R. Kumar, J. Rajput, S. Heddam, and A. Kuriqi, "Pre- and post-dam river water temperature alteration prediction using advanced machine learning models," Environ. Sci. Pollut. R., vol.29, pp.83321-83346, Jun. 2022.
DOI:10.1007/s11356-022-21596-x

[7] K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," EMNLP., pp.1724-1734, Oct. 2014.
DOI: 10.3115/v1/D14-1179