Accepted Manuscript

Research papers

Water level management of lakes connected to regulated rivers: an integrated modeling and analytical methodology

Tengfei Hu, Jingqiao Mao, Shunqi Pan, Lingquan Dai, Peipei Zhang, Diandian Xu, Huichao Dai

PII: S0022-1694(18)30365-2
DOI: https://doi.org/10.1016/j.jhydrol.2018.05.038
Reference: HYDROL 22815

To appear in: *Journal of Hydrology*

Received Date: 11 December 2017
Revised Date: 11 April 2018
Accepted Date: 15 May 2018

Water level management of lakes connected to regulated rivers: an integrated modeling and analytical methodology

Tengfei Hu[a, b], Jingqiao Mao[a*], Shunqi Pan[c], Lingquan Dai[d], Peipei Zhang[a], Diandian Xu[a], Huichao Dai[a]

[a] *College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China*

[b] *Nanjing Hydraulic Research Institute, Nanjing 210029, China*

[c] *School of Engineering, Cardiff University, Cardiff CF24 3AA, United Kingdom*

[d] *College of Hydraulic and Environmental Engineering, Three Gorges University, Yichang 443002, China*

*Corresponding author

Email: maojq@hhu.edu.cn (Jingqiao Mao)

Abstract: Reservoir operations significantly alter the hydrological regime of the downstream river and river-connected lake, which has far-reaching impacts on the lake ecosystem. To facilitate the management of lakes connected to regulated rivers, the following information must be provided: (1) the response of lake water levels to reservoir operation schedules in the near future and (2) the importance of different rivers in terms of affecting the water levels in different lake regions of interest. We develop an integrated modeling and analytical methodology for the water level management of such lakes. The data-driven method is used to model the lake level as it has the potential of producing quick and accurate predictions. A new genetic algorithm-based synchronized search is proposed to optimize input variable time lags and data-driven model parameters simultaneously. The methodology also involves the orthogonal design and range analysis for extracting the influence of an individual river from that of all the rivers. The integrated methodology is applied to the second largest freshwater lake in China, the Dongting Lake. The results show that: (1) the antecedent lake levels are of crucial importance for the current lake level prediction; (2) the selected river discharge time lags reflect the spatial heterogeneity of the rivers' impacts on lake level changes; (3) the predicted lake levels are in very good agreement with the observed data (RMSE ≤ 0.091 m; $R^2 \geq 0.9986$). This study demonstrates the practical potential of the integrated methodology, which can provide both the lake level responses to future dam releases and the relative contributions of different rivers to lake level changes.

Keywords: The Dongting Lake; Water level; Support vector regression; Input variable selection; Genetic algorithm; Orthogonal design

1 Introduction

Most of the major rivers in the world have been substantially changed by dam construction. The far-reaching impacts of river damming on the environment and ecosystems make it controversial. Dams can affect areas upstream and downstream of the rivers, on inundation, flow regulation, habitat fragmentation, etc. (Nilsson and Berggren, 2000; Nilsson et al., 2005). In particular, they significantly alter hydrological regimes of downstream rivers and river-connected lakes, for example, in terms of water level fluctuations, flood timing and duration.

Water level fluctuations play an important role in maintaining the structure, functioning and integrity of lake ecosystems (Coops et al., 2003; Leira and Cantonati, 2008). In dam-regulated rivers, relatively small water level fluctuations are often observed in the downstream areas (Magilligan and Nislow, 2005), which could negatively affect the ecosystems of river-connected lakes in various ways. For instance, the decreased amplitude of lake level fluctuations can lead to reductions in species richness and structural diversity of aquatic macrophytes (Geest et al., 2005; Wilcox and Meeker, 1991). Lake level stabilization can also dramatically change the spatial distribution and species composition of wetland vegetation (e.g., the succession of herbaceous to wooden wetlands). Due to lacustrine habitat deterioration, wetland habitat contraction and loss of wet-dry cycles, species abundance and richness of a variety of invertebrates, fishes, birds and mammals would as well diminish (Bunn and Arthington, 2002; Kingsford, 2000; Leira and Cantonati, 2008; Wilcox and Meeker, 1992). Moreover, water level drawdown in the downstream reaches arising from reservoir impoundment would accelerate the drainage of river-connected lakes, leading to earlier flood recession, extended duration of lake bottom exposure and potential wetland degradation (Wang et al., 2013; Zhang et al., 2012).

The Dongting Lake in China can be used as an example to illustrate the impacts of upstream dam regulation on the river-connected lake. The Three Gorges Dam (TGD) on the upper Yangtze River is one of the largest water resources projects in China and over the world (Yang et al., 2011). Since its first impoundment in June 2003, the TGD has been believed to be the main cause of many significant hydrological and ecological alterations, such as algal blooms in the reservoir tributary embayments (Mao et al., 2015) and changes in ecohydrological characteristics of mid-lower Yangtze reaches, river-connected lakes and Yangtze estuary (Chai et al., 2009; Dai et al., 2008; Gong et al., 2006; Yang et al., 2006; Zhang et al., 2012). The Dongting Lake, the second largest freshwater lake in China, is a Yangtze River-connected lake located downstream of the TGD. The lake and its surrounding wetlands are recognized as internationally important Ramsar sites, providing habitat for approximately 1,428 plant species, 114 fish species and 217 bird species (Xie et al., 2015). In general, there exists strong hydraulic interaction between the Yangtze River and the Dongting Lake. The hydrogeomorphic and ecological responses of the Dongting Lake to TGD operations have been well documented (Guan et al., 2014; Hu et al., 2015a; Hu et al., 2015b; Wu et al., 2013; Yuan et al., 2015). It is worth mentioning that the TGD, in addition to climate change and lakeshore development activities, accounts for hydrological regime alterations and some extreme drought events in this area (Dai et al., 2008). Such alterations may further result in severe environmental degradation, reduced biodiversity and water crises in the Dongting Lake region (Fang et al., 2006), indicating that the optimization of TGD operations is clearly necessary (Mao et al., 2016).

For the proper management of lakes connected to multiple dam-regulated rivers, the following questions must be answered: (1) how does the lake level respond to the scheduled dam releases in the near future? (2) which river plays the most important role in affecting the water levels in different lake regions of interest?

To deal with the first question calls for a modeling approach that relates remote river discharges to lake levels. In general, both physically based (e.g., hydrodynamic model) and data-driven (e.g., support vector regression, SVR) models can be used. The former is based on physical process descriptions with some simplifying assumptions (Abebe and Price, 2004), meaning that detailed topographical data are generally required. By contrast, the latter learns the input-output mapping from the training samples (Maier et al., 2010); therefore, only time series of the variable being investigated and its contributing factors are needed. Data-driven models clearly outperform their physically based counterparts in terms of computational efficiency (Lin et al., 2008). These models can thus be integrated into reservoir optimization models that minimize the negative impacts of reservoir operations on lake ecosystems. To model a lake using the data-driven method, the input variables are difficult to determine given that the response time of the lake level to different rivers can differ. Numerous combinations of time lagged river discharges that are potentially feasible need to be considered. In addition, it is necessary to calibrate the model during the evaluation of each candidate combination in order to avoid masking the candidate's real skill.

To answer the second question, one has to identify the relative contributions of different rivers to lake level variations. Understanding the different rivers' contributions is useful for carrying out cost-effective reservoir operations to satisfy the water demand of the lake in key periods (e.g., during reservoir impoundment). However, the water level at a lake site is dependent on the discharges of many different rivers. Some specific analytical techniques are needed to extract the influence of a single factor (i.e., an individual river) from that of a set of factors (i.e., all the rivers).

This paper aims to develop an integrated modeling and analytical methodology for the water level management of lakes connected to regulated rivers. First, site-specific prediction models of lake levels are developed using the data-driven method, which considers the impacts of remote river discharges and antecedent lake levels. In the model development, a new search strategy is proposed to obtain the optimal input variable time lags and data-driven model parameters simultaneously. The developed models can provide the lake level responses to future reservoir operation schedules. Second, based on the lake level models, the orthogonal design and range analysis are used to identify the importance of different rivers in terms of affecting the lake level.

The integrated modeling and analytical methodology is applied to the Dongting Lake in China. The reasonability of the selected input variable time lags is verified, and the performance of the lake level models (based on SVR) is fully assessed. The developed models are then used in a scenario where upstream dam releases in the following 10 days are scheduled. Next, the relative contributions of the Yangtze River and the Dongting Lake's major tributaries to lake level changes are analyzed. Given that rainfall is intentionally not considered in the lake level modeling, we also discuss the consequences of ignoring rainfall.

2 Material and methods

2.1 Study area and data collection

The Dongting Lake, located in the Yangtze River Basin, China (Fig. 1a), provides a wide range of ecosystem services, including drinking water supply, irrigation, fisheries and biodiversity conservation. The lake is one of the two large lakes that are directly connected to the Yangtze River (the other is Poyang Lake). Due to extensive reclamation and siltation, the area of the Dongting Lake had decreased from 4,350 km^2 in 1949 to 2,623 km^2 in 1995 (a 39.7% reduction) (Yin et al., 2007). Since the impoundment of the TGD in 2003, sediment interception by the reservoir has, to a large extent, prevented further reduction in the lake area (Hu et al., 2015a). The Dongting Lake Basin lies in a subtropical monsoon climate zone with an annual average temperature of ~18.6°C and an annual precipitation of 1,200-1,400 mm. The lake has distinct wet and dry seasons. The lake level in the dry season is much lower than that in the wet season, with a difference of over 10 m at Chenglingji.

The Dongting Lake is connected to the middle Yangtze River at the lake's northeastern end (i.e., Chenglingji, Fig. 1c). The connection is also made through some anastomosing distributary channels at three main avulsion nodes (i.e., Songzi, Taiping and Ouchi). In general, when the water level of the Yangtze River is lower, water flows from the lake into the river, and the lake level tends to decrease (i.e., emptying effect). By contrast, mainly during the wet season (April to October), the high water level in the Yangtze River limits the drainage of the lake (i.e., blocking effect). The Dongting Lake has four major tributaries, namely, the Xiang River, Zi River, Yuan River and Li River. The average annual water flowing into the Dongting Lake is 3.13×10^{11} m^3, of which the water from the Yangtze River accounts for 37.7% (Mao et al., 2016).

The hydrological data of the Dongting Lake and the related rivers from 2009 to 2012 were collected. Daily water levels of the lake were measured at lake stations No.1-5 (i.e., Chenglingji, Lujiao, Yingtian, Xiaohezui and Nanzui). Daily flow rates of the lake's four tributaries were obtained at river stations #1-4 (i.e., Xiangtan, Taojiang, Taoyuan and Shimen). The Qing River that joins the Yangtze River between the Gezhou Dam and Songzi node has small flow rates. In this study, the daily Yangtze River discharge (at #5) used in modeling consists of daily outflow discharges of the Gaobazhou Dam on the Qing River and the Gezhou Dam on the Yangtze River.

Both the Yangtze River and the Dongting Lake's tributaries are highly regulated by densely distributed dams (Fig. 1b). Thus, the Dongting Lake water level models to be developed are for the general and moderate flow and weather conditions, rather than extreme ones. It is decided that the data collected in 2010 and 2012 with slightly higher flood peaks are used for model training to obtain a wide validity, while the data in 2009 and 2011 are used for model testing. Table 1 presents the statistical characteristics of the hydrological data used in both periods. As can be observed in this table, the Yangtze River has significantly higher flow rates than the other rivers. The Xiang River and Yuan River contribute the most to the total tributary inflow to the Dongting Lake.

2.2 Integrated modeling and analytical methodology

This study develops an integrated modeling and analytical methodology to facilitate the water level management of lakes connected to multiple regulated rivers (e.g., the Dongting Lake).

As can be observed in Fig. 2, the data-driven method is used to model the lake water level, which considers the impacts of remote river discharges and antecedent lake levels. The data-driven method has an obvious advantage in providing quick predictions and is thus suitable to be integrated into a reservoir optimization model that attempts to improve the lake levels. Based on the genetic algorithm (GA), we propose a synchronized search for the optimal input variable time lags and data-driven model parameters. The synchronized optimization helps minimize the prediction error arising from model structural and parameter uncertainties. In the following step, site-specific prediction models of lake levels are trained using the optimized variable time lags and model parameters.

The developed models are used to provide the lake managers with the lake level responses to future reservoir operation schedules. Moreover, the relative contributions of different rivers are analyzed by using the orthogonal design.

2.3 Lake water level modeling

2.3.1 Problem formulation

Assuming that lake level variations are related to the discharges of rivers the lake is connected to, and the lake level at a time is also related to its states at the previous time steps, the daily water level at a lake station can be described as:

$$L_t = f(D^1_{t-m_1}, D^1_{t-m_1-1}, \ldots, D^1_{t-n_1}, \ldots, D^N_{t-m_N}, D^N_{t-m_N-1}, \ldots, D^N_{t-n_N}, L_{t-m_0}, L_{t-m_0-1}, \ldots, L_{t-n_0}) \quad (1)$$

where L_t is the water level at the lake station on day t; $D^i_{t\text{-}j}$ ($i = 1,\ldots,N$; $j = m_i,\ldots,n_i$) is the flow rate gauged at river station #i with a time lag of j days; L_{t-j} ($j = m_0,\ldots,n_0$) is the water level at the same lake station measured j days before day t. Notice that the minimum and maximum time lags, m_k and n_k ($k = 0,\ldots,N$), could vary across lake stations.

As can be observed in Eq. (1), rainfall is deliberately ignored in the lake level modeling, since the inclusion of rainfall could cause several problems. First, proper time lags of rainfall (at many rain gauges surrounding the lake) are very difficult to determine. Second, the computational time could significantly increase with the increase in the model input dimension. Moreover, future rainfall conditions have to be assumed before studying the lake level responses to reservoir operation schedules, which may introduce large prediction uncertainty.

2.3.2 Support vector regression

The regression function of Eq. (1) is estimated using SVR. SVR has been successfully applied to the modeling of environmental and water resources variables (Maier et al., 2010). For example, it has been used to predict lake water levels (e.g., Buyukyildiz et al., 2014; Çimen and Kisi, 2009), to predict river stages and discharges (e.g., Lin et al., 2006; Liong and Sivapragasam, 2002) and to estimate the relationship between river stage and discharge (e.g., Jain, 2012; Sivapragasam and Muttil, 2005).

Based on the statistical learning theory by Vapnik (1998), SVR is developed to solve non-linear regression estimation problems (Gunn, 1998). This technique employs structural risk minimization (SRM) rather than empirical risk minimization (ERM), which is often used in conventional artificial neural networks (ANNs). SRM attempts to minimize model complexity and empirical risk (i.e., training error) simultaneously and thus provides SVR with greater generalization capability (Kecman, 2001). Another advantage of SVR over conventional ANNs is that SVR has relatively few free parameters, leading to an easier calibration procedure (Khan and Coulibaly, 2006). However, the training process of SVR may be time-consuming when SVR is fed with large training dataset (Thissen et al., 2003).

Consider the dataset $\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_l,y_l)\}$, where $\boldsymbol{x}_i \in R^n$ is the input vector (e.g., remote river discharges and antecedent lake levels in this study) and $y_i \in R^1$ is the target output (e.g., current lake level). The underlying input-output relationship can be approximated by the non-linear function:

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^T \phi(\boldsymbol{x}) + b \tag{2}$$

where $\boldsymbol{\omega}$ is the weight vector, $\phi(\boldsymbol{x})$ is the embedding map that projects $\boldsymbol{x}$ into a high-dimensional feature space where linear regression can be performed, and b is the bias.

For the present application, the input vector must be mapped into the feature space due to the highly nonlinear relationship between the model inputs and output. The input-output relationship can be linearly estimated in a higher (possibly infinite) dimensional space.

Based on the linear ε-insensitive loss function ($\varepsilon > 0$ is the error threshold), the regression function is obtained by minimizing the regularized risk function (Vapnik, 1998):

$$\begin{aligned} \min_{\boldsymbol{\omega},b,\xi,\xi^*} \quad & \frac{\boldsymbol{\omega}^T\boldsymbol{\omega}}{2} + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \\ s.t. \quad & \boldsymbol{\omega}^T\phi(\boldsymbol{x}_i) + b - y_i \le \varepsilon + \xi_i \\ & y_i - \boldsymbol{\omega}^T\phi(\boldsymbol{x}_i) - b \le \varepsilon + \xi_i^* \qquad i = 1,\ldots.,l \\ & \xi_i, \xi_i^* \ge 0 \end{aligned} \tag{3}$$

where $C > 0$ is the regularization parameter determining the trade-off between model complexity $\boldsymbol{\omega}^T\boldsymbol{\omega}/2$ and training error $\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right)$, and the slack variables ξ_i and ξ_i^* are the lower and upper excess deviations, respectively.

Due to the possibly high dimensionality of $\boldsymbol{\omega}$, usually the dual problem of Eq. (3) is solved instead. The dual problem can be derived using the Lagrange multiplier technique:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \sum_{i=1}^{l} y_i\left(\alpha_i^* - \alpha_i\right) - \varepsilon\sum_{i=1}^{l}\left(\alpha_i^* + \alpha_i\right) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right)K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \\ s.t. \quad & \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right) = 0 \\ & 0 \le \alpha_i, \alpha_i^* \le C \qquad i = 1,....,l \end{aligned} \tag{4}$$

where α_i^* and α_i are Lagrange multipliers and $K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ is the kernel function.

The use of the kernel function avoids the 'curse of dimensionality'. There is no need to project the input vector into the high-dimensional feature space since the inner product in the feature space $\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ can be calculated directly from the training samples.

By solving Eq. (4), the regression function is

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right)K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) + b. \tag{5}$$

A symmetric, positive definite function that satisfies Mercer's theorem can be used as a kernel function (Gunn, 1998). Typical kernel functions include linear, polynomial, sigmoid and radial basis function (RBF). At present, there is no consensus as to which kernel is better than others (Buyukyildiz et al., 2014; Han et al., 2007). However, most SVR applications on hydrological modeling and forecasting have adopted the RBF kernel and obtained favorable performance (e.g., Çimen and Kisi, 2009; Khan and Coulibaly, 2006; Lin et al., 2006; Liong and Sivapragasam, 2002; Wei, 2015). In addition, the RBF has only one parameter to adjust, and it often shows better efficiency and performance than other kernels (Behzad et al., 2010; Dibike et al., 2001).

The RBF kernel is also used in this study, which takes the following form:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \exp\left(-\gamma \| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2\right) \tag{6}$$

where $\gamma > 0$ is the kernel parameter that determines the width of the kernel.

The error threshold ε, regularization parameter C and kernel parameter γ are user-defined SVR parameters. The LIBSVM software package (Chang and Lin, 2001) is used to solve the SVR function in this study.

2.4 Genetic algorithm-based synchronized search

Input variable selection (IVS) is a critical step in the development of data-driven models. Input variables should be relevant and non-redundant in order to avoid adding noise to the models and increasing model complexity. The omission of relevant input variables, on the other hand, can make the models inaccurate and unable to fully describe the system behavior (Galelli et al., 2014).

For the water level prediction of a lake connected to different rivers, proper time lags of discharges of these rivers (as well as the local water level) must be chosen. This can be a difficult task in view of the potentially long distance from the river stations to the lake stations. Taking the Dongting Lake as an example, there are five rivers surrounding the lake to be considered, and the longest distance from a river station to a lake station is ~390 km (Gezhou Dam to Chenglingji station). The complexity of IVS calls for a heuristic search algorithm (e.g., GA) to generate candidate input variable time lags, which can approach the optimal solution within a large search space (May et al., 2011).

Commonly used methods for evaluating candidate input variable time lags can be broadly grouped into filter, wrapper and embedded techniques (Guyon and Elisseeff, 2003). The filter techniques are independent of the designated data-driven method, and they assess the relevance of a model input based only on the available data (Liu and Motoda, 1998). This means that the response of model performance to the IVS outcome is completely ignored (Miller, 2002). In this regard, model-based wrapper and embedded algorithms can be more reliable. They evaluate the candidate time lags based on the corresponding model performance, as the data-driven method is integrated into the IVS procedure (Galelli et al., 2014). However, the model-based techniques generally need longer computational time and tend to mask the real skill of the candidates when the data-driven model is not calibrated for each of them (Maier et al., 2010). In addition, several relatively indirect ways to construct the model input can also be found in the literature (e.g., Baydaroğlu and Koçak, 2014; Baydaroğlu et al., 2017).

In a broad sense, data-driven model parameters can be viewed as a type of model input. In this spirit, a GA-based synchronized search for the optimal input variable time lags and data-driven model parameters (e.g., ε, C and γ of SVR in this study) is proposed (Fig. 3). The synchronized search falls into the category of model-based IVS algorithms due to the incorporation of the data-driven model into the search process. The search is implemented following the procedure below:

(1) Starting the search with the initial population (and the collected hydrological data); an individual in the population containing two floating numbers to indicate the time lags of each input variable and one floating number for each data-driven model parameter;

(2) Applying genetic operators (i.e., selection, crossover and mutation) to generate the

offspring population;

(3) Evaluating each individual in the offspring population by

   a) Dividing the individual into two parts, one for time lag information and the other for model parameters;

   b) Rounding up the two numbers for each input variable to obtain the variable's maximum and minimum time lags and then preparing the training dataset accordingly;

   c) Training the data-driven model with *n*-fold cross-validation to avoid overfitting;

   d) Using cross-validation root mean square error as the individual's fitness value;

(4) Checking whether the maximum generation has been reached;

(5) Ending the search and returning the optimal input variable time lags and model parameters if the answer is yes; otherwise, going back to Step (2).

2.5 Experimental setup

We applied the proposed modeling and analytical methodology to the Dongting Lake. As mentioned earlier, the model training period was 2010 and 2012 while the testing period was 2009 and 2011; the numbers of observations in the two periods were 731 and 720, respectively. It should be stressed that the synchronized optimization in Section 2.4 only used observations in the training period.

The synchronized optimization was separately implemented for each of the five lake stations (No.1-5) shown in Fig. 1c. Five river discharges gauged at river stations #1-5 were considered in the lake level modeling, i.e., N = 5 in Eq. (1). Three data-driven model parameters, i.e., ε, C and γ of SVR, were optimized along with the variable time lags. The 5-fold cross-validation was used to avoid the overfitting problem. In addition, all model inputs were linearly normalized to [0,1] to make sure they received equal attention in model training. The GA parameter values used in this study are shown in Table 2. As can be seen, 300 candidate combinations of variable time lags and model parameters evolved for 300 generations before returning the final optimization result. The GA search boundaries are also listed in Table 2.

The performance of the developed lake level models was assessed against multiple metrics, including root mean square error (RMSE) and coefficient of determination (R^2). The two metrics represent ‘squared errors’, which are apt to be dominated by large errors (Maier et al., 2010). Therefore, mean absolute error (MAE) and mean relative error (MRE) were also calculated to provide additional error information. The above four performance metrics are summarized in Table 3.

3 Results

3.1 Input variable time lags and SVR parameters

The synchronized search for the optimal input variable time lags and SVR parameters was separately applied to stations No.1-5 in the Dongting Lake (Fig. 1c). The optimization results are shown in Fig. 4 and Table 4, respectively.

According to Fig. 4, the strongest factor affecting the current lake level is the local lake levels at the previous time steps, ranging from three (at Lujiao, No.2) to eight days (at Nanzui, No.5). This is supported by the high correlation between the current lake level and its previous states (see Fig. 5). Fig. 5 also shows that the correlation exhibits a decreasing trend with time, which agrees with that the lake level on day *t* is most strongly affected by the lake level on day *t*-1 (Fig. 4).

Fig. 4 demonstrates that different rivers contribute differently to water level variations at a lake station. In addition, the time lags of a river are significantly different across the lake stations. These results reflect the spatial heterogeneity of the rivers' impacts on lake level changes.

The length of river discharge time lags ranges from the shortest one day (e.g., the Li River flow, D^4, to Yingtian, No.3) to the longest nine days (e.g., the Yangtze River flow, D^5, to Chenglingji, No.1). The time lag length is positively associated with the distance from the river station to the lake station and the amplitude of river discharge fluctuations. Therefore, the discharge of the Yangtze River, D^5, with the longest flow path and significant changes in flow magnitude, has the longest time lag length among the five river discharges.

As shown in Fig. 4, it takes approximately one to three days for the Xiang River flow, D^1, to reach lake stations Lujiao and Yingtian (No.2 and 3). The needed time to reach station Chenglingji (No.1) is often longer because of the increase in travel distance. A similar trend for the Zi River flow, D^2, can also be observed in this figure. It is interesting to note that the effects of the Xiang River and Zi River are identified by the GA in predicting the water levels at Xiaohezui and Nanzui (No.4 and 5), even though the confluence of each of the two rivers and the Dongting Lake lies downstream of the two stations. A possible explanation for this result is that the inflows from the two rivers can alter the downstream lake levels and, in turn, influence the upstream lake levels. Compared with the Xiang River flow, the water levels at stations No.4 and 5 are more responsive to the Zi River flow, which could be attributed to the relatively short distances from the Zi River's confluence to the two lake sites (Fig. 1c).

It is found that one day is generally insufficient for the Yuan River flow D^3 to reach Chenglingji (No.1), and the corresponding time lags are between two to nine days. For stations Yingtian, Xiaohezui and Nanzui (No.3-5), the time lags of D^3 lie between one day and six days. However, the water transport delay to reach station Lujiao (No.2) is shown to be much shorter (one or two days). One potential explanation for this significant difference is that Lujiao is located in a long and narrow channel (see Fig. 1c) with relatively high flow velocities; lake level changes at this site are sensitive to large inflows that require short travel time. Compared with station Nanzui (No.5), the water level at Xiaohezui (No.4) is relatively insensitive to the Li River flow D^4, which most likely results from the longer distance from the Li River's confluence to Xiaohezui (No.4). The Li River flow D^4 plays a limited role in affecting the water levels at stations Chenglingji, Lujiao and Yingtian (No.1-3) due to its small magnitude (see Table 1).

3.2 Lake level model performance

Fig. 6 compares the observed and predicted Dongting Lake water levels in the training and testing periods. A very good agreement between model predictions and observations can be found in both periods at each lake station. The lake level predictions are accurate even for the peak levels in the testing period. The consistent model performance arises from the fact that these lake level models are allowed to experience more severe floods in the training period. However, the predicted lake levels occasionally deviate from the observed data, and the model for station Yingtian (No.3) yields the largest proportion of these deviations. Fig. 7 presents the boxplots of lake level prediction errors (i.e., predictions minus observations) in the testing period. A majority of the errors (92.3%) vary between -0.10 m and 0.10 m. The models for stations Xiaohezui and Nanzui (No.4 and 5) produce the smallest errors, followed by those for Chenglingji and Lujiao (No.1 and 2).

The RMSE, R^2, MAE and MRE of the five lake level models are summarized in Table 5. These models can provide accurate predictions of daily Dongting Lake water level, with the maximum RMSE of 0.091 m and the minimum R^2 of 0.9986 in the testing period. The model for station Xiaohezui (No.4) has the best accuracy (RMSE = 0.037 m, MAE = 0.028 m and MRE = 0.0009), followed by, in sequence, the models for Nanzui (No.5), Chenglingji (No.1) and Lujiao (No.2). Although the model for Yingtian (No.3) presents relatively low performance, its prediction errors are still acceptable (RMSE = 0.091 m, MAE = 0.061 m and MRE = 0.0024). Such a model performance ranking is in common with the result obtained by merely considering the distribution of the prediction errors (Fig. 7). Interestingly, the five models have a very different order of performance when assessed against R^2, namely, the models for Chenglingji (No.1), Lujiao (No.2), Yingtian (No.3), Xiaohezui (No.4) and Nanzui (No.5) in descending order. Such discrepancies most likely result from the limited amplitude of water level variations at Xiaohezui and Nanzui (No.4 and 5), according to the definition of R^2. Table 5 also suggests that no manifest differences exist between training and testing RMSEs, meaning that the 5-fold cross-validation and SRM principle of SVR avoid overfitting effectively.

3.3 Lake level responses to future dam releases

The prediction of the Dongting Lake water level on day *t* relies on the availability of different river discharges and local water level on day *t*-1 (Fig. 4). The lake level on day *t*+1 can be predicted when the relevant measurements on day *t* are acquired on a real-time basis (i.e., real-time updating). However, to obtain the lake level responses to upstream reservoir operation schedules in the near future, the newly predicted lake levels need to be used instead of lake level observations as model inputs whenever possible (i.e., 'indirect' multi-step prediction).

The Dongting Lake water level variations over the course of a year could be characterized as four periods, namely the dry period (last Dec. to Mar.), water-level rise period (Apr. to May), wet period (Jun. to Jul.) and drawdown period (Aug. to Nov.). Taking station Chenglingji (No.1) as an example, we selected a 'time window' of 10 days for each period in 2009 to present the lake level responses. The observed flow rates of the Yangtze River and lake tributaries in each time window were considered the scheduled dam releases.

Fig. 8 compares the observed lake levels with the lake levels obtained from real-time updating and indirect multi-step prediction. The lake levels from real-time updating are closer to the observed data than the multi-step predictions. However, the accuracy of the multi-step prediction is still acceptable especially when the lake level remains low, rises or declines. Fig. 8 also reveals that, for the indirect multi-step prediction in each time window, the absolute prediction error does not necessarily enlarge with time.

3.4 Contributions of different rivers to lake level changes

The final step of the integrated methodology is to use the orthogonal design (Taguchi, 1987) and range analysis to identify the relative contributions of different rivers to lake level changes. The orthogonal design has been widely used in the field of design of experiments (e.g., Ghani et al., 2004; Kwak and Choi, 2002) due to its quick result and statistical rigor. This method can substantially reduce the number of needed experiments but still provide sufficient information. An orthogonal array of five factors at four levels ($L_{16}(4^5)$) was designed in Table 6. Each of the 16 model runs corresponded to a combination of river discharge variations. Based on the training data, a river discharge was altered by -15%, -5%, 5% and 15% under the levels of 1-4, respectively. Lake level variations were the differences in model-predicted lake levels corresponding to the changed and unchanged model inputs.

Fig. 9 shows the main effects of the five rivers obtained with the range analysis. Stations Chenglingji, Lujiao and Yingtian (No.1-3) see greater water level changes than the other two stations. This is in common with the characteristics of water level fluctuations at the five stations (Table 1). Fig. 9 also shows that the Yangtze River clearly plays a dominant role in affecting the lake levels at stations No.1-3. In addition, it seems reasonable that the lake levels at the three sites increase with the increase in lake tributary inflows. Note that the negligible effect of the Li River agrees with its small flow magnitude.

According to Fig. 9, lake level variations at stations Xiaohezui and Nanzui (No.4 and 5) are governed by both the Yangtze River and the Yuan River. Relative to stations No.1-3, the effect of the Yangtze River becomes less strong at the two sites. The reason is presumably that the lake levels at stations No.4 and 5 are about five meters higher than those at stations No.1-3 (Table 1). The Yuan River overtakes the other lake tributaries in terms of affecting the water levels at No.4 and 5, as its relatively large discharge flows past the two stations (Fig. 1c).

The above findings agree well with the relative magnitudes of correlation coefficients between the lake levels and river discharges (Table 7). For the lake levels at No.1-3, the correlation with the Yangtze River discharge (D^5) is obviously the greatest among the five river discharges. In the case of stations No.4 and 5, the correlation with the Yuan River discharge (D^3) turns noticeable.

4 Discussion

In the present work, the integrated modeling and analytical methodology was applied to the Dongting Lake in China, which is connected to multiple regulated rivers. In the development of the Dongting Lake water level models, we did not take into account the effect of rainfall. Even though the reasons for this have been given in Section 2.3.1, it is still interesting to investigate the consequences of ignoring rainfall in the lake level modeling.

Fig. 7 suggests that the developed lake level models produce large errors occasionally. Both the greatest overestimate (0.43 m) and the greatest underestimate (-0.64 m) occur at station Yingtian (No.3). The serious underestimates can probably be attributed to ignoring rainfall, which is not reflected by the river discharges, in the lake level modeling. Taking Yingtian (No.3) as an example, we further collected daily rainfall at weather stations P1 and P2 (Fig. 1c) to verify this assumption. Obviously, the runoff associated with rainfall at P1 and P2 has not been reflected by the river flow at station Xiangtan (#1). Fig. 10 shows the average daily rainfall at P1 and P2 and the serious underestimates of the observed lake levels (< -0.10

m) at Yingtian (No.3). It can be observed that a majority of the underestimates are closely related to the preceding rainfall, meaning that the model fails to capture the effect of rainfall over the areas downstream of the river stations.

However, the GA can find a trade-off solution when used to calibrate the lake level models without rainfall. Fig. 11 is the schematic diagram illustrating the trade-off solution: (1) initially in Phase 1, there is no rainfall in the river-lake system; a powerful model can 'perfectly' predict the lake levels using remote river discharges and antecedent lake levels; (2) in Phase 2, after a rainfall event is imposed, the model with its original parameter setting still can precisely predict the lake levels that are unaffected by the rainfall, but inevitably underestimates the raised water levels arising from runoff generation; (3) due to parameter optimization seeking to minimize the RMSE, the updated model in Phase 3 increases the predicted lake levels to reduce the underestimation. As stated earlier, the RMSE is dominated by large errors; the increase in model predictions thus caters for serious underestimates related to extreme rainfall, which eventually results in occasional large underestimates and overall slight overestimates. Fig. 12 shows the proportions of overestimated and underestimated lake levels. In accordance with the above speculation, the proportion of overestimates exceeds 50% at all lake stations. On average, 57.9% of the lake levels are overestimated and 42.1% are underestimated; the corresponding cumulative errors are 86.64 m and -61.22 m, respectively.

The performance-oriented parameter optimization ensures the high accuracy of the developed lake level models. Particularly for the lake level management, the simplified yet pragmatic models can better serve the purpose of providing the Dongting Lake water level responses to the upstream dam releases.

5 Conclusions

This study develops an integrated modeling and analytical methodology for the water level management of lakes connected to dam-regulated rivers, and applies the methodology to the Dongting Lake in China. The following conclusions can be drawn:

(1) The antecedent lake levels are the most important factor for the prediction of the current lake level;

(2) The river discharge time lags selected by the GA well describe the spatial heterogeneity of the rivers' impacts on lake level changes;

(3) The synchronized optimization is able to fulfill the potential of SVR, leading to highly accurate prediction of lake levels;

(4) The integrated methodology can provide the lake level responses to future dam releases and the relative importance of different rivers in terms of affecting the lake level.

Acknowledgments

The authors thank the anonymous reviewers and the editors for their constructive comments. This research is funded by the National Key Research and Development Program of China (2016YFC0402204), the National Natural Science Foundation of China (51379059) and the Fundamental Research Funds for the Central Universities. The support from the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) is also acknowledged.

References

Abebe, A.J., Price, R.K., 2004. Information theory and neural networks for managing uncertainty in flood routing. J. Comput. Civil Eng. 18 (4), 373-380.

Baydaroğlu, Ö., Koçak, K., 2014. SVR-based prediction of evaporation combined with chaotic approach. J. Hydrol. 508, 356-363.

Baydaroğlu, Ö., Koçak, K., Duran, K., 2017. River flow prediction using hybrid models of support vector regression with the wavelet transform, singular spectrum analysis and chaotic approach. Meteorol. Atmos. Phys. DOI 10.1007/s00703-017-0518-9.

Behzad, M., Asghari, K., Coppola, E.A., 2010. Comparative study of SVMs and ANNs in aquifer water level prediction. J. Comput. Civil Eng. 24 (5), 408-413.

Bunn, S.E., Arthington, A.H., 2002. Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity. Environ. Manage. 30 (4), 492-507.

Buyukyildiz, M., Tezel, G., Yilmaz, V., 2014. Estimation of the change in lake water level by artificial intelligence methods. Water Resour. Manag. 28 (13), 4747-4763.

Chai, C., Yu, Z.M., Shen, Z.L., Song, X.X., Cao, X.H., Yao, Y., 2009. Nutrient characteristics in the Yangtze River Estuary and the adjacent East China Sea before and after impoundment of the Three Gorges Dam. Sci. Total Environ. 407 (16), 4687-4695.

Chang, C.C., Lin, C.J., 2001. LIBSVM - A Library for Support Vector Machines (Version 3.21, December 2015). Software available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Çimen, M., Kisi, O., 2009. Comparison of two different data-driven techniques in modeling lake level fluctuations in Turkey. J. Hydrol. 378 (3-4), 253-262.

Coops, H., Beklioglu, M., Crisman, T.L., 2003. The role of water-level fluctuations in shallow lake ecosystems - workshop conclusions. Hydrobiologia 506 (1-3), 23-27.

Dai, Z.J., Du, J.Z., Li, J.F., Li, W.H., Chen, J.Y., 2008. Runoff characteristics of the Changjiang River during 2006: effect of extreme drought and the impounding of the Three Gorges Dam. Geophys. Res. Lett. 35 (7).

Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B., 2001. Model induction with support vector machines: introduction and applications. J. Comput. Civil Eng. 15 (3), 208-216.

Fang, J.Y., Wang, Z.H., Zhao, S.Q., Li, Y.K., Tang, Z.Y., Yu, D., Ni, L.Y., Liu, H.Z., Xie, P., Da, L.J., Li, Z.Q., Zheng, C.Y., 2006. Biodiversity changes in the lakes of the Central Yangtze. Front. Ecol. Environ. 4 (7), 369-377.

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Modell. Softw. 62, 33-51.

Geest, G.J.V., Coops, H., Roijackers, R.M.M., Buijse, A.D., Scheffer, M., 2005. Succession of aquatic vegetation driven by reduced water-level fluctuations in floodplain lakes. J. Appl. Ecol. 42 (2), 251-260.

Ghani, J.A., Choudhury, I.A., Hassan, H.H., 2004. Application of Taguchi method in the optimization of end milling parameters. J. Mater. Process. Tech. 145 (1), 84-92.

Gong, G.C., Chang, J., Chiang, K.P., Hsiung, T.M., Hung, C.C., Duan, S.W., Codispoti, L.A., 2006. Reduction of primary production and changing of nutrient ratio in the East China

Sea: effect of the Three Gorges Dam? Geophys. Res. Lett. 33 (7).

Guan, L., Wen, L., Feng, D., Zhang, H., Lei, G., 2014. Delayed flood recession in central Yangtze floodplains can cause significant food shortages for wintering geese: results of inundation experiment. Environ. Manage. 54 (6), 1331-1341.

Gunn, S.R., 1998. Support Vector Machines for Classification and Regression, Technical Report, University of Southampton, UK.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157-1182.

Han, D., Chan, L., Zhu, N., 2007. Flood forecasting using support vector machines. J. Hydroinform. 9 (4), 267-276.

Hu, C.H., Fang, C.M., Cao, W.H., 2015a. Shrinking of Dongting Lake and its weakening connection with the Yangtze River: analysis of the impact on flooding. Int. J. Sediment Res. 30 (3), 256-262.

Hu, Y.X., Huang, J.L., Du, Y., Han, P.P., Wang, J.L., Huang, W., 2015b. Monitoring wetland vegetation pattern response to water-level change resulting from the Three Gorges Project in the two largest freshwater lakes of China. Ecol. Eng. 74, 274-285.

Jain, S.K., 2012. Modeling river stage-discharge-sediment rating relation using support vector regression. Hydrol. Res. 43 (6), 851.

Kecman, V., 2001. Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. MIT Press, Cambridge, Massachusetts, USA.

Khan, M.S., Coulibaly, P., 2006. Application of support vector machine in lake water level prediction. J. Hydrol. Eng. 11 (3), 199-205.

Kingsford, R.T., 2000. Ecological impacts of dams, water diversions and river management on floodplain wetlands in Australia. Austral. Ecol. 25 (2), 109-127.

Kwak, N., Choi, C.H., 2002. Input feature selection for classification problems. IEEE T. Neural Networ. 13 (1), 143-159.

Leira, M., Cantonati, M., 2008. Effects of water-level fluctuations on lakes: an annotated bibliography. Hydrobiologia 613, 171-184.

Lin, B., Syed, M., Falconer, R.A., 2008. Predicting faecal indicator levels in estuarine receiving waters - an integrated hydrodynamic and ANN modelling approach. Environ. Modell. Softw. 23 (6), 729-740.

Lin, J.Y., Cheng, C.T., Chau, K.W., 2006. Using support vector machines for long-term discharge prediction. Hydrolog. Sci. J. 51 (4), 599-612.

Liong, S.Y., Sivapragasam, C., 2002. Flood stage forecasting with support vector machines. J. Am. Water Resour. As. 38 (1), 173-186.

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell, Massachusetts, USA.

Magilligan, F.J., Nislow, K.H., 2005. Changes in hydrologic regime by dams. Geomorphology 71 (1-2), 61-78.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ. Modell. Softw. 25 (8), 891-909.

Mao, J.Q., Jiang, D.G., Dai, H.C., 2015. Spatial–temporal hydrodynamic and algal bloom modelling analysis of a reservoir tributary embayment. J. Hydro-Environ. Res. 9 (2), 200-215.

Mao, J.Q., Zhang, P.P., Dai, L.Q., Dai, H.C., Hu, T.F., 2016. Optimal operation of a multi-reservoir system for environmental water demand of a river-connected lake. Hydrol. Res. 47 (S1), 206-224.

May, R., Dandy, G., Maier, H., 2011. Review of Input Variable Selection Methods for Artificial Neural Networks, in: Suzuki, K. (Eds.), Artificial Neural Networks - Methodological Advances and Biomedical Applications. InTech, pp. 19-44.

Miller, A.J., 2002. Subset Selection in Regression, second ed. Chapman and Hall, London, UK.

Nilsson, C., Berggren, K., 2000. Alterations of riparian ecosystems caused by river regulation. Bioscience 50 (9), 783-792.

Nilsson, C., Reidy, C.A., Dynesius, M., Revenga, C., 2005. Fragmentation and flow regulation of the world's large river systems. Science 308 (5720), 405-407.

Sivapragasam, C., Muttil, N., 2005. Discharge rating curve extension - a new approach. Water Resour. Manag. 19 (5), 505-520.

Taguchi, G., 1987. System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs. UNIPUB/Kraus International, New York, USA.

Thissen, U., Van Brakel, R., De Weijer, A.P., Melssen, W.J., Buydens, L.M.C., 2003. Using support vector machines for time series prediction. Chemometr. Intell. Lab. 69 (1-2), 35-49.

Vapnik, V., 1998. Statistical Learning Theory. John Wiley & Sons, New York, USA.

Wang, J.D., Sheng, Y.W., Gleason, C.J., Wada, Y., 2013. Downstream Yangtze River levels impacted by Three Gorges Dam. Environ. Res. Lett. 8 (4), 44012-44020.

Wei, C.C., 2015. Comparing lazy and eager learning models for water level forecasting in river-reservoir basins of inundation regions. Environ. Modell. Softw. 63, 137-155.

Wilcox, D.A., Meeker, J.E., 1991. Disturbance effects on aquatic vegetation in regulated and unregulated lakes in northern Minnesota. Can. J. Botany 69 (7), 1542-1551.

Wilcox, D.A., Meeker, J.E., 1992. Implications for faunal habitat related to altered macrophyte structure in regulated lakes in northern Minnesota. Wetlands 12 (3), 192-203.

Wu, H.P., Zeng, G.M., Liang, J., Zhang, J.C., Cai, Q., Huang, L., Li, X.D., Zhu, H.N., Hu, C.X., Shen, S., 2013. Changes of soil microbial biomass and bacterial community structure in Dongting Lake: impacts of 50,000 dams of Yangtze River. Ecol. Eng. 57, 72-78.

Xie, Y.H., Yue, T., Chen, X.S., Feng, L., Deng, Z.M., 2015. The impact of Three Gorges Dam on the downstream eco-hydrological environment and vegetation distribution of East Dongting Lake. Ecohydrology 8 (4), 738-746.

Yang, S.L., Milliman, J.D., Li, P., Xu, K., 2011. 50,000 dams later: erosion of the Yangtze River and its delta. Global Planet. Change, 75 (1-2): 14-20.

Yang, Z., Wang, H., Saito, Y., Milliman, J.D., Xu, K., Qiao, S., Shi, G., 2006. Dam impacts on the Changjiang (Yangtze) River sediment discharge to the sea: the past 55 years and after the Three Gorges Dam. Water Resour. Res. 42 (4).

Yin, H.F., Liu, G.R., Pi, J.G., Chen, G.J., Li, C.G., 2007. On the river-lake relationship of the middle Yangtze reaches. Geomorphology 85 (3-4), 197-207.

Yuan, Y.J., Zeng, G.M., Liang, J., Huang, L., Hua, S.S., Li, F., Zhu, Y., Wu, H.P., Liu, J.Y., He, X.X., He, Y., 2015. Variation of water level in Dongting Lake over a 50-year period: implications for the impacts of anthropogenic and climatic factors. J. Hydrol. 525, 450-456.

Zhang, Q., Li, L., Wang, Y.G., Werner, A.D., Xin, P., Jiang, T., Barry, D.A., 2012. Has the Three-Gorges Dam made the Poyang Lake wetlands wetter and drier? Geophys. Res. Lett. 39 (20).

List of Figures

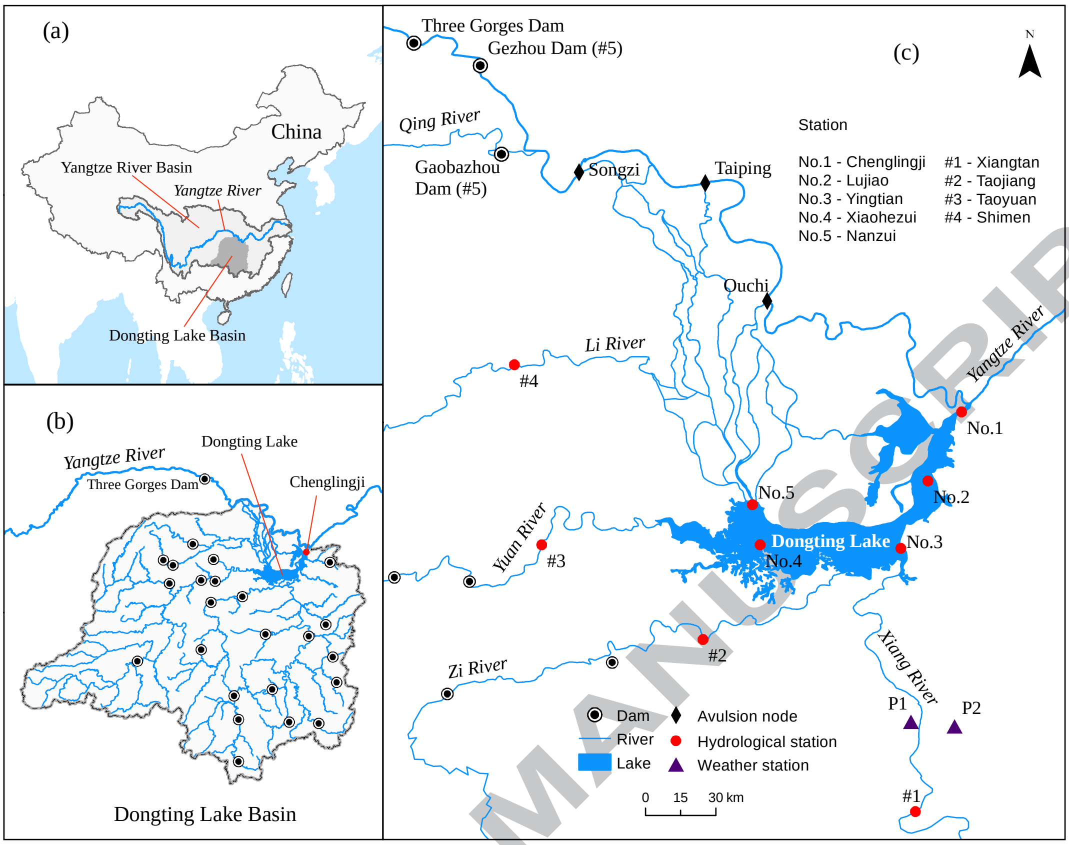

Fig. 1. Map of the study area. (a) Locations of the Yangtze River and the Dongting Lake Basin in the Yangtze River Basin, China; (b) river system and dam distribution in the Dongting Lake Basin; (c) the Yangtze River-Dongting Lake system, including distributary channels connecting the Dongting Lake to the Yangtze River at three main avulsion nodes (i.e., Songzi, Taiping and Ouchi) and the lake's four major tributaries (i.e., the Xiang River, Zi River, Yuan River and Li River).

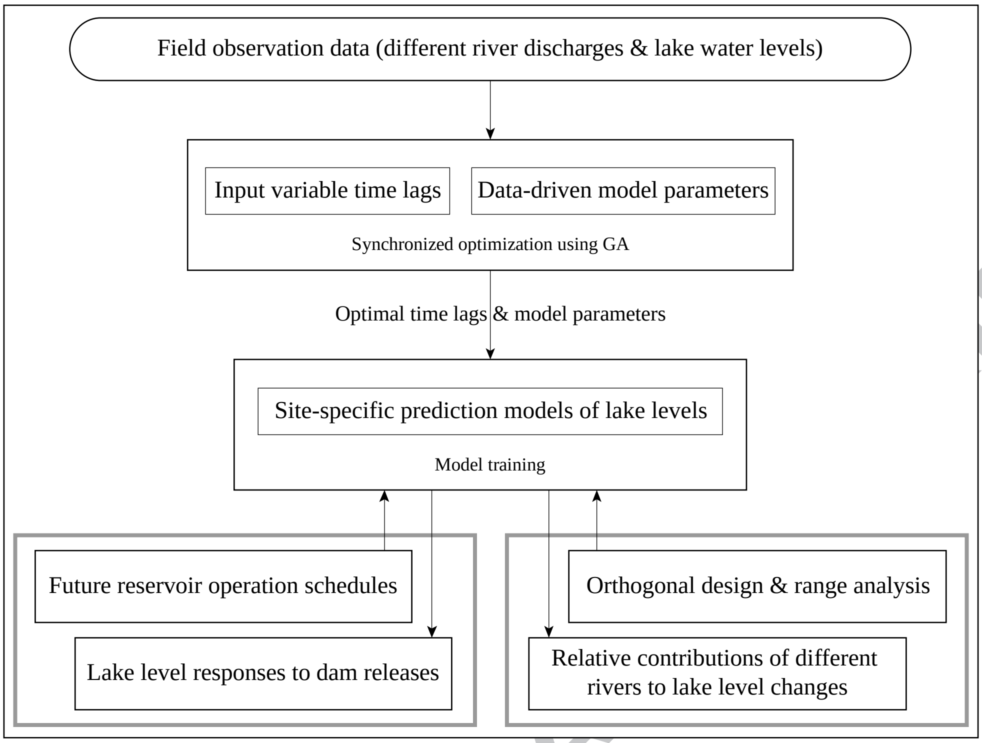

Fig. 2. Integrated modeling and analytical methodology for water level management of lakes connected to regulated rivers.

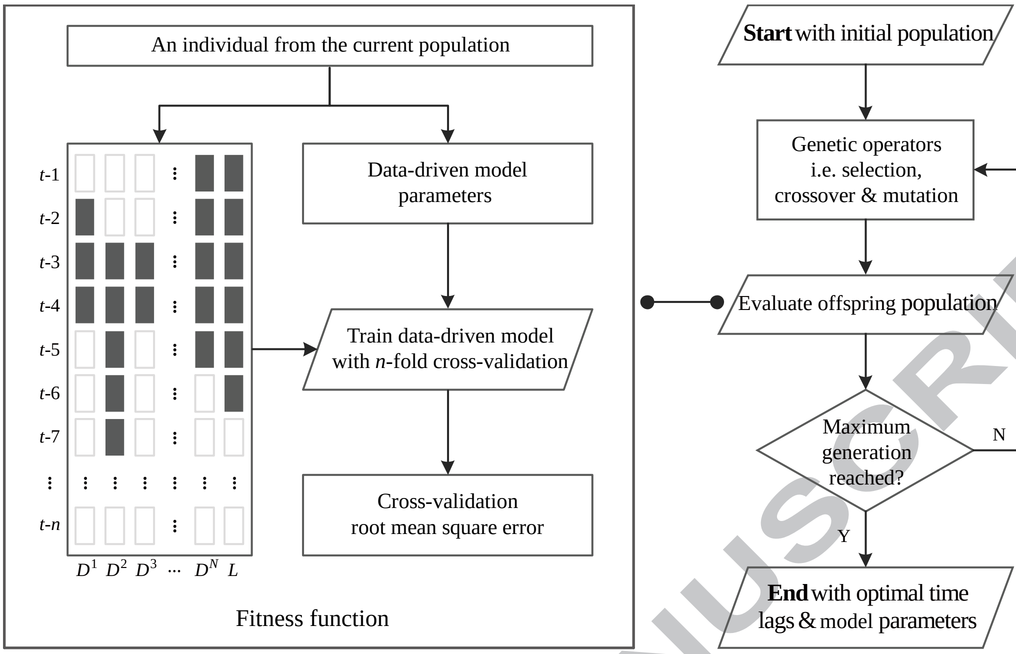

Fig. 3. Diagram of the GA-based synchronized search for the optimal input variable time lags and data-driven model parameters. In the Fitness function, D^1, D^2, D^3,…, D^N are discharges of N rivers; L is the local water level; gray blocks indicate the selected time lags.

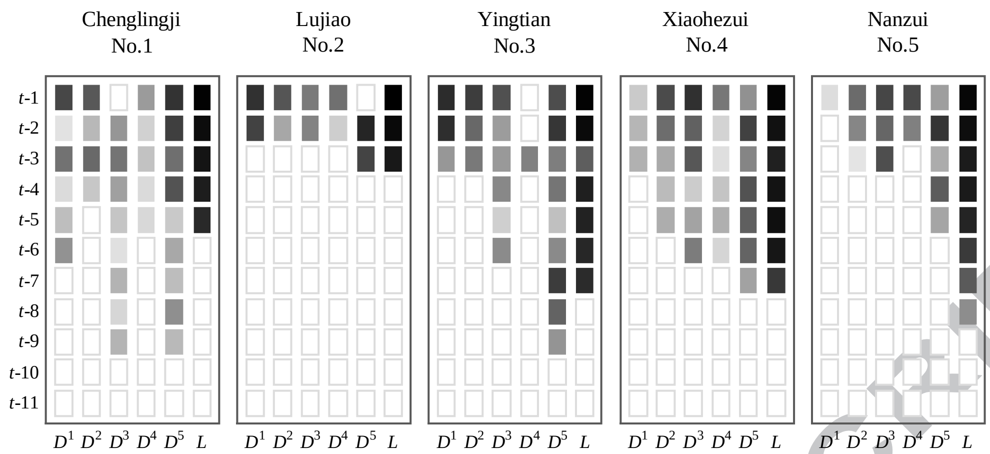

Fig. 4. The selected input variable time lags (in days) for the Dongting Lake level prediction. D^1, D^2, D^3, D^4, D^5 and L represent, respectively, the discharges of the Xiang, Zi, Yuan, Li and Yangtze River, and the local water level. Sensitivity analysis was conducted for the selected time lags. The model input corresponding to each time lag in the training period was altered by ±10%. The median value of the absolute differences in model-predicted lake levels was used to indicate the time lag's effect on lake level variations. All the time lags' effects were then ranked together. The darker the block is, the stronger its effect is.

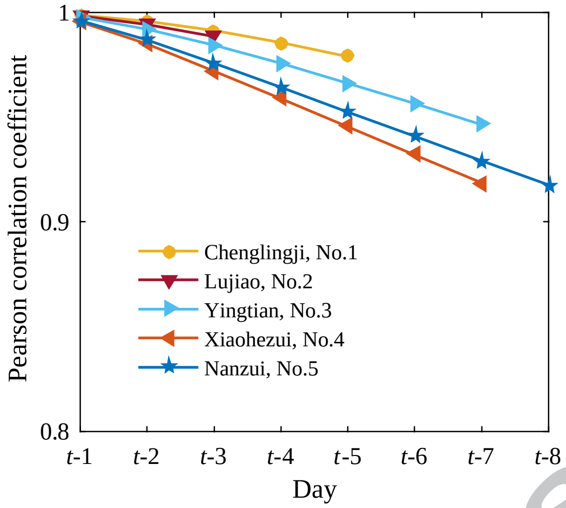

Fig. 5. Correlation between the current lake level and its states at the previous time steps.

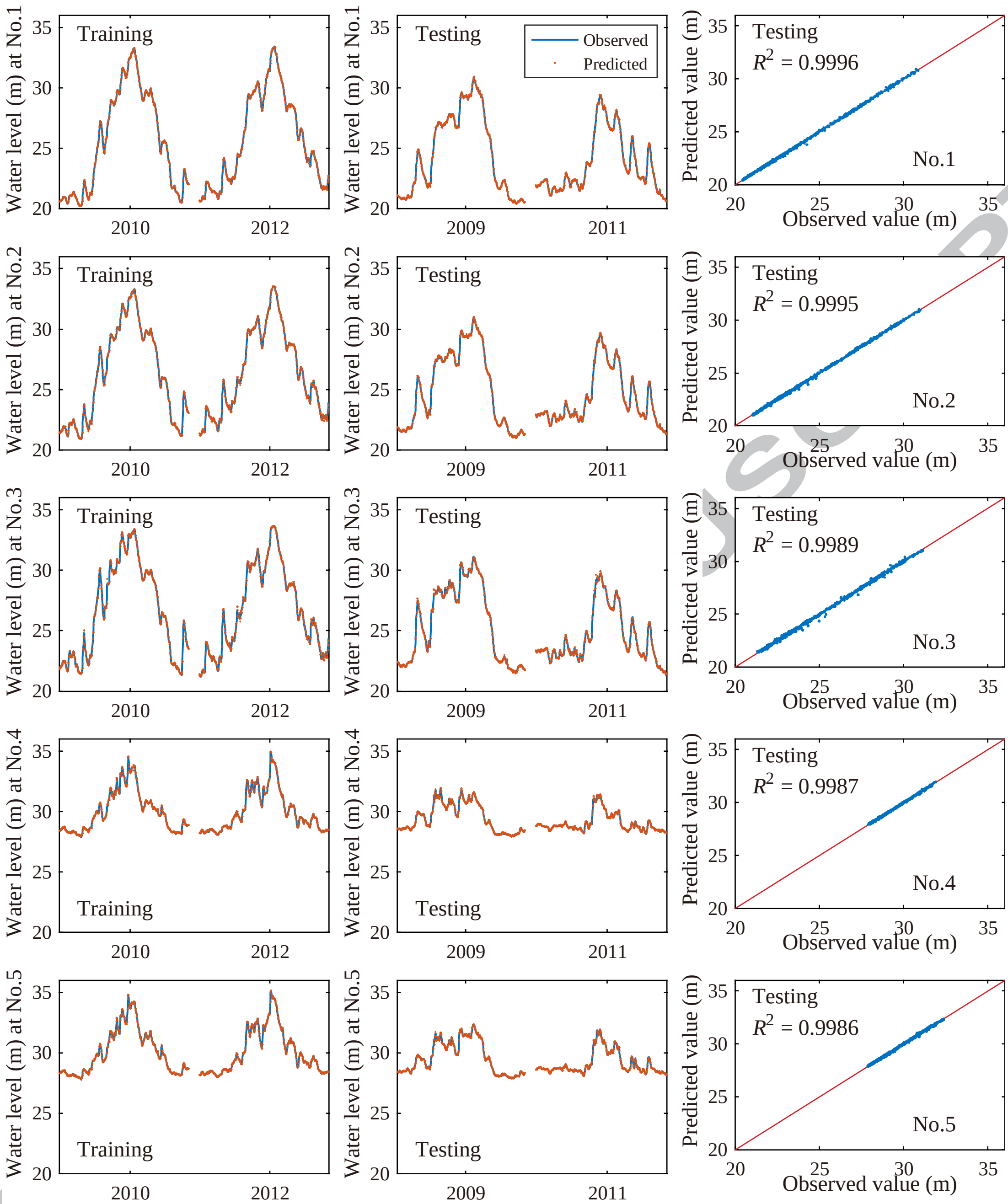

Fig. 6. Comparisons between the observed and predicted lake levels.

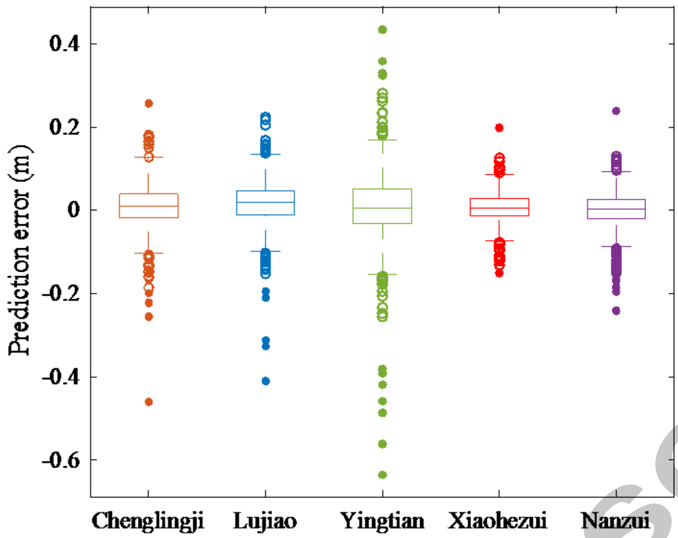

Fig. 7. Boxplots of the lake level prediction errors in the testing period.

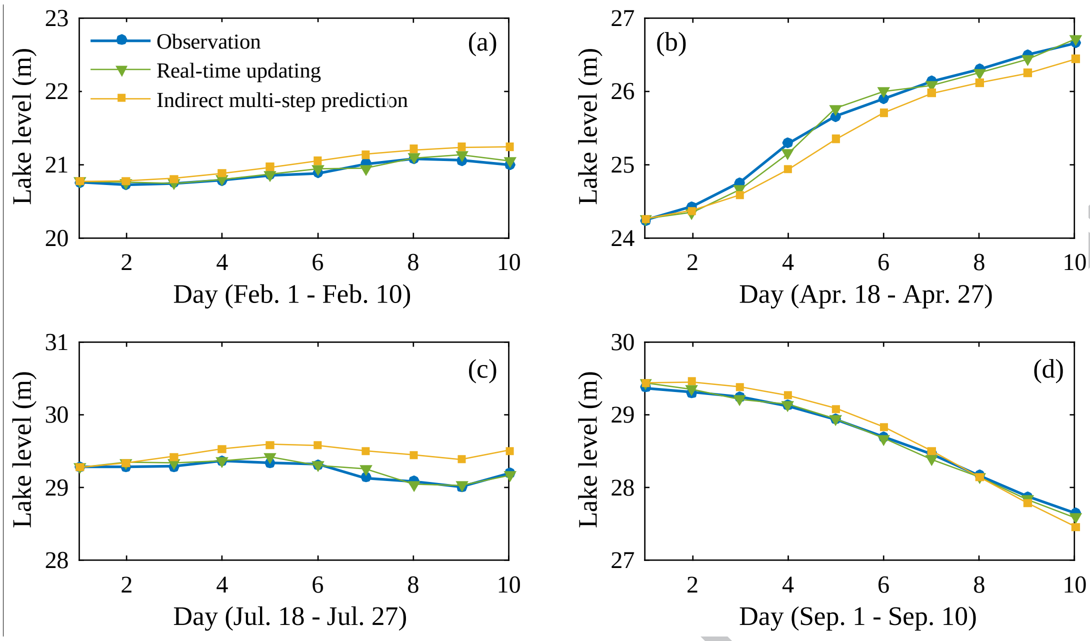

Fig. 8. Comparisons between the observed and predicted lake levels at Chenglingji in (a) dry period, (b) water-level rise period, (c) wet period and (d) drawdown period.

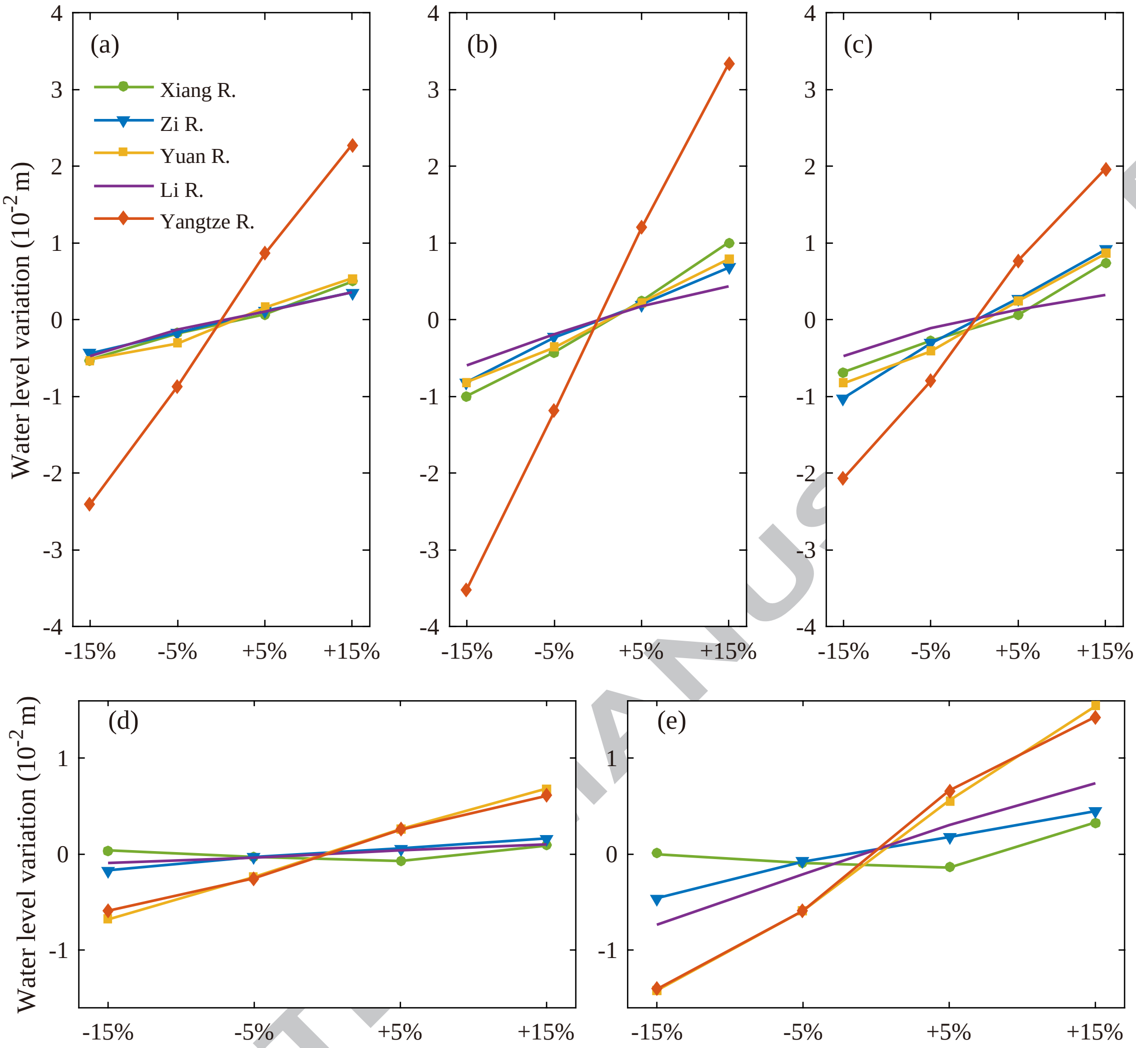

Fig. 9. Main effects of different rivers on the Dongting Lake level variations at (a) Chenglingji, (b) Lujiao, (c) Yingtian, (d) Xiaohezui and (e) Nanzui. The horizontal axis is the change in river discharge.

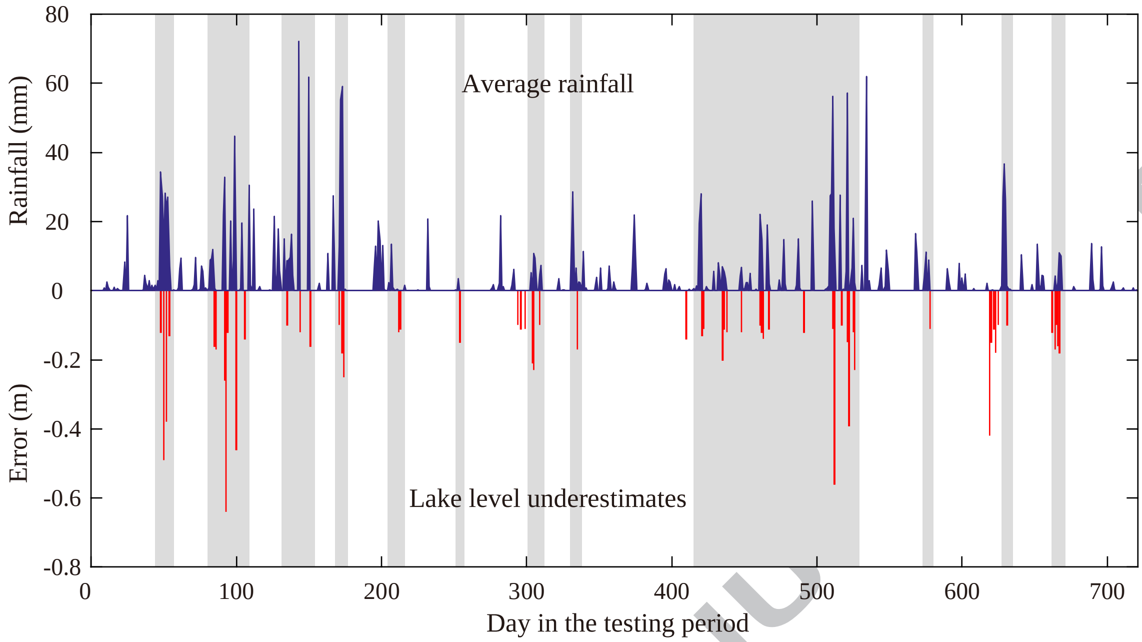

Fig. 10. Plots of average daily rainfall at stations P1 and P2 (top) and serious underestimates of the observed lake levels (< -0.10 m) at Yingtian (bottom). Gray columns indicate periods when the underestimates are closely related to the preceding rainfall.

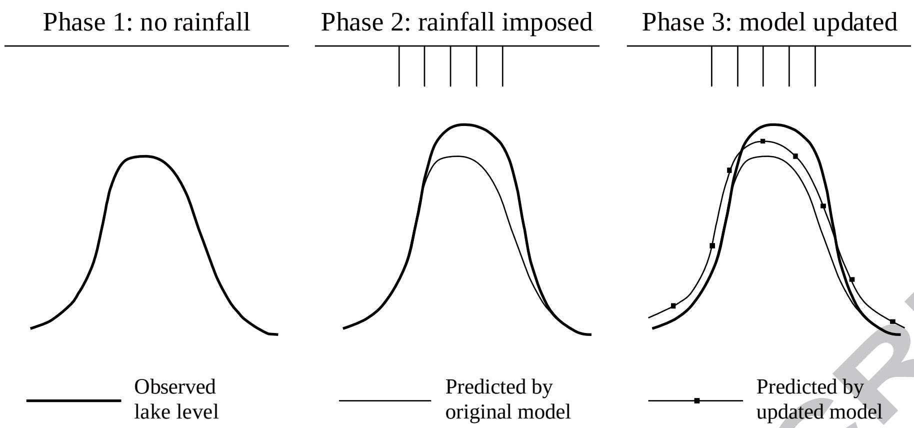

Fig. 11. Schematic diagram illustrating the changes in model behavior due to model parameter optimization. In Phase 1, the observed lake levels overlap the original model predictions.

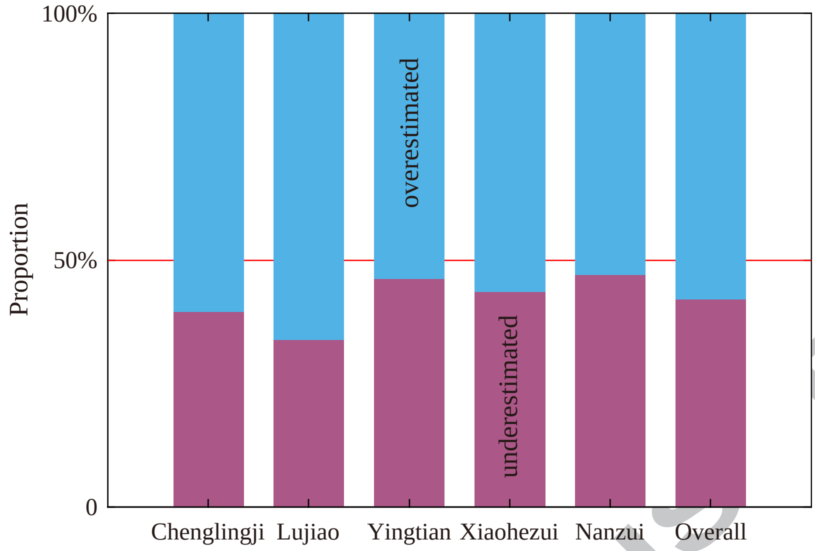

Fig. 12. Proportions of overestimated and underestimated lake levels in the testing period.

List of Tables

Table 1. Statistical characteristics of the hydrological data.

Station	Location	Data type	Dataset[a]	Minimum value	Maximum value	Mean value	Standard deviation
Chenglingji No.1	Dongting	Water level (m)	Training	20.21	33.40	25.66	3.95
			Testing	20.43	30.86	24.13	2.94
Lujiao No.2	Dongting		Training	20.90	33.51	26.31	3.67
			Testing	21.01	30.98	24.68	2.75
Yingtian No.3	Dongting		Training	21.21	33.67	26.69	3.66
			Testing	21.32	31.15	25.05	2.75
Xiaohezui No.4	Dongting		Training	27.89	34.93	29.99	1.68
			Testing	27.91	31.91	29.27	1.02
Nanzui No.5	Dongting		Training	27.78	35.14	30.08	1.80
			Testing	27.85	32.36	29.36	1.20
Xiangtan #1	Xiang R.	Flow rate (m^3/s)	Training	504.0	18400.0	2365.7	2421.3
			Testing	421.0	9090.0	1414.3	1186.9
Taojiang #2	Zi R.		Training	103.0	4450.0	729.1	645.4
			Testing	108.0	3090.0	549.2	449.1
Taoyuan #3	Yuan R.		Training	104.0	18600.0	2150.5	2167.3
			Testing	206.0	9390.0	1475.0	1355.4
Shimen #4	Li R.		Training	16.5	7330.0	485.3	521.5
			Testing	26.2	3850.0	347.6	359.9
Gezhou Dam #5	Yangtze R.		Training	5172.5	46975.0	13517.9	10126.3
			Testing	5097.5	40041.7	11498.0	7368.9
Gaobazhou Dam #5	Qing R.		Training	4.0	938.0	335.9	243.0
			Testing	7.9	960.0	327.5	240.6

[a] Training period: 2010 and 2012; testing period: 2009 and 2011

Table 2. GA parameter setting and search boundaries.

Item	Parameter	Value
GA	Maximum generation	300
	Population size	300
Search boundary	Lower boundary of the time lag	-0.5
	Upper boundary of the time lag	11
	Lower boundary of ε, C and γ	1×10^{-6}
	Upper boundary of ε, C and γ	1×10^{6}

Table 3. Summary of the performance metrics used in this study.

Name	Formula[a]
Root mean square error, RMSE	$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
Coefficient of determination, R^2	$R^2=1-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \Big/ \sum_{i=1}^{n}(y_i-\bar{y})^2$
Mean absolute error, MAE	$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$
Mean relative error, MRE	$\mathrm{MRE}=\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert$

[a] y_i, observed; $\hat{y}_i$, predicted; $\bar{y}$, mean of observations; n, number of observations

Table 4. Optimized SVR parameters.

Parameter	Chenglingji No.1	Lujiao No.2	Yingtian No.3	Xiaohezui No.4	Nanzui No.5
ε	0.0294	0.0616	0.0328	0.0317	0.0160
C	76023.7342	156077.1426	83939.2136	99908.1305	805.7511
γ	0.0008	0.0072	0.0024	0.0012	0.0528

Table 5. Summary of the site-specific lake level model performance.

Station	Training	Testing			
	RMSE (m)	RMSE (m)	R^2	MAE (m)	MRE
Chenglingji (No.1)	0.052	0.057	0.9996	0.041	0.0017
Lujiao (No.2)	0.069	0.061	0.9995	0.045	0.0018
Yingtian (No.3)	0.097	0.091	0.9989	0.061	0.0024
Xiaohezui (No.4)	0.041	0.037	0.9987	0.028	0.0009
Nanzui (No.5)	0.036	0.044	0.9986	0.032	0.0011

Table 6. Orthogonal array design and model simulation results.

Run	River discharge variation[a, b]					Median value of lake level variations (10^{-2} m)				
	D^1	D^2	D^3	D^4	D^5	No.1	No.2	No.3	No.4	No.5
1	1	2	3	3	2	-1.2	-1.9	-1.3	0.0	0.1
2	4	1	2	4	2	-0.6	-0.8	-1.0	-0.5	-0.7
3	2	2	4	4	1	-1.7	-2.8	-1.3	0.1	0.6
4	3	3	4	1	2	-0.5	-0.4	0.1	0.3	0.2
5	3	1	1	3	1	-3.0	-4.5	-3.6	-1.5	-3.2
6	3	2	2	2	4	1.9	3.0	1.3	0.2	0.3
7	2	3	2	3	3	0.7	1.0	0.6	0.1	0.4
8	4	3	3	2	1	-1.6	-2.1	-0.8	-0.2	-0.6
9	4	2	1	1	3	0.3	0.7	0.0	-0.5	-1.4
10	4	4	4	3	4	3.9	6.2	4.8	1.6	4.0
11	3	4	3	4	3	1.9	3.0	2.4	0.7	2.2
12	1	4	2	1	1	-3.2	-4.6	-2.6	-0.7	-2.4
13	1	3	1	4	4	1.8	2.3	1.2	0.1	0.8
14	2	4	1	2	2	-1.2	-1.8	-1.0	-0.9	-1.9
15	2	1	3	1	4	1.5	1.9	0.6	0.6	0.6
16	1	1	4	2	3	0.5	0.2	0.0	0.7	1.4

[a] D^1, D^2, D^3, D^4 and D^5 are as in Fig. 4

[b] Levels of 1, 2, 3 and 4 represent 15% decrease, 5% decrease, 5% increase and 15% increase, respectively

Table 7. The maximum correlation coefficients between the lake levels and river discharges at various time lags.

Station	D^1	D^2	D^3	D^4	D^5
Chenglingji (No.1)	0.42	0.47	0.58	0.49	0.88
Lujiao (No.2)	0.44	0.50	0.59	0.50	0.85
Yingtian (No.3)	0.48	0.53	0.61	0.51	0.81
Xiaohezui (No.4)	0.51	0.56	0.74	0.61	0.77
Nanzui (No.5)	0.45	0.51	0.66	0.57	0.84

- An integrated methodology is developed for lake water level management.
- Input variables and parameters of lake level models are optimized simultaneously.
- The antecedent lake levels are crucial to the prediction of the current lake level.
- The predicted lake levels agree very well with the observed data ($R^2 \geq 0.9986$).
- The relative contributions of different rivers to lake level changes are analyzed.