Yuan, Z, Liu, J, Liu, Y, Zhang, Q, Li, Y and Li, Z

A two-stage modelling method for multi-station daily water level prediction

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Yuan, Z, Liu, J, Liu, Y, Zhang, Q, Li, Y and Li, Z (2022) A two-stage modelling method for multi-station daily water level prediction. Environmental Modelling and Software, 156. ISSN 1364-8152

A Two-Stage Modelling Method for Multi-Station Daily Water Level Prediction

Zhi Yuan[a, b, c], Jingxian Liu[a, b], Yi Liu[a, b*], Qian Zhang[c*], Yue Li[d], Zongzhi Li[e]

[a]Hubei Key Laboratory of Inland Shipping Technology, School of Navigation, Wuhan University of Technology, 1040 Heping Avenue, Wuhan, Hubei 430063, PR China.

[b]National Engineering Research Centre for Water Transport Safety (WTSC), Wuhan University of Technology, 1040 Heping Avenue, Wuhan, Hubei 430063, PR China.

[c]Department of Electronics and Electrical Engineering, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK.

[d]College of Technology, Hubei Engineering University, Xiaogan, Hubei 432000, PR China.

[e]Department of Civil, Architectural, and Environment Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA.

Corresponding author: Yi Liu (liuyi_hy@whut.edu.cn); Qian Zhang (Q.Zhang@ljmu.ac.uk).

Abstract: Water level prediction is an essential task in inland water transportation and infrastructure operation. However, in recent years, the level of uncertainly in the current methods has increased significantly due to infrastructure operation and climate change, therefore, the need to develop more accurate models for multi-station daily water level prediction along the long and volatile inland rivers. This study proposes a two-stage modelling method to improve the accuracy and efficiency in simultaneous prediction of daily water levels for multiple stations. To achieve this goal, this paper takes Yangtze River trunk line as case study. First, we divide the 19 stations along the Yangtze River trunk line into 6 clusters by dynamic time warping (DTW) and hierarchical clustering algorithm (HCA). Then, the long short-term memory (LSTM) network and seasonal autoregressive integrated moving average (SARIMA) model are tailored to construct a multi-station daily water level prediction (MSDWLP) model for each cluster. Finally, to validate the proposed method, the daily water level data of 912 consecutive days from the 19 stations are analysed and predicted in detail. The results demonstrate that the proposed approach can yield more reliable forecasts than traditional deterministic models. Insight from the models can be used to predict daily water levels to better inform decision-making about waterborne transportation, water management, and water emergency response.

Key Words: water level prediction; multi-station; clustering; deep learning; Yangtze River

1. Introduction

Water level prediction related to rivers and coastal waters is highly valuable for waterborne transportation, water resources management, flood mitigation, and water emergency response (Kasiviswanathan et al. 2016; Wang and Zhang, 2018; Gabela and Sarmiento, 2020; Liu et al., 2021). In particular, high and low water levels are a major concern to the safe and efficient operation of inland shipping. Moreover, accurate and efficient water level modelling and prediction is essential for inland water transportation and infrastructure operation (Coraddu et al., 2017; Xu et al., 2017; Li et al., 2020a; Yuan et al., 2021a). As shown in Fig. 1, water levels prediction plays an important supporting role in ship speed control, trajectory prediction, fuel consumption analysis, cargo loading, navigation planning, hydroelectric power, agricultural irrigation and flood mitigation, which has become an important research topic in the fields of ship operation, navigation planning and infrastructure operation. However, accurate prediction of stations daily water level prediction along the long and volatile inland rivers has become a vital challenge since they are affected by many complex factors, such as climate, waterway topography, and periodic characteristics (Yuan et al., 2021b).

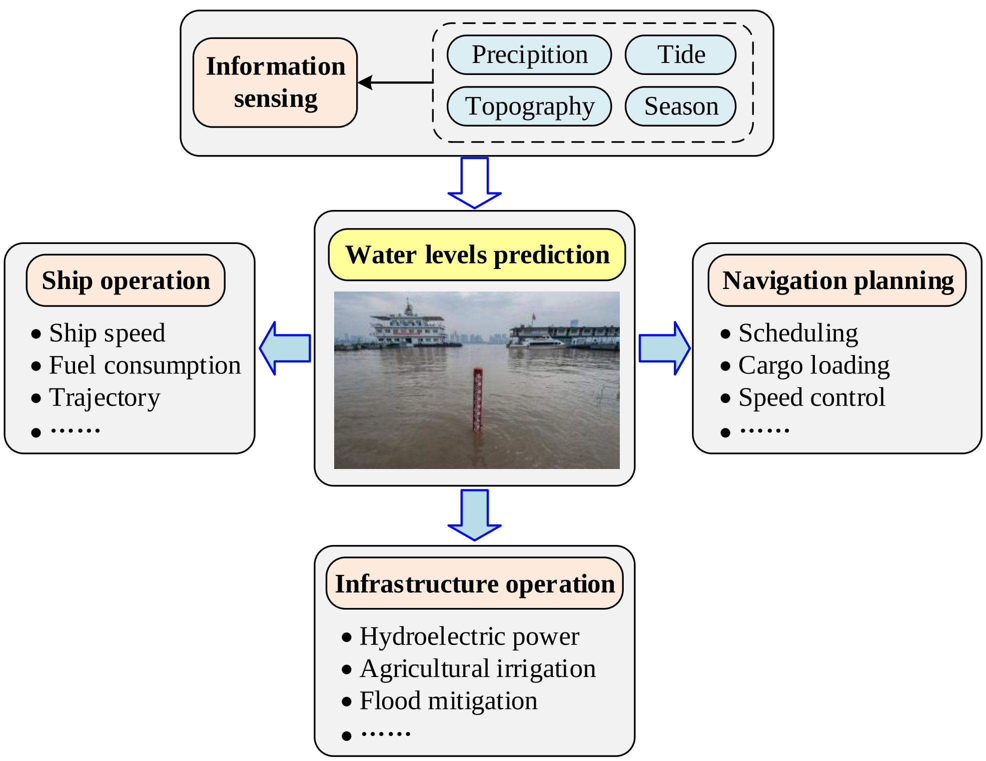


Fig. 1. Water levels prediction in inland water transportation and infrastructure operation.

By analyzing historical data of daily water level, an effective model can be established to predict the future daily water level accurately (Quilty and Adamowski, 2020; Yaseen et al., 2020; Zhu et al., 2020; Ehteram et al., 2021). In some existing literature, regression analysis methods and base models (such as method of lines, auto-regressive, etc.) have been widely used in water level correlation analysis and prediction. For instance, Wei (2015) introducted the locally weighted regression and the *k*-nearest neighbor models, and developed a methodology for formulating water level models to forecast river stages during typhoons. Paul et al. (2018) adapted the method of lines in addition with a newly embedded RKARMS(4,4) (RKAM(4,4) (Runge-Kutta arithmetic mean) and RKRMS(4,4) (Runge-Kutta root mean square)) technique for numerical prediction of water levels considering the effect of tide and surge related to a cyclone. Ebtehaja et al. (2019) presented a novel linear-based model for Lake Level Time Series forecasting and evaluated the performance of the methodology using two case studies of the Van Lake, in Turkey and the Michigan-Huron Lake, in North America. Chen et al. (2020) developed a hybrid model combining the auto-regressive (AR) analysis and the non-stationary tidal harmonic analysis model to improve short-term (with time scale of days) water level predictions in the tide-affected estuaries. The above models achieved good prediction results under certain circumstances. However, the water levels at inland river stations are all characterized by strong nonlinearities, which are difficult to be captured by linear models.

On the other hand, machine learning algorithms and techniques (Jordan and Mitchell, 2015) can accurately capture complex relationships and make predictions, and have been recently developed and implemented to water level prediction research. Zhong et al. (2017) established a hybrid ANN (Artificial Neural Network)-Kalman filtering model for forecasting the water level of Wuhan station, which locates at the middle section of the Yangtze River. Sahoo et al. (2019a) analyzed the suitability of Support Vector Regression for modelling monthly low flows time series for three stations in Mahanadi river basin, India. Zhu et al. (2020) used the feed forward neural network and deep learning technique to predict monthly lake water level. Yang et al. (2020) proposed an Edge COMputing-based Sensory NETwork for water level monitoring and prediction. Li et al. (2020b) proposed the weighted integration based on accuracy and diversity and kernel extreme learning machine algorithm to achieve the forecasting of Xiangjiang River and Yuanjiang River water level. Zhou et al. (2020) employed deep learning technique and multilayer perceptron to perform forecast of Nanjing navigable river's water-level fluctuation. Liu et al. (2021) proposed a hybrid Bayesian vine copula model for daily and monthly water level prediction. Besides, due to the superior learning and memory ability of the LSTM network, it has shown great advantages in time series modelling and predictive analysis (Zhang et al., 2018; Yuan et al., 2020; Adikari et al., 2021), which can provide valuable reference for water level series prediction. However, the above works mainly focus on the water level of a few stations or a certain segment, which can hardly be directly employed to forecast the water level of multi-station for long and volatile inland rivers, such as the Yangtze River trunk line, which contains 19 stations, as shown in Fig. 2.

Fig. 2. The Yangtze River trunk line.

Recently, a new method in the field of water levle series prediction is a combination of linear models and non-linear models (Moeen et al., 2017; Ebtehaja et al., 2019; Xu et al., 2019; Phan and Nguyen, 2020). Moeeni and Bonakdari (2017) combined the linear SARIMA model with the non-linear ANN model to develop a hybrid SARIMA-ANN model, and used the model to improve the prediction accuracy of the monthly inflow to the Jamishan dam reservoir in western Iran. Subsequently, Moeeni et al. (2017a) by considering the different deterministic terms (jump, trend and period) of monthly inflow time series, proposed a hybrid method based on the combination of SARIMA and adaptive neuro-fuzzy inference system (ANFIS). The prediction results approved a higher performance of the proposed SARIMA-ANFIS method in comparison with individual ones. Xu et al. (2019) proposed a combined Auto-regressive Integrated Moving Average-Recurrent Neural Network (ARIMA-RNN) model for water level prediction. Sahoo et al. (2019b) explored the suitability of a proposed LSTM-RNN and artificial intelligence method for low-flow time series forecasting, and used the method to forecast the daily discharged data collected of the Basantapur gauging station located on the Mahanadi River basin, India. Phan and Nguyen (2020) took advantange of linear and nonlinear models, and proposed a hybrid approach combining statistical machine learning algorithms and ARIMA for Red river water level forecasting. The effectiveness of the hybrid models has been verified through performance evaluation of the prediction water level.

In summary, some research on water level prediction have been conducted in existing literature, and some conclusions have been presented, including (1) linear models are difficult to capture the nonlinear relationship between water level series; (2) artificial intelligence methods show capability to capture nonlinear relationships and are widely used in water level prediction; (2) the hybrid model combining linear and nonlinear methods can effectively improve the accuracy of water level prediction. However, due to the complex interaction between waterways conditions and river dynamics, inland rivers' water levels are less predictable than ocean tides and rainfall (Chen et al., 2020). The current water level prediction models and methods rarely consider both the spatial and temporal changes of the water area, and it is difficult to be used for the daily water level prediction of multi-station along inland rivers. Therefore, to achieve simultaneous prediction of daily water level in multiple stations along inland rivers, the following two problems need to be explored and solved:

(1) Build and train a single model for each station, which increases the calculation cost;

(2) Build and train a single model for all stations, which reduces the prediction accuracy.

To address the above mentioned two problems, this study makes two main contributions, including:

(1) Clustering stations with similar characteristics, which reduces the calculation cost of modelling;

(2) Tailor a single prediction model for each cluster stations, including hybrid model, which improve the accuracy of simultaneous prediction of water level in multiple stations.

To achieve this goal, a two-stage divide-and-conquer method for daily water level of multi-station analysis and prediction is proposed, and a real-world case study of daily water level forecasting for 19 stations along the Yangtze River trunk line is presented. The hope is that the research framework, modelling strategy and experimental design presented herein can be informing for other researchers or practitioners to explore simultaneous predictions of various hydrology and water resources indicators from multiple related stations.

The remaining of this paper is organised as follows: In Section 2, the collected daily water level data of 19 stations are presented and analysed. In Section 3, the modelling strategy and methods are described. Section 4 provides experimental details concerning the case study. The detailed experiments and discussion about the water level modelling are presented in Section 5. Finally, conclusions and perspectives are drawn in Section 6. Table 1 summarises the abbreviations used in the paper.

Table 1. A list of abbreviations.

Abbreviation	Meaning	Abbreviation	Meaning
ANFIS	Adaptive Neuro-fuzzy Inference System	LSTM	Long Short-Term Memory
ADF	Augmented Dickey-Fuller	MA	Moving Average

AE	Absolute Error	MAE	Mean Absolute Error
AIC	Akaike Information Criterion	MAPE	Mean Absolute Percentage Error
ANN	Artificial Neural Network	MSDWLP	Multi-station Daily Water Level Prediction
AR	Auto-regressive	NSE	Nash–Sutcliffe efficiency coefficient
ARIMA	Auto-regressive Integrated Moving Average	RMSE	Root Mean Square Error
ARMA	Auto-regressive Moving Average	RKAM	Runge-Kutta Arithmetic Mean
DTW	Dynamic Time Warping	RKRMS	Runge-Kutta Root Mean Square
HCA	Hierarchical Clustering Algorithm	RNN	Recurrent Neural Network
IDE	Integrated Development Environment	SARIMA	Seasonal Auto-regressive Integrated Moving Average

2. Study area and data

2.1. Study area

The Yangtze River is the longest inland river in China, which is known as the "golden waterway". The annual freight volume of this river ranks first among the world's inland rivers (Notteboom et al., 2020), reaching 2.69 billion tons in 2019. The Yangtze River trunk line has become the key area of shipping and water research, where the prediction of daily water level along the line is one of the key problems that need to be solved urgently at present.

As shown in Fig. 3, the Yangtze River is 6,387 kilometres long and it is the longest river in Asia. The main stream of the Yangtze River traverses central China from west to east, between 90°33′-122°25′ east longitude and 24°30′-35°45′ north latitude. It flows through 11 provincial administrative regions, including Qinghai, Tibet, Sichuan, Yunnan, Chongqing, Hubei, Hunan, Jiangxi, Anhui, Jiangsu and Shanghai, and finally empties into the East China Sea. The trunk line of the Yangtze River is characterised by winding and uneven terrain. In some sections, the mountains are high, the valleys are deep and the current flows fast, while other sections have gentle slopes and gentle water. Therefore, the entire Yangtze River trunk line is divided into three water areas: upper reach, middle reach and lower reach. From the source to the Yangtze River Estuary, the a.s.l. (above sea level) of the Yangtze River mainstream gradually decreases. Among them, the section from Yibin station to the estuary is the Yangtze River trunk line, mainly including 19 stations, which can be seen in Fig. 2.

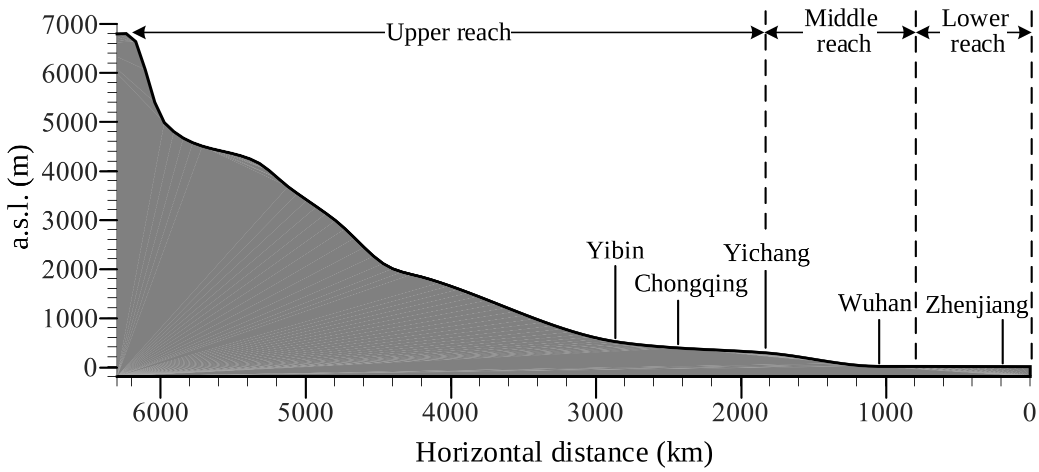


Fig. 3. Profile map of the Yangtze River mainstream.

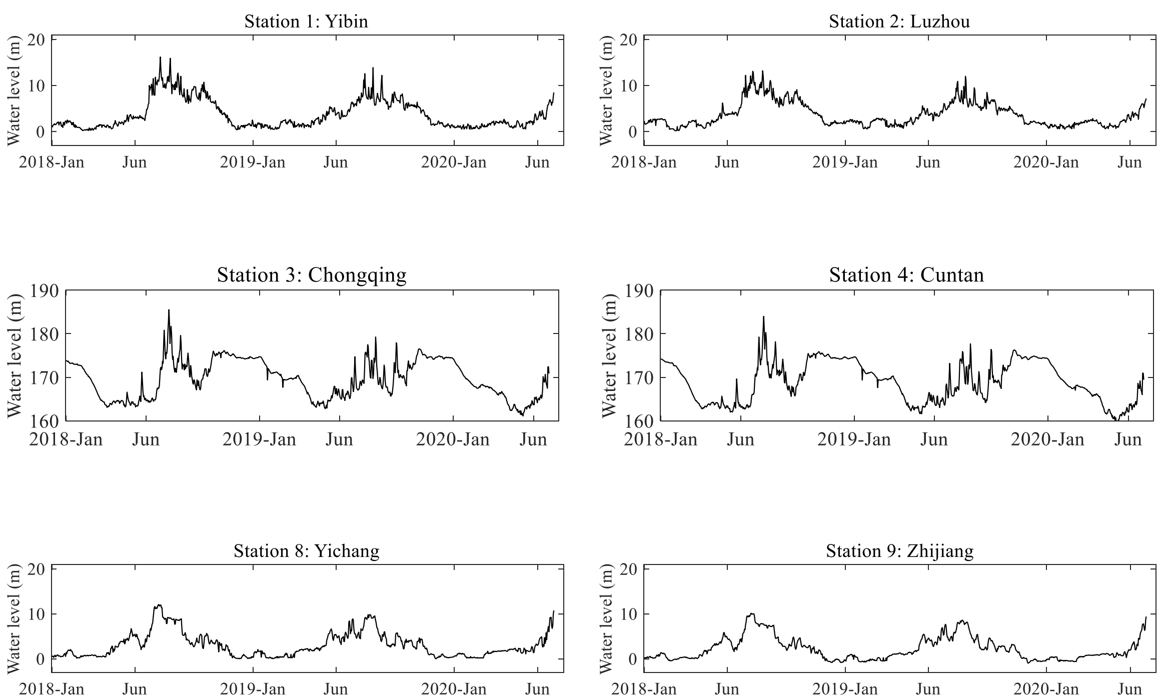
2.2. Water level data

In order to study the daily water level of the entire Yangtze River, this research collected the daily water level data of 19 stations along the main stream of the Yangtze River, i.e. Yibin, Luzhou, Chongqing, Cuntan, Fuling, Wanzhou, Maoping, Yichang, Zhijiang, Shashi, Jianli, Chenglingji, Hankou, Huangshi, Jiujiang, Anqing, Wuhu, Nanjing and Zhenjiang (as shown in Fig. 2). In this paper, the 19 stations are indexed sequentially from station 1 to station 19, where station 1 is Yibin and station 19 is Zhenjiang. The summary statistics for the 19 stations as shown in Table 2.

Table 2. Statistics of 19 Stations along the Yangtze River Trunk Line.

	Stations	a.s.l. (m)	Mileage (km)	Waterway level
Upper reach	Station 1~2	200~500	384	III
	Station 3~8	160~200	660	II
Middle reach	Stations 9~13	15~50	623.5	II
Lower reach	Stations 14~19	2~10	1020.3	I

From the topographical point of view, the 19 stations are distributed along three different parts of the Yangtze River trunk line: stations 1-8 located in the upper reach, stations 9-13 located in the middle reach and stations 14-19 located in the lower reach. It is worth noting that the terrain, topography and water flows of each station in different water areas have different characteristics. Specifically, the terrain in the upper reaches is high and steep, with an a.s.l. of 3000-4000m in some places, while the a.s.l. suddenly drops to 200-600m at the edge of the Sichuan Basin. The river enters the hilly area in the middle reach, with densely covered beaches and branching water flows. The valleys of certain river sections are several kilometres wide, and the river surface is 155-500m wide. The lower reaches of the river have gentle water and wide river surface and are with low mountains and wide valleys. In addition, the data of daily water level in different water areas have different value ranges. For example, the water levels of stations 3-7 are all above 145 meters, because they are in the reservoir area of the Three Gorges Dam. The highest water level of stations 5, 6 and 7 exceeds 170 meters, while the highest water level is less than 6 meters in station 19. The water level data come from the daily records issued by the Ministry of Transport of the People's Republic of China[1], which include the daily water level for two and a half years (912 days), form January 1, 2018 to June 30, 2020. The daily water level data of each station are shown in Fig. 4.



[1] http://www.mot.gov.cn/shuiluchuxing/changjiangshuiweigonggao/

Fig. 4. Daily water level of 19 stations.

It can be seen from Fig. 4 that the water level data are time-series data, because they present two obvious characteristics: the data of water level changes with time and the data are interrelated. Therefore, the analysis and prediction of the daily water level of 19 stations is the analysis and prediction of 19 different time series. To improve the efficiency in analysis and prediction of multi-station daily water level and improve the applicability of the proposed prediction method, it is considered to build one model for several stations with similar water level sequence rather than build one model for each station.

3. Methodology

The research framework is as shown in Fig. 5. Firstly, the daily water level data of 19 stations along the Yangtze River trunk line are collected. After specific analysis, some feature information of the water level data is obtained. To avoid building separate models for all stations, a strategy of divide and conquer is proposed. The DTW algorithm is emploed to measure the similarity of the water level data of multiple stations, and a HCA is utilised to group the stations with similar characteristics according to the similarity matrix. Then, the MSDWLP models are constructed based on the LSTM network and the SARIMA method for different clusters. To improve the prediction accuracy of each station, StatsModels algorithm (Lemenkova, 2019) is employed to decompose the water level series (training set) into trend, period and residual. For the complex water levels with shorter periodicity (2~20 days, because a voyage time of the ship in the Yangtze River trunk line is about 20 days), LSTM is used to approximate the trend and SARIMA is used to approximate the residual term (sum of period and residual). Finally, the daily water levels of the 19 stations in the next 7 days are predicted and analysed in detail.

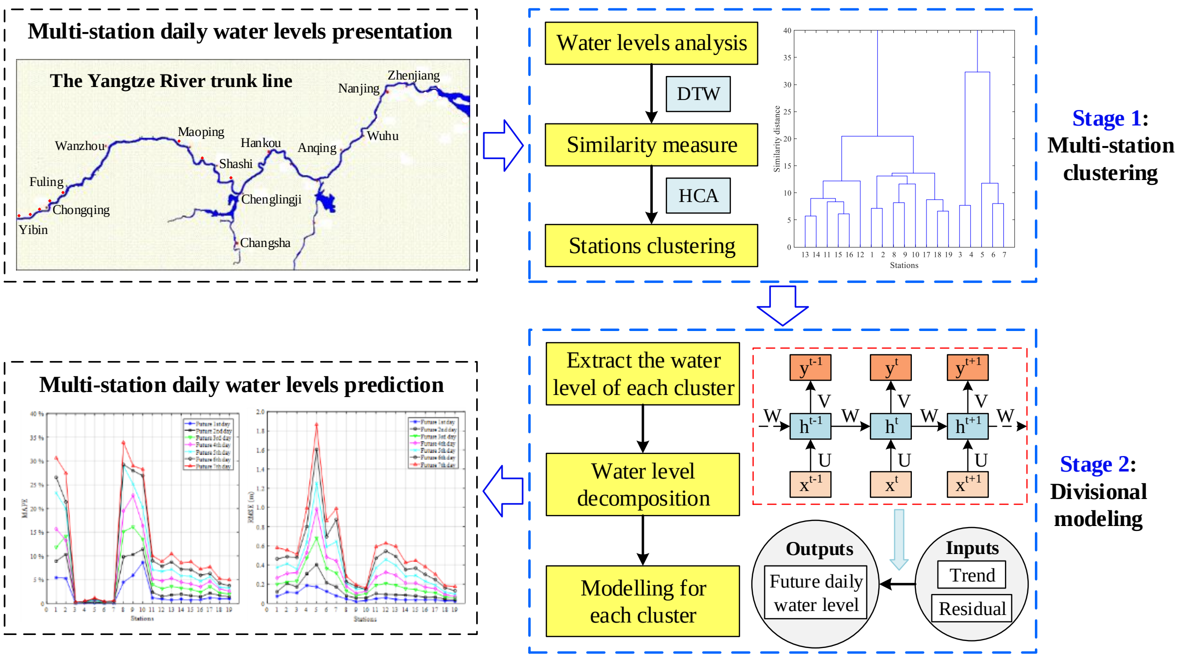


Fig. 5. The research framework for daily water level prediction of multi-station along the Yangtze River trunk line.

3.1. Modelling strategy

In this paper, a two-stage modelling strategy is proposed as shown in Fig. 6.

At the first stage, the water level data from multiple stations are clustered based on similarity measurement using a hierarchical clustering algorithm. Stations with a similar trend in water level data would be integrated into a single prediction model.

At the second stage, daily water level of each cluster will be firstly decomposed into an additive model, for the clusters whose data have a long periodicity, such as yearly or longer, or have no obvious periodicity, one may preferentially adopt an efficient LSTM network in modelling. Contrarily, a LSTM-SARIMA method is proposed for the clusters with shorter periodicity. In this method, a daily water level will be decomposed into a long-term trend, a period change trend and residual. Then, different parts of the decomposition are processed using different methods. In particular, the long-term trend is modelled using the LSTM network, and the residuals including period and residual is modelled with SARIMA. And the prediction results of each part are combined to obtain the integral water level prediction.

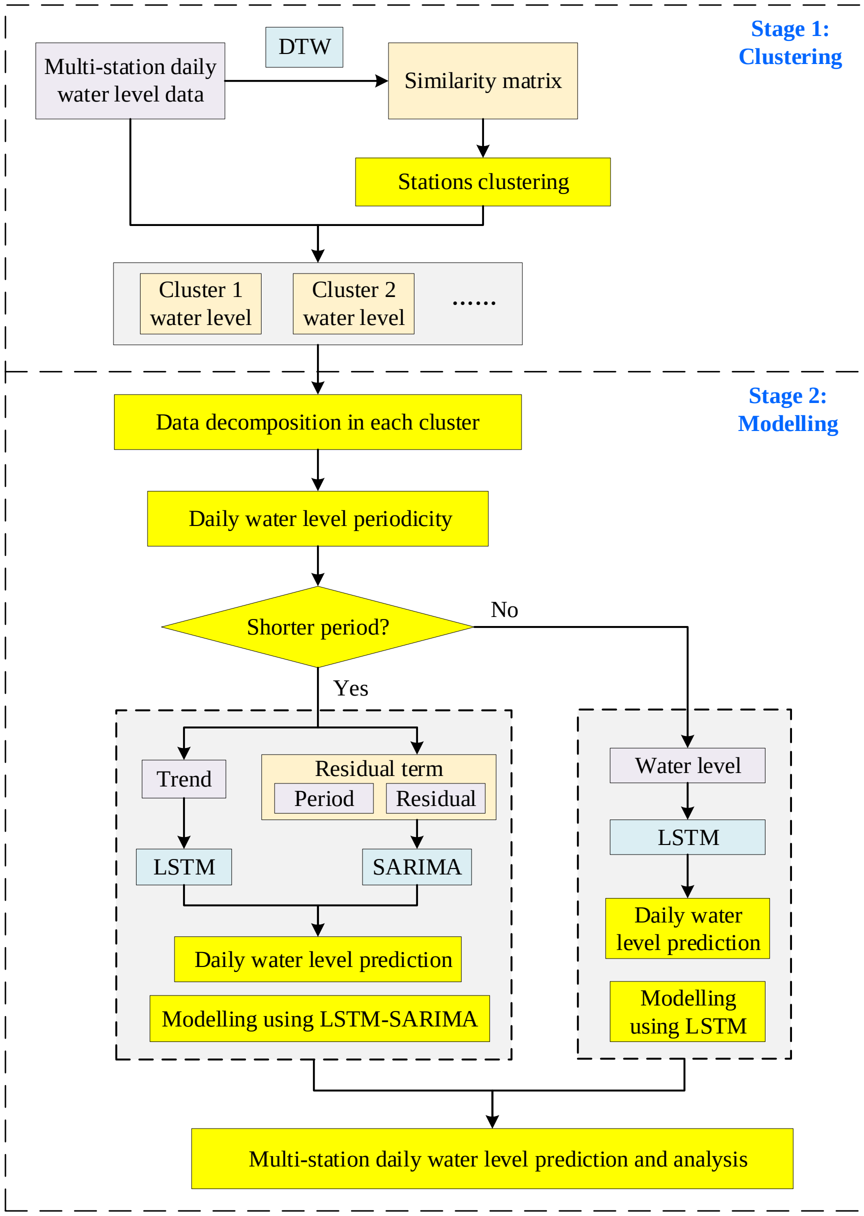


Fig. 6. The proposed two-stage modelling strategy.

3.2. Clustering method

HCA (Zhou et al., 2017) is a clustering technique that does not need to consider the selection of the number and positions of initial clusters. The hierarchical clustering algorithm is very intuitive, i.e. clustering layer by layer. The core of an agglomerative clustering algorithm is to start with individual clusters and merge the two closest clusters at each step. The clustering algorithm used for grouping stations is designed based on the hierarchical clustering principle and is shown below.

Algorithm 1. The clustering algorithm for stations.

1. Calculate the similarity matrix of the water level of stations

2. Repeat

3. Merge the two nearest clusters

4. Update the similarity matrix to reflect the proximity between the new cluster and the original clusters

5. Until only one cluster remains

It can be seen from Algorithm 1 that we need to first calculate the similarity matrix between the water levels of the 19 stations. It is worth noting that the number of the water level data is the same for all stations. That is to say, the water level data of the 19 stations are 19 time series of equal length. Therefore, we can employ the DTW algorithm to calculate the similarity distance between the water level sequences of different stations. DTW is a dynamic programming algorithm suitable for accurately calculating the similarity between multiple series (Dürrenmatt et al., 2013; Yu et al., 2018). In addition, DTW has no parameter restrictions and is robust to time. Assuming that S_1 and S_2 are the water level sequences of station 1 and station 2, the similarity distance matrix between S_1 and S_2 is calculated as follows:

$$(DP[i][j])^2 = \begin{cases} (S_1[0] - S_2[0])^2 & i = 0, j = 0 \\ (S_1[0] - S_2[j])^2 + DP[0][j-1] & i = 0 \\ (S_1[i] - S_2[0])^2 + DP[i-1][0] & j = 0 \\ (S_1[i] - S_2[j])^2 + \min(DP[i-1][j], DP[j-1][i], DP[i-1][j-1]) & i, j > 0 \end{cases}$$

where DP is the similarity matrix, i and j are the indices of S_1 and S_2, $DP[i][j]$ is the similar distance between $S_1[i]$ and $S_2[j]$.

3.3. Predictive modelling methods

3.3.1. Water level time series decomposition

In general, time series prediction technology is to study the variation trend and law of the target series by processing the historical data, so as to predict the data of the future time. Assuming that $T(t)$ represents the long-term trend item of the time series $y(t)$, $P(t)$ represents the item of periodic change trend, and $R(t)$ represents the random interference item, there are three common time series models (Wang et al., 2017) as follows:

(1) Addition model: $y_t = T_t + P_t + R_t$.

(2) Multiplication model: $y_t = T_t \cdot P_t \cdot R_t$.

(3) Mixed model: $y_t = T_t \cdot P_t + R_t$ or $y_t = T_t + P_t \cdot R_t$.

where y_t is the observation record of the target sequence.

In this work, the addition format and StatsModels algorithm were employed for water level time series decomposition, and the water level of each station was decomposed into trend, period and residual: $W(t) = T(t) + P(t) + R(t)$. The specific decomposition steps are as follows:

Step 1: Decompose trend items by using the centralized moving mean method.

Step 2: Subtract the trend term from the original water level, average the values at the same frequency in each period to obtain the periodic term, and further centralize to obtain the period term of the original water level.

Step 3: Calculate the residuals: $R(t) = W(t) - T(t) - P(t)$.

3.3.2. LSTM

The LSTM network is one type of RNN, which was designed to tackle the problem of gradient dispersion existing in the conventional RNNs. It has strong ability of learning and memory, and can efficiently process the sample data of time series (Fang et al., 2020; Yuan et al., 2020).

The LSTM network contains three control gates: $Igate_t$, $Fgate_t$ and $Ogate_t$, and two transmission states: C_t and Hid_t, where t is time. $Igate_t$ represents the input gate at time t, which is a control gate from the previous long-state information to the current long-state information. It is used to control how much new information is saved. $Fgate_t$ represents the forget gate at time t, which is also a control gate from the previous long-state information to the current long-state information. It is used to control how much history information is forgotten. $Ogate_t$ represents the output gate at time t, which is a control gate from the current information to the output-state information. The summation of the long-state information and the short-state information is the current information. C_t is named “cell state”, which memorises information, and C_{t-1} is the cell memory at the time point $t-1$. Hid_t is named “hidden state”, which represents the output of the hidden node. In_t represents the input sequence that can be a series with one dimension or multiple dimensions. f is a sigmoid activation function and h is a hyperbolic tangent activation function. The calculation of control gates and transmission states are shown in equations (1)-(5) (Gers et al., 2000).

$$Igate_t = f(W_I In_t + U_I Hid_{t-1} + b_I) \tag{1}$$

$$Fgate_t = f(W_F In_t + U_F Hid_{t-1} + b_F) \tag{2}$$

$$Ogate_t = f(W_O In_t + U_O Hid_{t-1} + b_O) \tag{3}$$

$$C_t = Fgate_t \odot C_{t-1} + Igate_t \odot h(W_C In_t + U_C Hid_{t-1} + b_C) \tag{4}$$

$$Hid_t = Ogate_t \odot h(C_t) \tag{5}$$

where $W_{\{I,F,O,C\}}$ are the weight matrices linking the input layer with the hidden layer, $b_{\{I,F,O,C\}}$ are offset weight matrices and $U_{\{I,F,O,C\}}$ are the self-looping weight matrices of the hidden layer (Yuan et al., 2021a).

The specific modelling steps for daily water level using the LSTM network are designed as follows.

Step 1: Extracting the data of the daily water level of each station according to the station index.

Step 2: Normalising data. The data are normalised to be between 0 and 1 to eliminate the adverse effects of singular sample data.

Step 3: Generating time series. In this step, the time series is generated according to time, time steps and batch size. The format of the time series is [samples, time steps, features], which the LSTM network adapts to.

Step 4: Setting the parameters of the LSTM network, including the number of neurons, the number of hidden layers, epochs (training iterations), activation function, training function and loss function.

Step 5: Training and optimising the initial LSTM network until termination criteria are satisfied.

Step 6: Predicting the future data in a single step or multiple steps using the trained LSTM network. It should be noted that the prediction results need to be denormalised.

Step 7: Results evaluation and analysis. The prediction accuracy is analysed by calculating some performance measures.

3.3.3. SARIMA

The SARIMA model is an evolution model of seasonal or periodic data based on the ARIMA model (Sun et al., 2020). The ARIMA model refers to the model established by transforming non-stationary time series into a stationary time series, and then regressing the lag value of the dependent variable (Phan and Nguyen, 2020), and the present value and lag value of the random error term, including AR process, Moving Average (MA) process, Auto-regressive Moving Average (ARMA) process and ARIMA process (Velasco and Lazakis, 2020).

The expression of the SARIMA model can be written as $SARIMA(p,d,q)(P,D,Q)_s$, where, p, P, q, and Q represent the maximum lag order of non-seasonal, seasonal, autoregressive and moving average, respectively; d and D represent the order of the differentiate and seasonal difference, respectively; s denotes the period of the seasonal time series. It is worth noting that in the process of modelling and analysis using the SARIMA method, the values of parameters p, P, d, D, q, and Q are not very large, and d and D usually take 0 and 1 to meet the requirements.

The specific modelling steps for time series using the SARIMA model are designed as follows.

Step 1: Visual analysis of time series data. The time series diagram of the data is drawn to visualise the trend of sequence changes over time.

Step 2: Test the stationarity of the data. In statistics, the Augmented Dickey-Fuller (ADF) test is a common and effective method for testing sequence stationarity.

Step 3: Sequence stabilisation. The stabilisation of the time series is to eliminate the trend effect and seasonal effect of the sequence, and the differencing method is the most common method to achieve sequence stabilisation.

Step 4: Model order determination, that is, determining the parameters of the SARIMA model: p, P, d, D, q and Q. This is a very critical step, this paper uses the network search method to systematically select the optimal values of the parameters. At the same time, the Akaike Information Criterion (AIC) is selected as the criterion for selecting the best model parameters. The AIC not only improves the degree of model fitting, but also introduces a penalty term to make the model parameters as few as possible, which is helpful to reduce the possibility of overfitting.

Step 5: Building of the SARIMA model according to the optimal parameters determined in step 4.

Step 6: Model testing. Verify if the residuals of the model are correlated and they are normally distributed with zero mean. If not, the model can be further improved.

Step 7: Data prediction. Use the constructed SARIMA model to predict the future data of the time series.

Step 8: Results evaluation and analysis.

3.4. Data partitioning

In this study, water level data with 912 rows and 19 columns were collected, which came from the daily water level records of 19 stations on the Yangtze River trunk line for 912 days (from January 1, 2018 to June 30, 2020). In our strategy, the daily water level data of two and a half years were treated as time series, and were divided into two parts. Therefore, the 80% data from the first 2 years (January 1, 2018 to December 31, 2019) were used as training data, and the 20% data from the next half year (January 1, 2020 to June 30, 2020) were used as testing data.

In our methodology, the training data is for LSTM network and SARIMA model training, and the testing data is for models validation. In the constructed models, the variables presented to the models are vectors composed of the consecutive days daily water level data from the 19 stations. In specifically, the daily water levels of training set are cyclically extracted in chronological order, and constructed into vectors with length $timestep + n\_features$ and presented to the basic LSTM network, where, the first $timestep$ variables are the input and the last $n\_features$ variables are the output. The daily water levels of testing set are divided into 26 groups and presented to the trained models, where, the length of each group of variables is $timestep$, and 26 groups of water level prediction values for 7 consecutive days are obtained through loop prediction. The framworks of LSTM model training and testing are shown in Fig. 7 and Fig. 8, where, WL represents water level, ts represents $timestep$, and other symbols are described in Section 3.3.2.

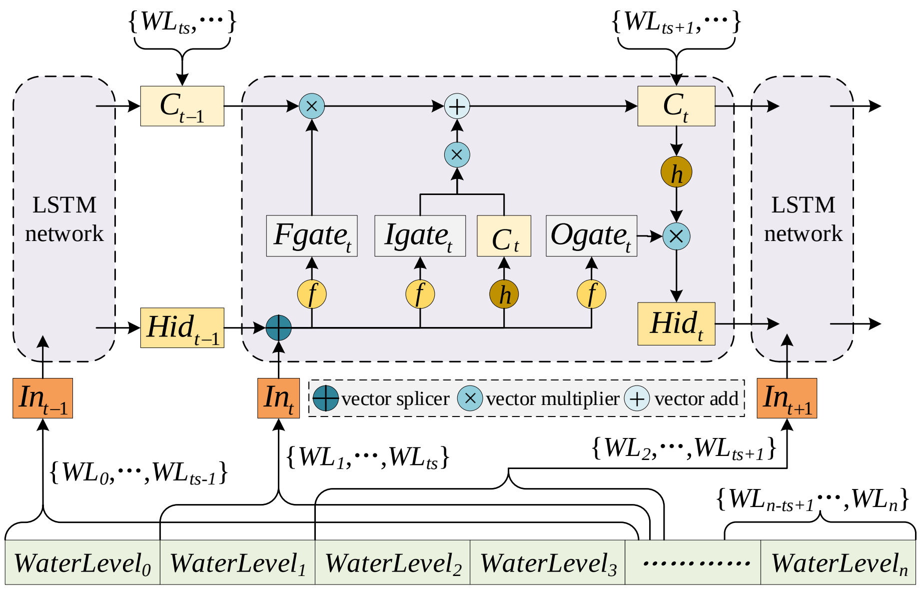


Fig. 7. Training framework of LSTM network.

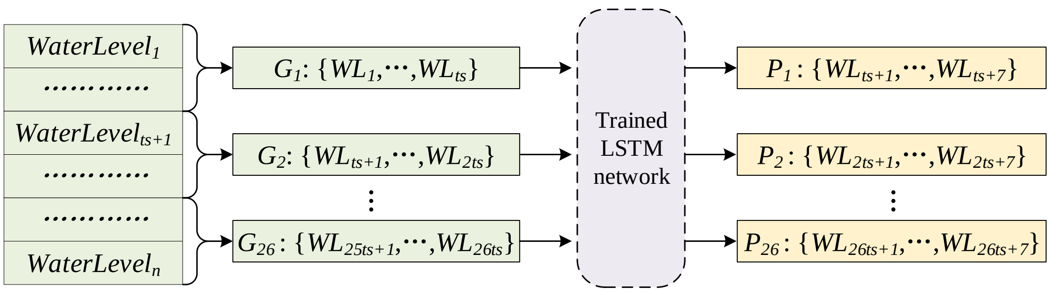


Fig. 8. Testing framework of trained LSTM network.

3.5. Performance measures

To assess the accuracy and reliability of the proposed MSDWLP models, the prediction results are analyzed by several performance measures (Ebtehaja et al., 2019), including *AE* (Absolute Error), *MAE* (Mean Absolute Error), *MAPE* (Mean Absolute Percentage Error), *RMSE* (Root Mean Square Error) and *NSE* (Nash–Sutcliffe efficiency coefficient), as shown in Equations (6)-(10).

$$AE = |z_t - \hat{z}_t| \tag{6}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|z_t - \hat{z}_t| \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|z_t - \widehat{z_t}|}{z_t} \times 100\% \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(z_t - \widehat{z_t})^2} \tag{9}$$

$$NSE = 1 - \frac{\sum_{t=1}^{n}(z_t - \widehat{z_t})^2}{\sum_{t=1}^{n}(z_t - \overline{z_t})^2} \tag{10}$$

where t represents the time index of a datum, $t = 1,2,\cdots,n$, n represents the length of the prediction sequence, $\widehat{z_t}$ and z_t are the predicted value and the measured value of the tth datum, and $\overline{z_t}$ is the mean of z_t.

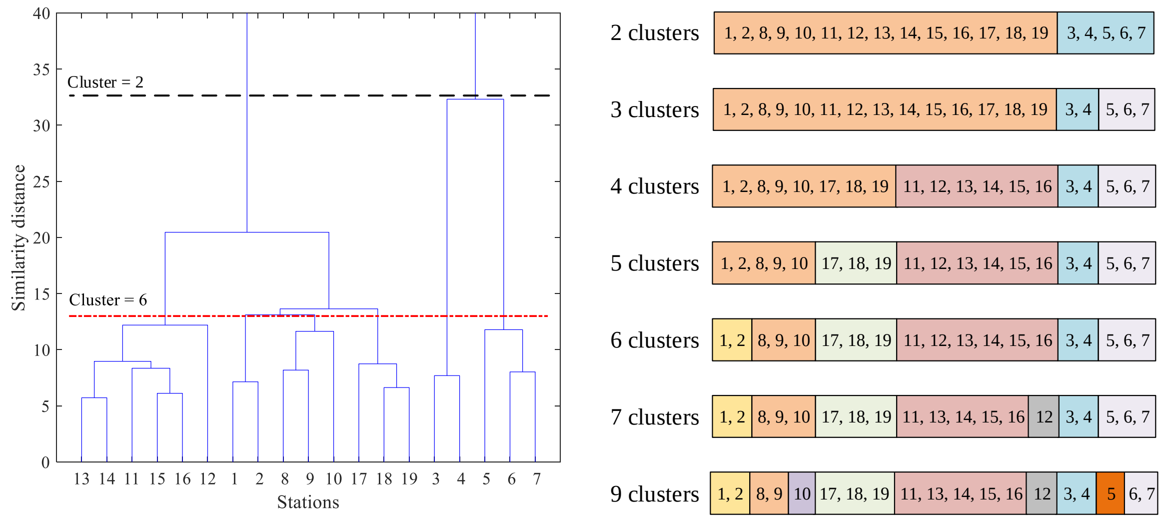
In addition, to supplement these performance evaluation, scatter plots and time series plots are also inserted for a graphical demonstration.

4. Experimental details

In this section, details concerning the experiment study are presented, mainly including stations clustering, model settings and experimental design. The data analysis and modelling experiments were conducted using a desktop PC with Intel Core i7-7700 CPU and 16GB RAM main memory. Its operating system was 64-bit Windows 10 and the programming language employed was Python 3.7, where a Python IDE (integrated development environment) Spyder and an open-source library Keras and sklearn were employed.

4.1. Stations clustering

First of all, the DTW-based hierarchical clustering method was employed for the clustering of daily water level sequences from 19 stations. For the daily water level data (912 consecutive days from the 19 stations, which from January 1, 2018 to June 30, 2020), using the DTW and clustering algorithms introduced in Section 3.2, the hierarchical clustering tree of the 19 stations obtained from the similarity distance matrix is shown in Fig. 9 (a). Different clustering results can be obtained by moving a cross-cut line up and down. For example, moving the cross-cut line to the top, as shown by the black line in the figure, the 19 stations are divided into two rough clusters: stations 3, 4, 5, 6 and 7 as one cluster and the other stations as another cluster. In fact, stations of 3, 4, 5, 6 and 7 are all located in the Three Gorges Reservoir area, which have much higher daily water levels (>145 meters) than the other stations. Moreover, the overall water levels at stations 3 and 4 are higher and more complex than that of stations 5, 6, and 7, which can be seen from Fig. 4. Therefore, stations 3, 4, 5, 6 and 7 are further divided into two clusters: stations 3 and 4 as one cluster and stations 5, 6 and 7 are with another cluster. At this time, the 19 stations were divided into 3 clusters. Moreover, moving the cross-cut line down to the position of the red line, 6 clusters can be obtained. Fig. 9 (b) shows 7 different clustering results for 19 stations, including 2, 3, 4, 5, 6, 7 and 9 clusters.



(a) Hierarchical clustering tree (b) Clustering results

Fig. 9. Hierarchical clustering tree and clustering results of 19 stations.

As can be seen from Fig. 9, 19 stations can be clustered into at least 2 clusters and at most 19 clusters (moving the cross-cut line to the bottom of Fig. 9 (a)). However, for multi-station daily water level prediction, the fewer the clusters, the less distinct the water level features of the stations are; the more clusters, the more expensive the modelling and computation. For example, with 9 clusters, stations 5, 6, and 7 are divided into two clusters. However, their water level ranges and trends did not change significantly (as shown in Fig. 4), and one model could be constructed for all 3 stations, as well as stations 8, 9, and 10. Therefore, considering the performance and computational cost, 6-cluster result is accepted for follow-up research (in additional experiments, it is also verified that 6-cluster is an ideal clustering results).

4.2. Model settings

According to the proposed divide-and-conquer modelling strategy, 4 types of MSDWLP models are constructed and tested. To verify the effectiveness of the proposed method which use LSTM for the trend components and SARIMA for the residual component in our strategy, different MSWLP models for are set comparison, including MSDWLP_S, MSDWLP_CS, MSDWLP_CH1 and MSDWLP_CH2, which are described as follows:

MSDWLP_S denotes the MSDWLP model that constructed by using LSTM alone for all 19 stations without, including one daily water level prediction model.

MSDWLP_CS denotes the MSDWLP models that constructed by using LSTM alone for each cluster based on stations clustering results, including 6 daily water level prediction models.

MSDWLP_CH1 denotes the MSDWLP models that constructed by combining LSTM and SARIMA for each cluster based on stations clustering results, where the LSTM is used for the trend components and SARIMA is used for the residual term components (period and residual), including 6 daily water level prediction models.

MSDWLP_CH2 denotes the MSDWLP models that constructed by combining LSTM and SARIMA for each cluster based on stations clustering results, where the SARIMA is used for the trend components and LSTM is used for the residual components (period and residual), including 6 daily water level prediction models.

In the next sub-section, the model parameters settings of the models are introduced.

4.3. Parameter settings

The optimal parameters are crucial to the performance improvement of the model. For a high-performing LSTM network, there are multiple hyperparameters and activation functions that need to debugged and optimized (Yuan et al., 2020). In the experiments, the parameters and their candidate values are described as follows:

Neurons: the number of neurons of the LSTM network.

Time step: the length of the input time series of the LSTM network.

n_features: the length of the output time series of the LSTM network.

Batch size: the size of each input data of the LSTM network.

Training epochs:the training times of the LSTM network, that is, number of iterations.

Learning rate: control the rapid convergence of the LSTM network model to the optimum, generally ranging from 0.001 to 0.1.

Activation functions: *relu*, a rectified linear unit function; *linear,* a linear activation function; *tanh*, a hyperbolic tangent function; *softsign*, similar to *tanh* but smoother; *sigmoid*, a common s-type function.

Loss functions: *mae*, mean absolute error; *mse,* mean squared error; *mape,* mean absolute percentage error; *msle,* mean squared logarithmic error; *hinge* and *squared hinge.*

Optimisation algorithms: *rmsprop*, root mean square propagation optimiser; *adam*, adaptive moment estimation; *adamax*, a variant of *adam* with infinity norm; *nadam*, nesterov-accelerated adaptive moment estimation; *adagrad*, adaptive gradient algorithm; *adadelta*, extension of *adagrad* with smaller learning rate.

In addition, to avoid overfitting of the network, the early stopping method is used to improve the generalization performance of the LSTM network. The specific steps are as follows: (1) divide the original training data set into training set and validation set; (2) calculate the error of the validation set in each Batch size period, and stop training if the error increases; (3) use the parameters from the previous iteration result as the final parameters of the LSTM.

Parameters settings are into different groups, and a network search algorithm (NSA) is designed based on the minimum error criterion and AIC to systematically select the optimal values of the parameters. For example, Table 3 records the *NSE* of the LSTM network parameter setting experiments for the daily water level prediction of station 1, where, activation function, optimiser function, time step, batch size, and neurons number are determined in group 1, 2, 3, 4 and 5. As shown in Table 3, the following parameters settings is found to be appropriate: the neurons was set to 150, the batch size was set to 36, the time step was set to be 6, the activation function was the rectified linear unit function "*relu*", the loss function was set to *mse*, and the optimization algorithm was the nesterov-accelerated adaptive moment estimation optimiser "*nadam*".

Table 3. Experimental records of LSTM network parameter settings (for station 1).

Group	Neurons	Batch size	Time step	Activation function	Optimization algorithm	*NSE*
1	150	36	5	***linear***	*rmsprop*	0.784
	150	36	5	***relu***	*rmsprop*	**0.801**
	150	36	5	***sigmoid***	*rmsprop*	0.692
	150	36	5	***tanh***	*rmsprop*	0.746
	150	36	5	***softsign***	*rmsprop*	0.684
2	150	36	5	*relu*	***adagrad***	0.802
	150	36	5	*relu*	***adadelta***	0.794
	150	36	5	*relu*	***adam***	0.782
	150	36	5	*relu*	***adamax***	0.806
	150	36	5	***relu***	***nadam***	**0.815**

3	150	36	**4**	*relu*	*nadam*	0.818
	150	36	**6**	***relu***	***nadam***	**0.868**
	150	36	**7**	*relu*	*nadam*	0.851
	150	36	**8**	*relu*	*nadam*	0.836
4	150	**24**	6	*relu*	*nadam*	0.835
	150	**30**	6	*relu*	*nadam*	0.838
	150	**42**	6	*relu*	*nadam*	0.842
	150	**48**	6	*relu*	*nadam*	0.841
5	**130**	36	6	*relu*	*nadam*	0.814
	140	36	6	*relu*	*nadam*	0.817
	160	36	6	*relu*	*nadam*	0.825
	170	36	6	*relu*	*nadam*	0.821

Similarly, $SARIMA(p,d,q)(P,D,Q)_s$ model has 7 key parameters that need to be set (Sun et al., 2020), namely the maximum lag order of non-seasonal p, the maximum lag order of autoregressive q, the maximum lag order of seasonal P, the maximum lag order of moving average Q, the order of the differentiate d, the order of seasonal difference D, and the period of the time series s. The best parameters p, P, d, D, q and Q are also determined by a network search method, which is based on a variant of Hyndman-Khandakar algorithm (Hyndman et al., 2008).

After a series of preliminary experiments for different models, the following parameter settings were found to be appropriate: the Neurons was set to 150, the batch size was set to 36, n_features was set to 1 (cyclic rolling method is used to predict the daily water level in future days), training epochs was set to 1000, learning rate was set to 0.002, and other parameter settings are shown in Table 4.

Table 4. Optimal parameter settings for different MSDWLP models.

Model	LSTM				SARIMA		
	Time step	Activation function	Loss function	Optimization algorithm	(p,d,q)	(P,D,Q)	s

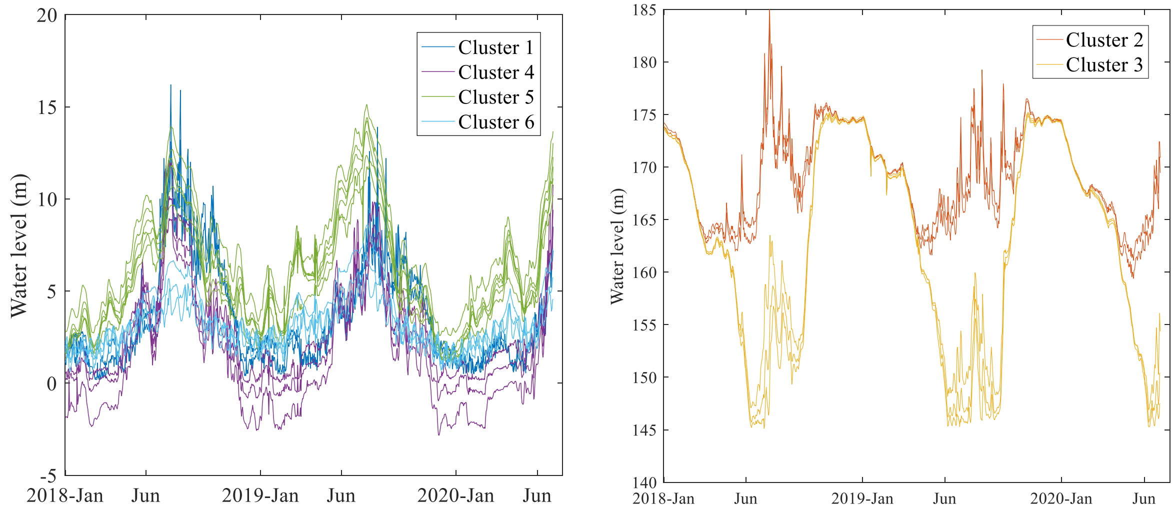
MSDWLP_S		8	*relu*	*mse*	*nadam*	--	--	--
MSDWLP_CS	Cluster 1	6	*linear*	*mse*	*nadam*	--	--	--
	Cluster 2	6	*relu*	*mse*	*adam*	--	--	--
	Cluster 3	5	*linear*	*mse*	*adagrad*	--	--	--
	Cluster 4	10	*tanh*	*mse*	*adagrad*	--	--	--
	Cluster 5	8	*relu*	*mse*	*nadam*	--	--	--
	Cluster 6	10	*linear*	*mse*	*rmsprop*	--	--	--
MSDWLP_CH1	Cluster 1	6	*relu*	*mse*	*nadam*	(2,1,1)	(3,1,0)	7
	Cluster 2	6	*linear*	*mse*	*adam*	(2,1,1)	(3,1,0)	8
	Cluster 3	5	*relu*	*mse*	*nadam*	--	--	--
	Cluster 4	10	*tanh*	*mse*	*rmsprop*	--	--	--
	Cluster 5	8	*relu*	*mse*	*nadam*	--	--	--
	Cluster 6	10	*relu*	*mse*	*nadam*	(2,0,1)	(2,1,1)	15
MSDWLP_CH2	Cluster 1	7	*linear*	*mse*	*adam*	(2,2,1)	(3,2,0)	7
	Cluster 2	8	*relu*	*mse*	*nadam*	(2,1,2)	(3,1,1)	8
	Cluster 3	5	*tanh*	*mse*	*rmsprop*	--	--	--
	Cluster 4	10	*relu*	*mse*	*rmsprop*	--	--	--
	Cluster 5	8	*relu*	*mse*	*adagrad*	--	--	--
	Cluster 6	15	*linear*	*mse*	*nadam*	(2,1,1)	(2,0,1)	15

5. Results and discussion

5.1. The results of stations clustering

Firstly, through the proposed Algorithm 1, 19 stations are clustered into 6 clusters, with cluster 1: stations 1 and 2, cluster 2: stations 3 and 4, cluster 3: stations 5, 6 and 7, cluster 4: stations 8, 9 and 10, cluster 5: stations 11, 12, 13, 14, 15 and 16, and cluster 6: stations 17, 18 and 19. Fig. 10 demonstrates the daily water levels of 6 clusters of 19 stations. As shown in Fig.

10, based on the 6-cluster result, the daily water level sequences within the same cluster show very similar characteristics in the data distribution and change trend.



(a) Clusters 1, 4, 5 and 6 (b) Clusters 2 and 3

Fig. 10. Visualization of 6 clusters result of 19 stations.

5.2. The results of daily water level prediction in 19 stations

Table 5 records the daily water level prediction performance of different models for the test set, in which, the parameters of each model are set according to Section 4.3. From the values of performance measures in Table 5, the following conclusions can be drawn:

(1) For each lead time, the models of MSDWLP_CS, MSDWLP_CH1 and MSDWLP_CH2, which were constructed based on the clustering results of 6 clusters, exhibited more accurate predictions in terms of the *MAE*, *MAPE*, *RMSE* and *NSE* than MSDWLP_S. Therefore, the multi-station clustering was an important factor that resulted in a more accurate prediction for our case study.

(2) For each lead time, the combined models, MSDWLP_CH1 and MSDWLP_CH2, both provided low errors (*MAE*, *MAPE* and *RMSE*) and high *NSE*. It can be seen that the combined models present improvements in the daily water level prediction accuracy.

(3) For each lead time, the MSDWLP_CH1 model, which is proposed in this paper, further improved the performance of daily water level prediction for multi-stations. For example, the *MAE* and *RMSE* on the 1st day were only 0.06m and 0.07m, and the *NSE* reached 0.999; on the 7th day, they were 0.41m, 0.45m, and 0.946, respectively.

The results in Table 5 present support to the proposed two-stage divide-and-conquer method for multi-station daily water level analysis and prediction. To further verify this claim, we select a special example, the first 7 days of testing data, which are from January 1, 2020 to January 7, 2020, and summarize the detailed prediction results of the stations in each cluster as follows.

Table 5. Test set performance for different MSDWLP models.

Model	MAE (m)	MAPE (%)	RMSE (m)	NSE
1 Day Lead Time				
MSDWLP_S	0.36	15.14	0.41	0.691
MSDWLP_CS	0.23	9.42	0.25	0.785
MSDWLP_CH1	0.06	2.61	0.07	0.999
MSDWLP_CH2	0.14	5.89	0.16	0.946
3 Day Lead Time				
MSDWLP_S	0.62	26.14	0.71	0.602
MSDWLP_CS	0.48	21.24	0.62	0.716
MSDWLP_CH1	0.19	7.71	0.21	0.972
MSDWLP_CH2	0.30	10.23	0.38	0.915
7 Day Lead Time				
MSDWLP_S	1.21	58.48	1.46	0.561
MSDWLP_CS	1.13	51.51	1.32	0.643
MSDWLP_CH1	0.41	11.78	0.45	0.946
MSDWLP_CH2	0.56	23.75	0.67	0.836

5.2.1 Cluster 1: a LSTM-SARIMA model

Cluster 1 contains stations 1 and 2. It can be seen from Fig. 4 that the water level data of station 1 and station 2 are not changing smoothly, but the daily water level appears to have a 7 days period. Firstly, using the addition model, the daily water level (training set) of station 1 is decomposed into trend and residual. Where, the residual term contains the period of small amplitude and residual. This was achieved by implementing the method of day period decomposition. From the trend part, we can find that it can reflects the changing trend of the daily water level, and is much smoother than the original water level data.

Then, the LSTM neural network was tailored and implemented to build a model for the trend prediction, and the SARIMA method was used for predictive modelling for the residual part. According to the experiments in Section 4.3, the following parameters settings were used: for the LSTM network, the number of neurons was set to 150, the batch size was set to 36, the time step was set to be 6, the number of epochs was set to 1000; the activation function was the rectified linear unit function "relu", the loss function was set to mse, and the optimiser function was the Nesterov-accelerated adaptive moment estimation optimiser "nadam"; for the SARIMA model, $s = 7$ and $p = 2$, $P = 3$, $d = 1$, $D = 1$, $q = 1$ and $Q = 0$.

Finally, the trend predicted by LSTM, and the residual predicted by SARIMA were added to obtain the predicted daily water level. As an example, the water level prediction results of stations 1 and 2 for certain 7 days (January 1, 2020 to January 7, 2020) are shown in Fig. 11 (a). It can be seen that the prediction accuracy is very good for the future 7 days, where the absolute error for the future 1st day is less than 0.07m and the one for the future 2nd day is less than 0.13m in station 1. Moreover, the absolute error are only 0.03m and 0.04m in station 2.

5.2.2 Cluster 2: a LSTM-SARIMA model

Cluster 2 includes stations 3 and 4, which are located in the Three Gorges reservoir area. As with cluster 1, the water level after decomposition also has an obvious period 8 days. In the same way, the LSTM network was used for forecasting the trend, and SARIMA was used for forecasting the residual. All the parameters of the LSTM network are the same as those in cluster 1. Except for $s = 8$, the parameters of SARIMA are also consistent with those in cluster 1. The predicted results of the daily water level of the future 7 days (January 1, 2020 to January 7, 2020) for stations 3 and 4 are shown in Fig. 11 (b). It can be seen that the prediction is very accurate for the first two days, where the absolute error for the future 1st day is less than 0.14m and the absolute error for the future 2nd day is less than 0.13m.

5.2.3 Cluster 3: a LSTM model

Cluster 3 contains three stations: 5, 6 and 7, which are also located in the Three Gorges reservoir area. Unlike clusters 1 and 2, after the water level was decomposed, no short period was found. However, good water level predictions can still be obtained using only the LSTM network. The predicted results of certain future 7 days (January 1, 2020 to January 7, 2020) for station 4 are as shown in Fig. 11 (c). The LSTM network settings were as follows: the batch size was set to 36, the number of neurons was set to 150, the epochs was set to 1000, the mean square error function was selected as loss function, the rectified linear unit function was selected as the activation function, the Nesterov-accelerated adaptive moment estimation was selected as the training function, and the step-times was set to 5. It is worth noting that the parameters are consistent with the LSTM network for clusters 1 and 2, except for the time step being 5. Although the absolute errors of stations 5, 6 and 7 are larger than those of stations 1, 2, 3 and 4, their absolute percentage errors are very small, because the daily water level of the three stations has always been above 145m. It can be found the prediction accuracy of the developed model is very good in forecasting the future several days, especially for stations 6 and 7.

5.2.4 Cluster 4: a LSTM model

Cluster 4 has three stations: no. 8, no. 9 and no. 10. Similar to cluster 3, the decomposed water levels have no short period. When the LSTM network is used for the daily water level prediction of cluster 4 of stations, the prediction results for the next 7 days (January 1, 2020 to January 7, 2020) are shown in Fig. 11 (d). In the experiments, all the parameters of the LSTM network are the same with those in the previous clusters, except that the time step is 10. It can be observed the developed model can well predict the water level for future several days, especially for the future 1st and 2nd days.

5.2.5 Cluster 5: a LSTM model

Cluster 5 includes six stations: no. 11, 12, 13, 14, 15 and 16. Most of these six stations are located in the lower reach of the Yangtze River trunk line. Although the water level data of the six stations have no obvious periodic characteristics, they are all relatively smooth. The LSTM network was set to have the same parameters as those for previous clusters, except for the time step being 8. The prediction results of the six stations in the future 7 days (January 1, 2020 to January 7, 2020) are shown in Fig. 11 (e). From Fig. 11 (e), we can see that the absolute errors of the six stations in the future 1st and 2nd days are all less than 0.09m. More importantly, the absolute errors of each station in the next 3 days are no more than 0.18m, and in the next 7 days are no more than 0.59m.

5.2.6 Cluster 6: a LSTM model

Cluster 6 contains stations 17, 18 and 19. After the decomposition of the water level data, an obvious period of 15 days was found. Similarly, the LSTM network for forecasting the trend, it was set to the same parameters as those for cluster 4. And the SARIMA was used for forecasting the residual term including period and residual , it settings were found as follows: $p = 2$, $P = 2$, $d = 0$, $D = 1$, $q = 1$, $Q = 1$, and $s = 15$. The absolute errors of the six stations do not exceed 0.03m in the future 1st day and 0.06m in the future 2nd day. The prediction results of the three stations of cluster 6 for the future 7 days (January 1, 2020 to January 7, 2020) are shown in Fig. 11 (f).

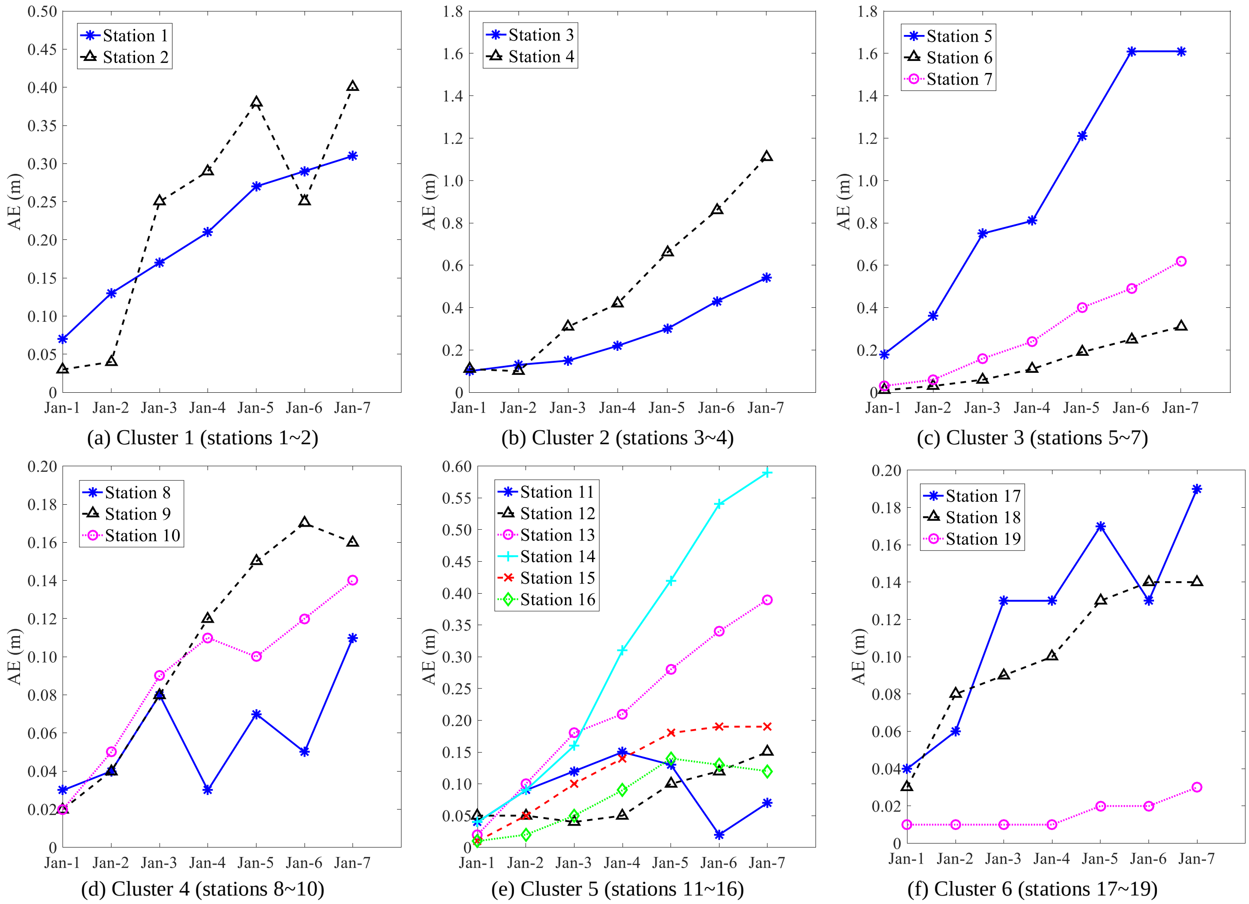


Fig. 11. The *AE* of water level of the future 7 days (January 1, 2020 to January 7, 2020) for 6 clusters.

Table 6. Configuration of MSDWLP_CH1.

Cluster	Stations	Shorter period	Prediction model
1	1, 2	7 days	LSTM [6, 36, 150, 1000]-SARIMA(2,1,1)(3,1,0)$_7$
2	3, 4	8 days	LSTM [6, 36, 150, 1000]-SARIMA(2,1,1)(3,1,0)$_8$
3	5, 6, 7	No	LSTM [5, 36, 150, 1000]
4	8, 9, 10	No	LSTM [10, 36, 150, 1000]
5	11, 12, 13, 14, 15, 16	No	LSTM [8, 36, 150, 1000]
6	17, 18, 19	15 days	LSTM [10, 36, 150, 1000] -SARIMA(2,0,1)(2,1,1)$_{15}$

In summary, in our MSDWLP_CH1, clusters 1, 2 and 6 adopt the combined model LSTM-SARIMA, while clusters 3, 4 and 5 used the LSTM model, where all the models have shown good prediction results. The water level prediction models and their main structural parameters of all clusters are shown in Table 6.

As a specific example, January 1, 2020 to January 7, 2020, the *MAE* and *MAPE* of 19 stations for a certain prediction case, are also tested and analyzed. In general, as the number of days increases, the prediction errors *MAE* of each station also increase. However, in the stations of each cluster, good water level prediction results have been achieved. In addition, to verify advantages of the proposed method, it is compared with two models MSDWLP_S and MSDWLP_CS, and the performances (mean of 10 experiments) of different models can be shown in Table 7. Compared with the proposed method, both MSDWLP_S and MSDWLP_CS have relatively high prediction errors. When MSDWLP_S is used to predict the case, the *MAE* and *MAPE* have further increased in all stations. When MSDWLP_CS is employed, the *MAE* and *MAPE* have greatly increased in stations 1~4 and 17~19. Obviously, the method MSDWLP_CH1 proposed in this paper has better accuracy in 19 stations daily water level prediction. More importantly, fewer model parameters are required for multi-station daily water level prediction in our method in that the clustering in the first stage reduces the number of objects that modelling in the second stage.

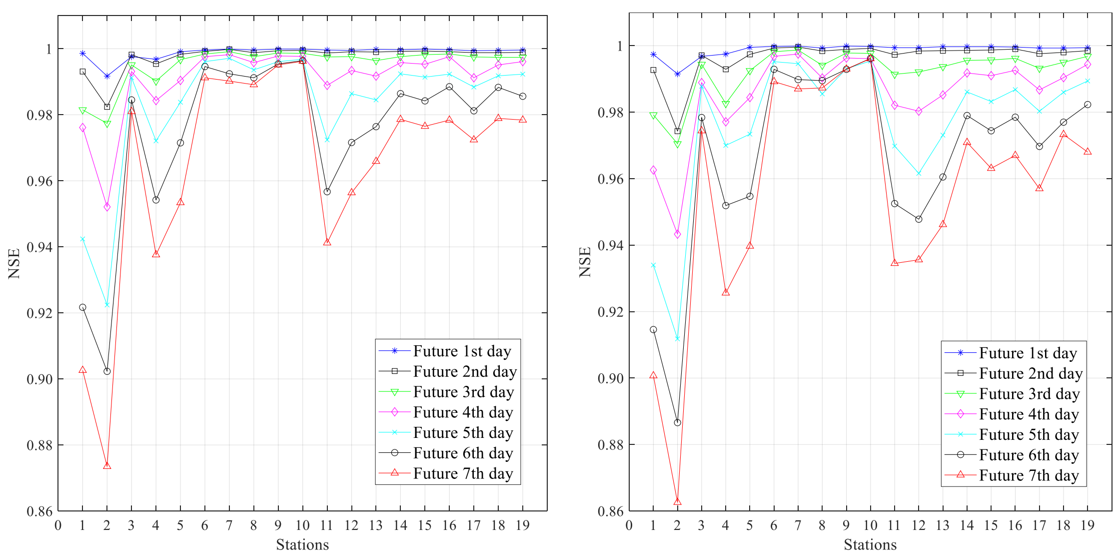
Table 7. *MAE* and *MAPE* of future 7 days (January 1, 2020 to January 7, 2020) with different MSDWLP models.

Station	*MAE* (m)			*MAPE* (%)		
	_S	_CS	_CH1	_C	_CS	_CH1
1	0.44	0.37	0.21	48.13	40.19	22.27
2	0.48	0.38	0.23	54.12	42.79	26.80
3	2.46	1.98	0.27	1.62	1.14	0.15
4	2.42	1.83	0.51	1.41	1.05	0.29
5	1.54	0.93	0.93	0.89	0.54	0.54
6	1.05	0.14	0.14	0.58	0.08	0.08
7	1.12	0.29	0.29	0.63	0.16	0.16
8	0.13	0.06	0.06	18.22	7.79	7.79
9	0.11	0.08	0.08	98.68	49.98	49.98
10	0.14	0.09	0.09	9.84	5.63	5.63

11	0.16	0.09	0.09	7.89	3.91	3.91
12	0.15	0.08	0.08	5.43	2.67	2.67
13	0.26	0.22	0.22	18.31	11.96	11.96
14	0.38	0.31	0.31	25.06	20.37	20.37
15	0.19	0.12	0.12	13.14	8.55	8.55
16	0.12	0.08	0.08	8.74	5.15	5.15
17	0.23	0.18	0.12	19.02	14.58	9.74
18	0.19	0.17	0.10	16.17	14.30	8.19
19	0.11	0.07	0.02	12.82	7.81	1.66

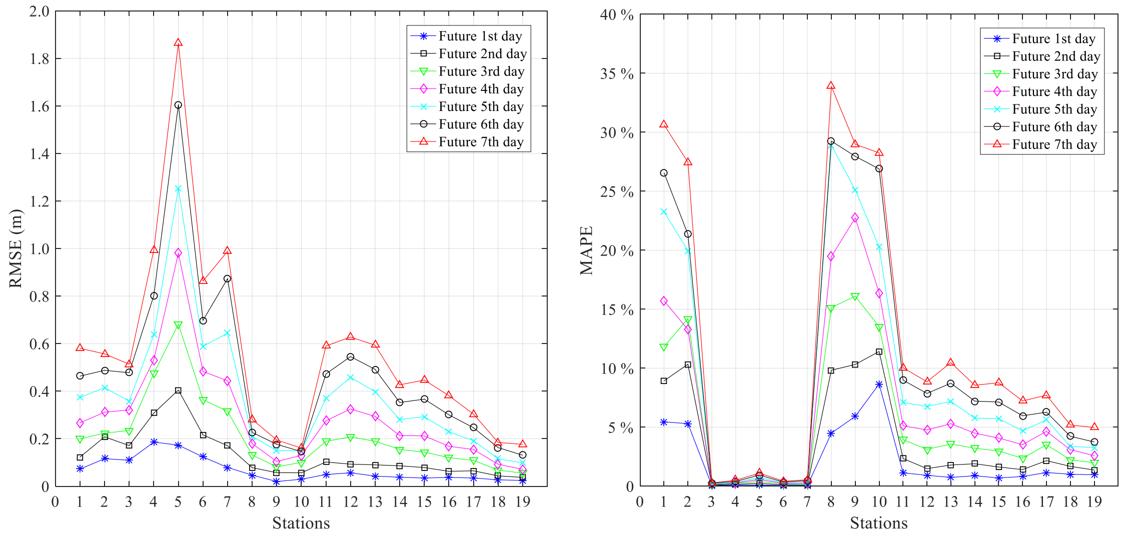
5.3. Discussion and insights

The predictive performance *NSE* of the developed MSDWLP_CH1 against all the training data (ranging from January 1, 2018 to December 31, 2019, 730 days) and testing data (ranging from January 1, 2020 to June 30, 2020, 182 days) is shown in Fig. 12. Among them, the testing data containing 182 days of daily water level is divided into 26 groups for prediction and analysis, and the output of each group includes 7 days. As can be seen from Fig. 12, the *NSE* of all training and testing are over 0.97 in the future 3 days. Moreover, the *NSE* of all stations in the future 7 days are basically above 0.90, except for 0.887 and 0.863 in the future 6th and 7th days at station 2, and most of them exceed 0.95.



(a) Training (b) Testing

Fig. 12. Modelling performance in *NSE* against all training and testing data.



(a) *RMSE* (b) *MAPE*

Fig. 13. The *RMSE* and *MAPE* of the developed models against all testing data.

Fig. 13 illustrates the prediction *RMSE* and *MAPE* of 19 stations in all testing data. As can be seen from Fig. 13 (a), at all stations, the *RMSE* becomes greater if the prediction time is longer. This means that the models always perform the best in predicting the near future, such as the future 1^{st} day, and become less accurate when predicting the far future, such as the future 7^{th} day. The bigger *RMSEs* appear in stations 4, 5, 6 and 7, where the stations are located in the Three Gorges Reservoir area and their daily water levels are above 145 meters all year round. The smaller *RMSEs* appear in stations 8, 9 and 10, which belong to the fourth cluster. Except for the stations in cluster 2 and cluster 3, the *RMSEs* of other stations for the future 1st day are less than 0.12m, and the *RMSEs* for the future 3rd day are less than 0.20m. From Fig. 13 (b), it can be seen the smallest *MAPE* appears in stations 3, 4, 5, 6 and 7, and the large *MAPEs* appear in stations 1, 2, 8, 9 and 10. Stations 1 and 2 belong to the first cluster and are located in the upper reach of the Yangtze River. Stations 8, 9 and 10 belong to the fourth cluster and are located in the middle reach. In fact, these two river sections are the most curved sections and the water regime is the most complicated. The variation of *MAPE* in other stations is relatively small, especially in the lower reach of the Yangtze River, such as stations 14~19.

In general, as the number of days increases, the prediction errors of each station also increase. The method proposed in this paper showed decent accuracy in 19 stations daily water level prediction, and more importantly, fewer model parameters are required in that the clustering in the first stage reduces the number of modelling objects in the second stage.

5. Conclusions

In this paper, the daily water level data of multiple stations on the Yangtze River trunk line have been collected, analysed and predicted in order to better inform decision-making about safe and effective waterborne transportation, water management, and emergency response. A new two-stage divide-and-conquer modelling strategy has been proposed. In the first stage, the DTW algorithm is employed to calculate the similarity distance of water level series, and the hierarchical clustering algorithm is used to divide 19 stations into 6 clusters according to the similarity matrix. In the second stage, the LSTM network and SARIMA model are tailored and implemented to build the MSDWLP models for each cluster. In particular, for the clusters with short periodicity and obvious seasonal change trend, the daily water level is decomposed into long-term trend, periodic change trend, and residual . Then the MSDWLP_CH1 model is employed for better results, in which, the long-term trend part is approximated by the LSTM network, and the residual term part is approximated by the SARIMA model. In the validation experiments, the daily water levels of 19 stations in the future 7 days are predicted. The results show that the proposed analysis and modelling method can be well applied to the case of the Yangtze River trunk line. The *RMSE* of the prediction is no more than 0.12m for the future 1st day and is no more than 0.26m for the future 3rd day. The average *MAPE* across 19 stations is 2.03% for the future 1st day and is 6.91% for the future 7 days.

The water levels of inland rivers is affected by many complex factors, such as a.s.l., waterway topography, periodic characteristics, and flood control and drought resistance strategies, which make it difficult to elicit conventional predictive models. In the proposed two-stage method, firstly, 19 stations were clustered into 6 categories according to the similar characteristics of daily water level, which reduced the influence of a.s.l. and waterway topographic changes on daily water level change, thereby reducing the complexity of multiple stations' daily water level prediction. Secondly, a prediction model was constructed for each cluster stations according to the periodic characteristics of daily water level, which reduced the number of prediction models and the parameters that need to be determined, thereby improving the accuracy of daily water level prediction. Moreover, the anticipation of the water level variation will support decision making in planning and operation of waterborne transportation, water management, and emergency response.

Meanwhile, a potential improvement in the following study is to employ more factors in the proposed method, and which is flexible to use other data, such as precipitation and tide, as predictors if available. The inland reginal and oceanic climate variables play a critical role in providing valuable information for the multi-station river water level prediction. Additionally, the performance of the proposed approach at smaller timescales, for example, hourly forecasting, is worth exploring.

Declaration of Competing Interset

The authors declare that they have no know competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 51709219), the National Key Research and Development Program of China (Grant No. 2018YFC1407400), the Qingdao Research Institute of Wuhan University of Technology (Grant No. 2019A02), and the China Scholarship Council (CSC) (Grant No. 201906950086).

References

Adikari, K. E., Shrestha, S., Ratnayake, D. T., Budhathoki, A., Mohanasundaram, S., Dailey, M. N., 2021. Evaluation of artificial intelligence models for flood and drought forecasting in arid and tropical regions. Environmental Modelling & Software, 144, 105136.

Chen, Y., Gan, M., Pan, S., Pan, H., Zhu, X., Tao, Z., 2020. Application of auto-regressive (AR) analysis to improve short-term prediction of water levels in the Yangtze estuary. Journal of Hydrology, 590, 125386.

Coraddu, A., Oneto, L., Baldi, F., Anguita, D., 2017. Vessels fuel consumption forecast and trim optimisation: a data analytics perspective. Ocean Engineering, 130, 351-370.

Dürrenmatt, D. J., Giudice, D. D., Rieckermann, J., 2013. Dynamic time warping improves sewer flow monitoring. Water Research, 47(11), 3803-3816.

Ebtehaja, I., Bonakdaria, H., Gharabaghib, B., 2019. A reliable linear method for modeling lake level fluctuations. Journal of Hydrology, 570.

Ehteram, M., Ferdowsi, A., Faramarzpour, M., Al-Janabi, A. M. S., Al-Ansari, N., Bokde, N. D., Yaseen, Z. M., 2021. Hybridization of artificial intelligence models with nature inspired optimization algorithms for lake water level prediction and uncertainty analysis. Alexandria Engineering Journal, 60(2), 2193-2208.

Fang, Z., Wang, Y., Peng, L., Hong, H., 2020. Predicting flood susceptibility using LSTM neural networks. Journal of Hydrology, 125734.

Gabela, J, Sarmiento, L., 2020. The effects of the 2013 floods on germany's freight traffic. Transportation Research Part D Transport and Environment, 102274.

Gers, F.A., Schmidhuber, J., Cummins, F., 2000. Learning to forget: continual prediction with LSTM. Neural Comput. 12 (10), 2451–2471.

Hyndman, Rob J., Khandakar, Y., 2008. Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 027.

Jordan, M. I., Mitchell, T. M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.Kasiviswanathan, K. S., He, J., Sudheer, K. P., Tay, J. H., 2016. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. Journal of Hydrology, 536, 161-173.

Lemenkova, P., 2019. Testing linear regressions by StatsModel Library of Python for oceanological data interpretation. Aquatic Sciences and Engineering, 34(2), 51-60.

Li, X., Jiang, W., Duan, D., 2020a. Spatio-temporal analysis of irrigation water use coefficients in china. Journal of Environmental Management, 262, 110242.1-110242.8.

Li, Y., Shi, H., Liu, H., 2020b. A hybrid model for river water level forecasting: cases of Xiangjiang River and Yuanjiang River, China. Journal of Hydrology, 587, 124934, doi: 10.1016/j.jhydrol.2020.124934.

Liu, Z., Cheng, L., Lin, K., Cai, H., 2021. A hybrid bayesian vine model for water level prediction. Environmental Modelling & Software,142, 105075.

Moeeni, H., Bonakdari, H., 2017. Forecasting monthly inflow with extreme seasonal variation using the hybrid SARIMA-ANN model. Stochastic environmental research and risk assessment, 31(8), 1997-2010.

Moeeni, H., Bonakdari, H., Ebtehaj, I., 2017. Integrated SARIMA with Neuro-Fuzzy Systems and Neural Networks for Monthly Inflow Prediction.Water Resources Management, 31(7).

Notteboom, T., Yang, D., Xu, H., 2020. Container barge network development in inland rivers: a comparison between the Yangtze River and the Rhine River. Transportation Research Part A: Policy and Practice, 132, 587–605.

Paul, G. C., Senthilkumar, S., Pria, R., 2018. An efficient approach to forecast water levels owing to the interaction of tide and surge associated with a storm along the coast of bangladesh. Ocean Engineering, 148, 516–529.

Phan, T. T. H., Nguyen X. H., 2020. Combining statistical machine learning models with arima for water level forecasting: the case of the red river. Advances in Water Resources, 142, 103656.

Quilty, J., Adamowski, J., 2020. A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. Environmental Modelling & Software, 130, 104718.

Sahoo, B. B., Jha, R., Singh, A., Kumar, D., 2019a. Application of support vector regression for modeling low flow time series. Ksce Journal of Civil Engineering, 23, 923-934.

Sahoo, B. B., Jha, R., Singh, A., Kumar, D., 2019b. Long short-term memory (lstm) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophysica, 67, 1471-1481, doi: 10.1007/s11600-019-00330-1.

Sun Q., Wan J., Liu S., 2020. Estimation of Sea Level Variability in the China Sea and Its Vicinity Using the SARIMA and LSTM Models. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 3317-3326.

Tourian, M. J., Tarpanelli, A., Elmi, O., Qin, T., Brocca, L., Moramarco, T., Sneeuw, N., 2016. Spatiotemporal densification of river water level time series by multimission satellite altimetry. Water Resources Research, 1140-1159.

Velasco C., Lazakis I., 2020. Real-time data-driven missing data imputation for short-term sensor data of marine systems. a comparative study. Ocean Engineering, 218, 108261.

Wang, D., Wei, S., Luo, H., Yue, C., Grunder, O., 2017. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Science of The Total Environment, 580, 719-733.

Wang, K., Zhang, A., 2018. Climate change, natural disasters and adaptation investments: inter- and intra-port competition and cooperation. Transportation Research Part B: Methodological, 117, 158-189.

Wei, C. C., 2015. Comparing lazy and eager learning models for water level forecasting in river-reservoir basins of inundation regions. Environmental Modelling & Software, 63, 137-155.

Xu, G., Cheng, Y., Liu, F., Ping, P., Sun, J., 2019. A Water Level Prediction Model Based on ARIMA-RNN. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (Big Data Service). IEEE.

Xu, X., Zhang, X., Fang, H., Lai, R., Zhang, Y., Huang, L., Liu, X., 2017. A real-time probabilistic channel flood-forecasting model based on the Bayesian particle filter approach. Environmental Modelling & Software, 88, 151-167.

Yang, T. H., Wang, C. W., & Lin, S. J., 2020. ECOMSNet–An edge computing-based sensory network for real-time water level prediction and correction. Environmental Modelling & Software, 131, 104771.

Yaseen, Z. M., Naghshara, S., Salih, S. Q., Kim, S., Malik, A., Ghorbani, M. A., 2020. Lake water level modeling using newly developed hybrid data intelligence model. Theoretical and Applied Climatology, 141, 1285-1300.

Yu, Z., Bedig, A., Montalto, F., Quigley, M., 2018. Automated detection of unusual soil moisture probe response patterns with association rule learning. Environmental Modelling & Software, 105, 257-269.

Yuan, Z., Liu, J., Liu, Y., Zhang, Q., Liu, W., 2020. A multi-task analysis and modelling paradigm using LSTM for multi-source monitoring data of inland vessels. Ocean Engineering, 213, 107604.

Yuan, Z., Liu, J., Zhang, Q., Liu, Y., Yuan, Y., Li, Z., 2021a. Prediction and optimisation of fuel consumption for inland ships considering real-time status and environmental factors. Ocean Engineering, 221, 108530.

Yuan, Z., Liu, J., Zhang, Q., Liu, Y., Yuan, Y., Li, Z., 2021b. A Practical Estimation Method of Inland Ship Speed Under Complex and Changeful Navigation Environment. IEEE Access, 9, 15643-15658.

Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J., 2018. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. Journal of Hydrology, 561, 918–929.

Zhong, C., Guo, T., Jiang, Z., Liu, X., Chu, X., 2017. A hybrid model for water level forecasting: A case study of Wuhan station. 2017 4th International Conference on Transportation Information and Safety (ICTIS), Banff, AB, 2017, 247-251.

Zhou, S., Xu, Z., Liu, F., 2017. Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering, IEEE Transactions on Neural Networks and Learning Systems, 28(12), 3007-3017.

Zhou, T., Jiang, Z., Liu, X., Tan, K., 2020. Research on the long-term and short-term forecasts of navigable river's water-level fluctuation based on the adaptive multilayer perceptron. Journal of Hydrology, 591, 125285.

Zhu, S., Hrnjica, B. I., Ptak, M., Adam Choiński, Sivakumar, B., 2020. Forecasting of water level in multiple temperate lakes using machine learning models. Journal of Hydrology, 585, 124819.