Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China

Bing Li, Guishan Yang, Rongrong Wan, Xue Dai and Yanhui Zhang

ABSTRACT

Modeling of hydrological time series is essential for sustainable development and management of lake water resources. This study aims to develop an efficient model for forecasting lake water level variations, exemplified by the Poyang Lake (China) case study. A random forests (RF) model was first applied and compared with artificial neural networks, support vector regression, and a linear model. Three scenarios were adopted to investigate the effect of time lag and previous water levels as model inputs for real-time forecasting. Variable importance was then analyzed to evaluate the influence of each predictor for water level variations. Results indicated that the RF model exhibits the best performance for daily forecasting in terms of root mean square error (RMSE) and coefficient of determination (R^2). Moreover, the highest accuracy was achieved using discharge series at 4-day-ahead and the average water level over the previous week as model inputs, with an average RMSE of 0.25 m for five stations within the lake. In addition, the previous water level was the most efficient predictor for water level forecasting, followed by discharge from the Yangtze River. Based on the performance of the soft computing methods, RF can be calibrated to provide information or simulation scenarios for water management and decision-making.

Key words | artificial neural networks, lake water level, Poyang Lake, random forests, support vector regression, variable importance analysis

Bing Li
Guishan Yang (corresponding author)
Rongrong Wan
Xue Dai
Yanhui Zhang
Key Laboratory of Watershed Geographic Sciences,
Nanjing Institute of Geography and Limnology,
Chinese Academy of Sciences,
Nanjing 210008,
China
E-mail: *gsyang@niglas.ac.cn*

Bing Li
Xue Dai
University of Chinese Academy of Sciences,
Beijing 100049,
China

INTRODUCTION

Lake water level forecasting has important applications for identifying the main influencing factors of water level fluctuations, determination of the watershed hydrological cycle variation trends under projections of global climate changes, integration of reservoir management schemes, and ensuring sufficient freshwater supply (Wantzen *et al.* 2008; Hu *et al.* 2008; Kourgialas *et al.* 2015). However, lake water level variations involve a complex nonlinear process, which integrates precipitation, discharge from tributaries, topography, and so on. The variations become even more complex when the lake interacts with a large river (e.g., the interaction of the Poyang Lake with the Yangtze River). Reliable and accurate forecasting of lake water level has always been a challenge for hydrologists and water resource managers.

In recent decades, numerous forecasting techniques, including physically based hydrodynamic models (e.g., CHAM, MIKE21, and EFDC), time series analysis (e.g., auto-regressive moving average and auto-regressive integrated moving average), and soft computing methods (e.g., artificial neural networks (ANNs), support vector regression (SVR), and model trees), have been developed to simulate

doi: 10.2166/nh.2016.264

hydrological time series. (Belmans *et al.* 1983; Hsu *et al.* 1995; Dawson & Wilby 2001; Alvisi *et al.* 2006; Khan & Coulibaly 2006; Altunkaynak 2007; Lai *et al.* 2013; Li *et al.* 2013). In particular, physically based hydrodynamic models exhibit the best performance in forecasting water level. However, these methods require detailed terrain data, as well as complex boundaries and parameters as input, and are computationally expensive and limited to restricted duration (Li *et al.* 2015). Time series analysis is more complex and unreliable than the neural network model (Altunkaynak 2007). In addition, time series analysis does not consider the nonstationary and nonlinear characteristics of data structure (Kumar & Maity 2008). It is difficult to use nonlinear and complex exhibition of model variables for accurate quantification of uncertainty associated with the predictions, which often mislead water resource managers during decision-making (Aqil *et al.* 2007; Mustafa *et al.* 2012). Soft computing methods are capable of capturing complex nonlinear relationships between inputs and outputs without the need for explicit knowledge of the physical process, and they also avoid the creation of extremely complex models in the rare cases when all information is available (Trichakis *et al.* 2011). Soft computing methods, particularly ANN and SVR, have been successfully applied to solve nonlinear problems in hydrological series simulations, such as groundwater level forecasting (Daliakopoulos *et al.* 2005; Yoon *et al.* 2011; Gholami *et al.* 2015), rainfall prediction (Chau & Wu 2010), and surface water level/discharge forecasting (Altunkaynak 2007; Callegari *et al.* 2015).

The random forests (RF) model has been proposed as a new soft computing method by Breiman (2001). RF handles nonlinear and non-Gaussian data well, is amenable to model interpretation, and is free of over-fitting problems as the number of trees increases. Furthermore, RF provides a measure of the relative importance of descriptors, which can be further utilized in variable selection (Genuer *et al.* 2010). In the past few years, RF has been employed to simulate suspended sediment concentration and soil organic carbon stocks (Francke *et al.* 2008; Were *et al.* 2015). However, soft computing methods with different algorithms may have different levels of adaptability for diverse problems. For example, Yoon *et al.* (2011) found that the performance of ANN is better than that of SVM in the model training and testing stages when predicting groundwater level in a coastal aquifer. Rodriguez-Galiano *et al.* (2015) found that the RF method performs better than ANN and SVM in predicting and mapping mineral prospectivity. Were *et al.* (2015) concluded that RF has the highest tendency for overestimation, and that SVR is the best model for predicting soil organic carbon stocks. However, few studies have compared the adaptability and accuracy of different soft computing methods for hydrological series forecasting, especially for highly nonlinear water level forecasting. In the present work, the RF model was first utilized for forecasting water level fluctuations and then compared with commonly used ANN, SVR, and a linear model (LM) in terms of accuracy.

Poyang Lake, the largest freshwater lake in China, is fed by five main tributaries and is connected to the Yangtze River, whose blocking effect (even intrusion) greatly affects water level variations in the lake. In recent decades, intensified global climate changes and anthropogenic activities have greatly altered Poyang Lake's water regime to some extent (Guo *et al.* 2008), with more frequent occurrence of floods and droughts, which take on a trend of sharp transformation (Guo *et al.* 2012; Li & Zhang 2015). Building a dam has been proposed in the downstream area of the lake to alleviate severe droughts and flood risk in Poyang Lake (Huang *et al.* 2015). A few researchers have studied Poyang Lake water level forecasting (Jiang & Huang 1997; Lan 2014; Li *et al.* 2015). However, few have taken into account the effects of both time lag and the previous hydrological status of the lake. The time lag effect was proved to be important for hydrological series forecasting (Aqil *et al.* 2007; Chau & Wu 2010; Bao *et al.* 2014). Chau & Wu (2010) found notable differences at 1-, 2-, and 3-day ahead by using partial autocorrelation for daily rainfall prediction using an ANN model. Cross-correlation has been used to determine lag times of precipitation and discharge (Yoon *et al.* 2011; Li *et al.* 2015). The trial-and-error method was also utilized to obtain the most sensitive time lag (Hipni *et al.* 2013). Therefore, an accurate water level forecasting model that considers the previous hydrological status and the time lag effect is required to provide suggestions for the development and management of water resources. Such a model can also help identify the main factors that influence water levels in Poyang Lake.

The specific objectives for this paper were: (1) to determine a model of highest accuracy by comparing RF with ANN, SVR, and the LM model, and incorporating discharge from lake catchment tributaries and the Yangtze River, the

time lag effect and the previous hydrological status for water level forecasting; and (2) to explore the relative importance of each predictor for different water level stations within the lake. The proposed model provides a useful tool for water resource management and for identifying the major influencing factors for lake water level fluctuations.

MATERIALS AND METHODS

Study area

Poyang Lake is located at 115 °47′–116 °45′E and 28 °22′–29 °45′N on the southern bank of the Yangtze River, which is the second largest river in the world (Figure 1). It is fed primarily by five tributaries: the Gan, Fu, Xin, Rao, and Xiu Rivers, and is freely connected with the Yangtze River at Hukou. It has a subtropical monsoon climate with an average annual temperature of 17.6 °C and a mean annual precipitation of 1,450–1,550 mm, which is mostly concentrated in summer, leading to considerable intra-annual water level variations. As the streamflow varies by season, the surface area of Poyang Lake can fluctuate greatly from less than 1,000 km^2 in the dry season to approximately 4,000 km^2 during the rainy season (Shankman *et al.* 2006), when it can be described as 'flooding like sea, while drying like thread'. Poyang Lake has an angular surface from south to north, with five representative hydrological

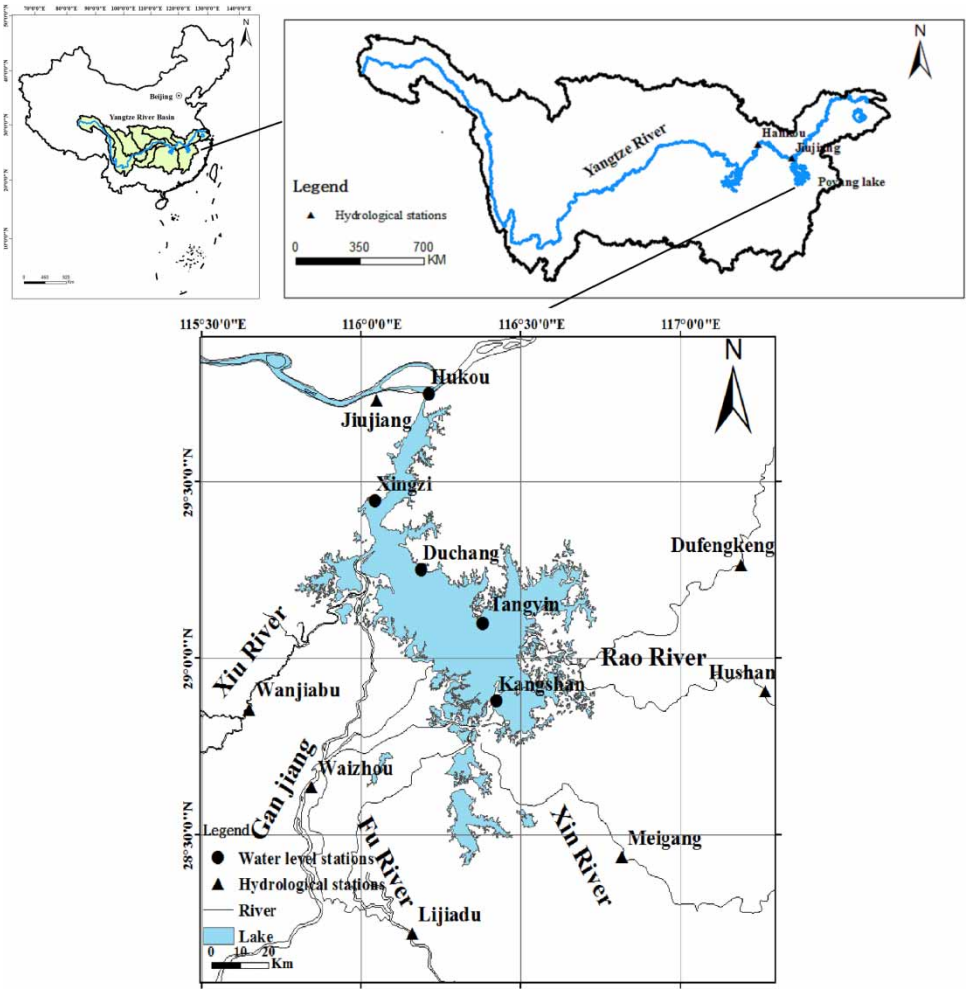

Figure 1 | Location of study area and hydrological stations.

stations in different parts: Hukou, Xingzi, Duchang, Kangshan, and Tangyin (from north to south) (Figure 1).

Data collection

As shown in Figure 1, Jiujiang station is the closest representative of the Yangtze River to affect water level variations within the lake. Meanwhile, given the missing discharge data in Jiujiang station prior to 1988, Hankou station was chosen to be a substitute in the model, as it has significant correlation with Jiujiang station (correlation coefficient = 0.995). The data applied in this study include the following: (1) daily discharge observations of six hydrological stations in the lake's catchment and the Yangtze River from 1955 to 2012, namely, Waizhou (wz), Hushan (hs), Dufengkeng (dfk), Lijiadu (ljd), Meigang (mg), and Wanjiabu (wjb) stations of the upstream tributaries (the Gan, Fu, Xin, Rao, and Xiu Rivers) and Hankou station of the Yangtze River (Table 1); and (2) daily water level observations of five gauge stations within the lake, namely, Hukou, Xingzi, Duchang, Tangyin, and Kangshan stations (from north to south). Observation at Tangyin station started in 1962; thus, only the 1962–2012 data were considered from this station. The observations on water level and discharge were obtained from the Hydrological Bureau of Jiangxi Province. Figure 1 shows the locations of these hydrological stations.

Table 1 | Characteristics of input and output data used in this study

Inputs and outputs	Stations	Duration	Drainage area[a] (10^4 km^2)
Gan River	Waizhou	1955–2012	8.12
Fu River	Lijiadu	1955–2012	1.58
Xin River	Meigang	1955–2012	1.55
Rao River			
Chang River	Dufengkeng	1955–2012	1.5
Lean River	Hushan	1955–2012	
Xiu River			
Liao River	Wanjiabu	1955–2012	1.48
Yangtze River	Hankou	1955–2012	–
Poyang Lake	Hukou	1955–2012	–
	Xingzi	1955–2012	–
	Duchang	1955–2012	–
	Tangyin	1962–2012	–
	Kangshan	1955–2012	–

[a]Data are from the study on Poyang Lake (the Editorial Committee of 'Study on Poyang Lake' 1987).

Soft computing methods

RF model

The RF model employs the strategy of a random selection of a subset of m predictors to grow a binary tree, where each tree is grown on a bootstrap sample of the training set (Breiman 2001) (Figure 2). This algorithm is an extension of bagging, and a competitor to boosting (Polikar *et al.* 2012). Regression trees imply no assumptions of distribution of data (Francke *et al.* 2008). For each tree, the response data were grouped into two descendant nodes to maximize homogeneity, and the best binary split was selected. Each descendant node of the selected split was treated similarly to the original node, and the process continued recursively until a stop criterion was met. All the trees were grown to their maximum sizes, and final predictions were obtained from the averaged results (Breiman 2001). In RF modeling, three parameters need to be specified: (1) the number of trees to grow in the forests (n_{tree}), which is the most important parameter of RF; (2) the number of randomly selected predictor variables at each node (m_{try}); and (3) the minimal number of observations at the terminal nodes of the trees (*nodesize*). The default number of trees was 500, although more stable results for estimating variable importance

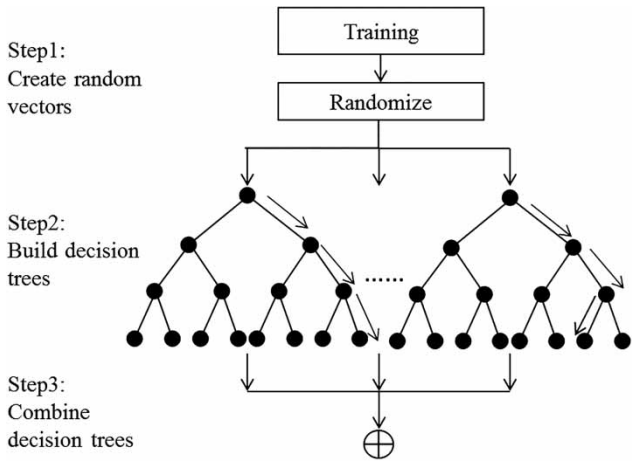

Figure 2 | Demonstration of the RF methodology (Malekipirbazari & Aksakalli 2015).

could be achieved with a higher number of trees (Were *et al.* 2015). As well, the importance of each predictor is measured by increased mean squared errors (MSEs) as the predictors were excluded one by one from RF models. The relative importance of each predictor is determined from 100 runs of the RF models and normalized to 100% to provide a simple basis for comparison in different stations. In this paper, 500 parameter sets including n_{tree}, m_{try}, and *nodesize* for the RF model were tried and the one with the highest accuracy was selected.

SVR model

SVR is a forecasting model based on the structural risk minimization principle, and aims at minimizing a bound on the generalization error (Smola & Schölkopf 2004; Vapnik 2013). Several advantages of SVR are its improved generalization ability, unique and globally optimal architectures, and the ability to be rapidly trained (Lan 2014). One of the highlights of SVR is flexibility, depending on different types of kernel function such as the linear, polynomial, and radial basis function (RBF) kernel (Lan 2014). The linear kernel is a special case of RBF, and RBF can better handle the case when the relationship between inputs and outputs is nonlinear (Lin *et al.* 2006). Moreover, the RBF kernel involves fewer numerical difficulties than the polynomial kernel, which has more hyperparameters than the RBF kernel (Lin *et al.* 2006). Hence, the commonly used RBF kernel is adopted in the present study. The input data are projected using the RBF kernel to hyperspace, where a complex nonlinear relationship can be simply presented and solved (Yoon *et al.* 2011; Wei 2012) (Figure 3). Given the training data $\{(x_1, y_1), \cdots (x_{1n} y_n)\}$, where x_i and y_i are the input and output data, respectively. The goal of ε-SVR is to determine a function f(x) that has the most ε deviation from the input data and that is as flat as possible (Smola & Schölkopf 2004). The formula of the RBF kernel is:

$$f(x, w) = \sum_{j=1}^{n} w_j \exp\left(-\gamma \|x - x_j\|^2\right),$$

where γ is a parameter and vector x_j is the input of the training data. The unknown vector of w is determined to minimize the function:

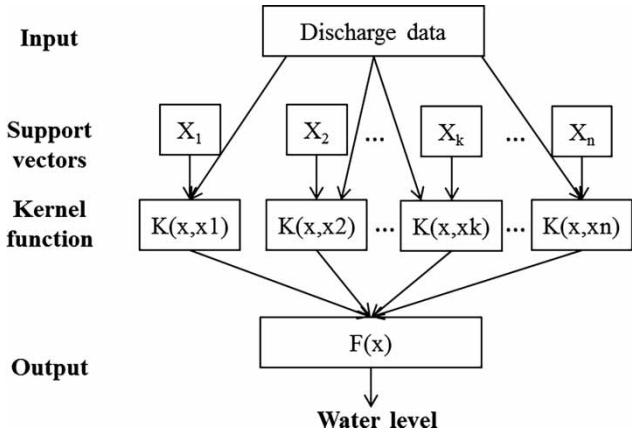

Figure 3 | Demonstration of the SVR methodology (Yoon *et al.* 2011).

$$\min_{w \in R} \cdots \frac{1}{2}\|w\|^2 + C \bullet \sum_{i=1}^{n} \max\left(|y_i| - f(x_i, w) - \varepsilon, 0\right)$$

where C (cost) >0 controls the tradeoff between the flatness of f(x), and deviations greater than ε are tolerated.

An internationally recognized uniform method for SVM parameter optimization has not been established. This study adopted the most commonly used method, in which γ, C and ε are calibrated in a certain range by grid search in R statistical environment. Similarly, 500 pairs of parameters were tried and the set with the best performance was selected.

ANN model

The ANN model was initially formed as a simplified model parallel to the 'biological' model. It simulates the characteristics of the human neural network to deal with distributed parallel information. A common ANN architecture consists of input, hidden, and output layers with node activation functions (Hsu *et al.* 1995) (Figure 4). ANNs with one hidden layer are commonly used in hydrologic modeling (Dawson & Wilby 2001). Although numerous ANN algorithms have been proposed, such as Elman recurrent, RBF, and Hopfield neural networks (Kecman 2001), the error back-propagation algorithm (BPA) introduced by Rumelhart *et al.* (1985) may be the most commonly applied algorithm, which adjusts the connecting weights in the direction where the error

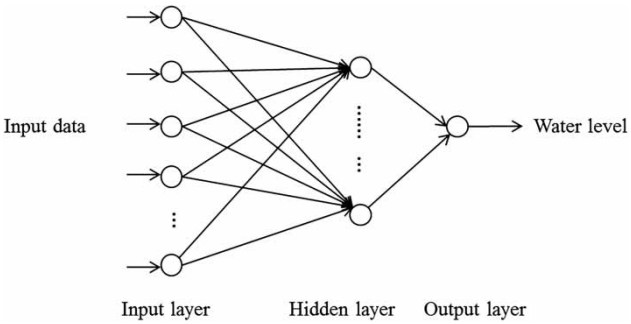

Figure 4 | Demonstration of the ANN methodology (Were *et al.* 2015).

performance function decreases most rapidly to minimize the training error (Sulaiman *et al.* 2011; Trichakis *et al.* 2011). Through iterative propagation of errors back to the network, the differences between the output and target are consistently sent to the learning process, to automatically adjust and readjust the connection weights between the elements, until a desirable network output is achieved (Dawson & Wilby 2001; Trichakis *et al.* 2011; Were *et al.* 2015). In the present study, a multilayer perception network with one hidden layer using the BPA learning algorithm was trained to establish the ANN model. The activation function consists of a log-sigmoid function in the hidden layer and a linear function in the output layer. The input is normalized by subtracting each column of the data set by its mean value, and divided by the standard deviation. Furthermore, a trial-and-error method based on performance value from the training stage was applied to determine the optimal hidden nodes (*size*), learning rate, and momentum. Similar to the RF model, 500 parameter sets were generated and tested to determine the optimal parameter set. The set with the smallest hidden neurons giving the best performance was chosen.

Model training and evaluation

Input–output scenarios

For all of the soft computing methods regarded in this study, the daily discharge observations of the seven hydrological stations were used as input variables, and the water level of the five gauge stations within Poyang Lake was chosen as output in each model. These specific variables were selected based on the unique hydrological processes of Poyang Lake, in which water level variations are mainly influenced by discharge from catchment and the Yangtze River. In addition, describing the hydrological processes through a nonlinear model is necessary (Chen *et al.* 2015). For simplicity, the daily water level forecasting of Hukou station was used as an example for evaluating the performance of RF, SVR, ANN, and LM. Furthermore, to examine the effect of time lag and previous lake water level on forecasting performance, three input scenarios were developed: (1) the current daily discharge from the seven hydrological stations of the Poyang Lake tributaries and the Yangtze River; (2) the daily discharge of the seven stations between day (t) and (t-5); and (3) on the basis of scenario 2, the average lake water level over the previous week (wl7) was incorporated. The trial-and-error method was then used and the time lag with the highest accuracy of water level forecasting for Poyang Lake was determined as most sensitive.

Performance evaluation

The popular v-fold cross-validation, which provides a good trade-off between model over-fitting and under-fitting, was employed to evaluate the performance of the candidate models (Yoon *et al.* 2011; Hipni *et al.* 2013). The entire data sets (daily records from 1955 to 2012) were randomly partitioned into v equal-sized subsets. During each modeling process, one of the partitions was used for validation, while the others were used for training. Furthermore, the modeling process was repeated v times, and the performance metrics were averaged to achieve the final performance. In reference to similar studies, using a v of 5, 10, and 20 could result in slightly different error estimates, which are often not significant (Feng *et al.* 2005; El-Shafie & Noureldin 2010; Hipni *et al.* 2013). Therefore, five-fold cross-validation is used to evaluate the performance of the models regarded in this work to reduce computing time.

Among several criteria that are commonly used for model performance evaluation, such as the root mean square error (RMSE), coefficient of determination (R^2), mean absolute relative error, Nash–Sutcliffe efficiency coefficient (NSCE), and Akaike information criterion, RMSE, R^2, and NSCE were selected in this study (Lin *et al.* 2006; Ghorbani *et al.* 2010; Lan 2014). RMSE measures the residual value between the measured and forecasted lake water level, and records in real

units of the water level. R^2 measures the degree of colinearity between the observed and predicted values. It also describes the proportion of the total variance in the observed data that can be explained by the model. The NSCE is a popular index to assess the predictive power of hydrological models. In this work, the NSCE was used for evaluating the sensitivities of each parameter set. An RMSE value of 0, R^2 and NSCE value of 1 are pursued in the best forecast models. The formulas for RMSE, R^2, and NSCE are listed as follows:

$$RMSE = \sqrt{\sum_{1}^{n} (y_i - \tilde{y}_i)^2 / n}$$

$$R^2 = \sum_{i=1}^{n} \left[(y_i - \bar{y}_i)(\tilde{y}_i - \bar{\tilde{y}}_i) \right]^2 / \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\tilde{y}_i - \bar{\tilde{y}}_i)^2 \right]$$

$$NSCE = 1 - \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where y_i is the observed water level, $\tilde{y}_i$ is the forecasted water level, and $\bar{y}_i$ indicates the average water level.

The terms 'training' and 'testing' of soft computing models correspond to the calibration and validation of physically based hydrodynamic model. Data preparation and analysis were conducted using Microsoft Excel 2007 and R 3.1.3. Specifically, we have used the implementation and optimization of RF, ANN, and SVR available in the 'randomForests', 'e1071', and 'nnet' package, respectively, in R statistical environment (Team 2014). The model parameter sensitivity analysis was conducted by generating 500 model parameter sets for each model. The model performance was estimated using here the NSCE value for each parameter set.

RESULTS AND DISCUSSION

Comparison of the models for daily water level forecasting

Figure 5 shows the relationship between measured water level and those predicted by the RF, SVR, ANN, and LM using five-fold cross-validation. The RF simulation results exhibit tighter data distribution with better linearity than its counterparts concerned, not only in the training stage ($R^2 = 0.993$) but also in the testing stage ($R^2 = 0.965$). The high value of R^2 in the testing period implies that the RF model is the most efficient for water level forecasting. In addition, ANN and SVR models have similar performance in the testing stage ($R^2 = 0.953$), the LM is also applied where the distribution of the data are exponential and far from linear (RMSE = 1.19 m). In general, all the soft computing methods displayed desirable performances with respect to RMSE and R^2, which indicated that the models were fully trained with the 58-year data to provide satisfactory forecasting. In addition, the RMSEs of the ANN and RF models were found to vary with the number of trees and units in the hidden layer (Figure 6); RF regression had a relatively low RMSE for the testing stage (minimum of 0.7 m) and especially for the training stage (minimum of 0.32 m). Moreover, the RMSE was stable for each neuron in the hidden layers (Figure 6), which indicated that the five-fold cross-validation can provide a good trade-off between over-fitting and under-fitting (Huang *et al.* 2007). The RF regression performance was also stable when the number of trees reached 30. Therefore, RF was chosen to further establish the daily water level forecasting model for the three scenarios.

Effect of time lag on daily water level forecasting

In this subsection, a series of time lags is selected to predict daily water level using RF and five-fold cross-validation under scenario 2 for each station. For simplicity, only the performance during the testing stage is displayed (Table 2). The models with a time lag of 4 days for both the Yangtze River (Hankou station) and the catchment tributaries exhibited the best performance, with the lowest RMSE values of 0.51, 0.55, 0.56, and 0.46 m for Hukou, Xingzi, Duchang, and Tangyin, respectively. Thus, the time lag of discharge from the tributaries and the Yangtze River has a substantial influence on water level variations. The inflow for different intervals of time has significant influence on the predicted flow/water level (Aqil *et al.* 2007). The R(t-3)T(t-3) time lag of the Kangshan station was obtained, which slightly outperforms the time lag of R(t-4)T(t-4) (Table 2). Simply put, the R(t-4)T(t-4) time lag was chosen for the forecasting model of all the five gauge stations in scenarios 2 and 3.

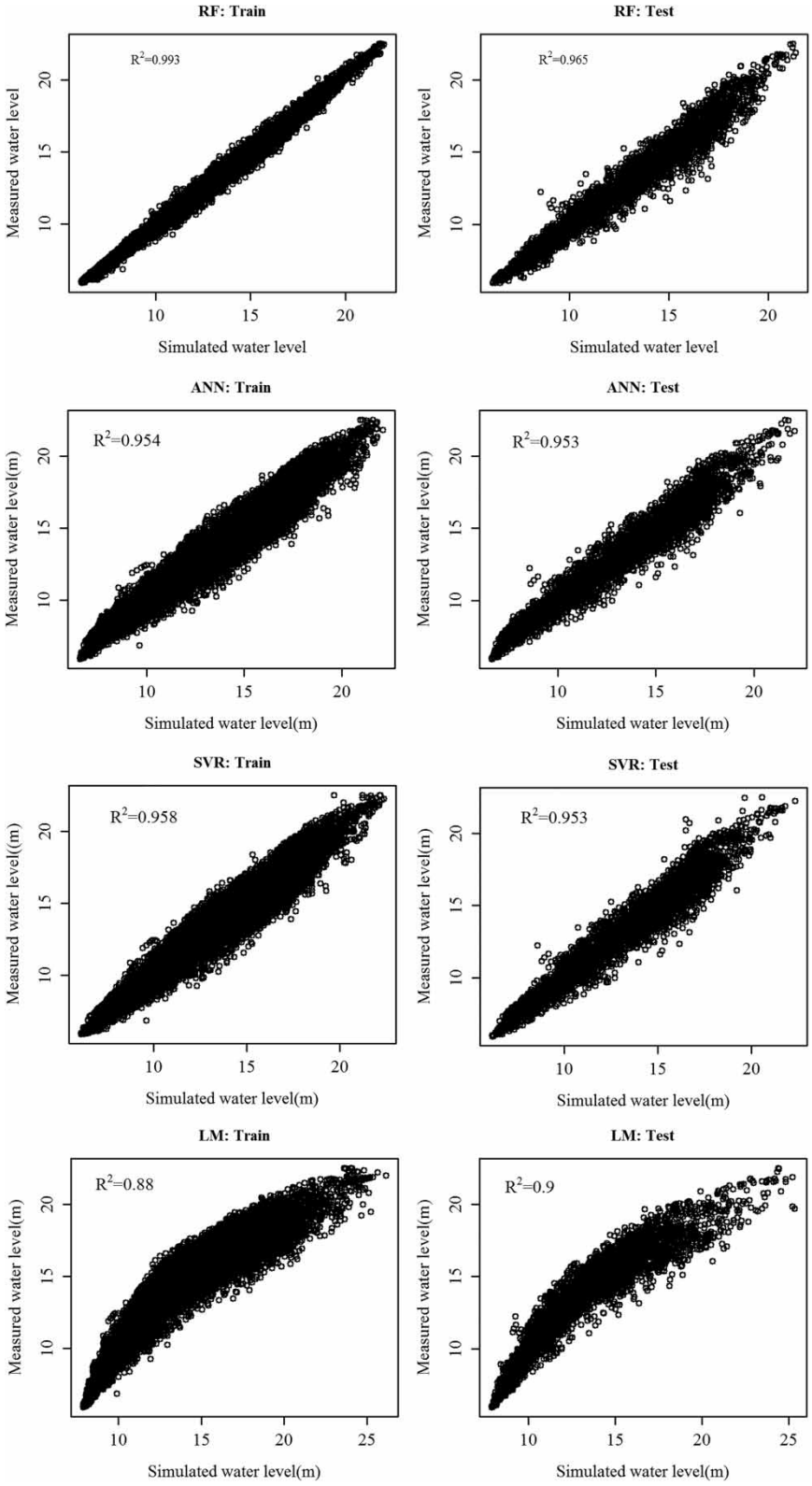

Figure 5 | Forecasting performance of RF, SVR, ANN, and LM for forecasting water level.

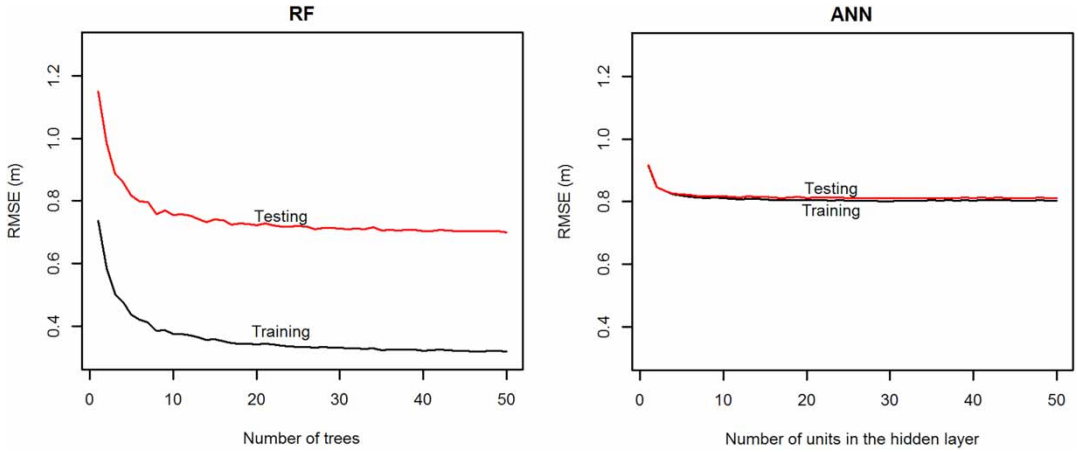

Figure 6 | Comparison of RF and ANN for RMSE variations and model stabilization.

Table 2 | The performance evaluation for scenario 2 using RF and five-fold cross-validation for daily water level forecasting

	Hukou		Xingzi		Duchang		Tangyin		Kangshan	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
R(t)T(t-1)	0.69	0.967	0.71	0.957	0.69	0.943	0.58	0.938	0.52	0.917
R(t)T(t-2)	0.67	0.968	0.69	0.959	0.68	0.945	0.57	0.940	0.52	0.918
R(t)T(t-3)	0.66	0.969	0.68	0.96	0.68	0.946	0.56	0.941	0.52	0.916
R(t)T(t-4)	0.66	0.97	0.68	0.961	0.67	0.946	0.57	0.940	0.54	0.913
R(t)T(t-5)	0.65	0.971	0.68	0.96	0.68	0.945	0.57	0.939	0.55	0.908
R(t-1)T(t-1)	0.62	0.973	0.66	0.963	0.66	0.948	0.54	0.946	0.50	0.926
R(t-1)T(t-2)	0.62	0.974	0.64	0.965	0.65	0.950	0.53	0.947	0.49	0.927
R(t-1)T(t-3)	0.61	0.974	0.64	0.965	0.64	0.951	0.53	0.948	0.50	0.925
R(t-1)T(t-4)	0.60	0.975	0.64	0.965	0.64	0.951	0.54	0.946	0.51	0.920
R(t-1)T(t-5)	0.60	0.975	0.60	0.975	0.65	0.950	0.55	0.944	0.53	0.916
R(t-2)T(t-2)	0.57	0.978	0.60	0.969	0.61	0.956	0.50	0.954	0.46	0.935
R(t-2)T(t-3)	0.56	0.978	0.60	0.97	0.60	0.957	0.50	0.954	0.47	0.932
R(t-2)T(t-4)	0.56	0.978	0.60	0.969	0.61	0.955	0.51	0.952	0.49	0.927
R(t-2)T(t-5)	0.56	0.978	0.61	0.968	0.63	0.953	0.52	0.949	0.51	0.921
R(t-3)T(t-3)	0.53	0.981	0.57	0.973	0.57	0.961	0.47	0.959	**0.45**	**0.938**
R(t-3)T(t-4)	0.53	0.98	0.57	0.972	0.59	0.960	0.48	0.957	0.47	0.933
R(t-3)T(t-5)	0.54	0.98	0.59	0.97	0.60	0.957	0.50	0.953	0.49	0.926
R(t-4)T(t-4)	**0.51**	**0.982**	**0.55**	**0.974**	**0.56**	**0.963**	**0.46**	**0.961**	0.46	0.937
R(t-4)T(t-5)	0.53	0.98	0.58	0.972	0.59	0.959	0.49	0.956	0.48	0.930
R(t-5)T(t-5)	0.53	0.981	0.56	0.973	0.57	0.961	0.47	0.959	0.47	0.933

R and T represent the Yangtze River and tributaries, respectively, and the lowest RMSE and highest R^2 are in bold font.

Effect of input scenarios on daily water level forecasting

Scenario 3 produced the best results among all the three scenarios for all five hydrological stations (Table 3). For the training stage, the values of R^2 are close to 1 for all five stations. Although scenario 3 has the highest R^2, only a slight difference is observed among the three scenarios. By contrast, RMSE decreased from scenario 1 (0.29 m on average) to scenario 3 (0.14 m on average). The forecasting precision in the testing stage is relatively lower than the training stage. Similarly, R^2 increased, whereas RMSE decreased from scenario 1 to scenario 3, with an average RMSE of 0.25 m. Thus, scenario 3 utilizes more information from the input data, indicating that the time lag effect of water level responses to discharge, and previous water level, should be incorporated in establishing a water level forecasting model. The RMSE decreased by 63.8%, 64.4%, 60.6%, 64.4%, and 54.7% in Hukou, Xingzi, Duchang, Tangyin, and Kangshan, respectively, from scenario 1 to scenario 3. Nevertheless, when previous water level data are missing, scenario 2 can also attain a desirable level of precision (Table 3). In addition, for the five hydrological stations within Poyang Lake, the R^2 values slightly decreased from north to south (i.e., longer distance from the Yangtze River) (Table 2). Kangshan has the lowest level of forecasting precision, which indicates that the discharge of the Yangtze River greatly affects the water level within the lake and its effect gradually decreases in the upstream direction. Similar results were obtained by Li *et al.* (2015) using the back-propagation neural network. Thus, scenario 3 is considered the best among the three scenarios for all five water level stations. In other words, the RF algorithm and five-fold cross-validation comprise the best model to forecast the daily water level in Poyang Lake, when the inputs include the 4-day time lag of the Yangtze River, the daily discharge of the tributaries, and the previous water level within the lake.

Source of uncertainty

RF, SVR, and ANN models for forecasting water level fluctuations were compared based on continuous measured data quality controlled by the Hydrological Bureau of Jiangxi Province. The models were calibrated with five-fold cross-validation data in order to reduce the uncertainty. In addition, the training/testing data set represents relatively real-time hydrological processes in lake water level fluctuations, which incorporated the influence of lake catchment

Table 3 | Performance of RF in the training and testing sets using five-fold cross-validation

		Output water level				
Stage	**Metrics**	**Hukou**	**Xingzi**	**Duchang**	**Tangyin**	**Kangshan**
Scenario 1						
Training	R^2	0.994	0.992	0.989	0.988	0.984
	RMSE	0.31	0.32	0.32	0.26	0.24
Testing	R^2	0.966	0.955	0.94	0.935	0.914
	RMSE	0.69	0.73	0.71	0.59	0.53
Scenario 2						
Training	R^2	0.996	0.995	0.993	0.993	0.988
	RMSE	0.23	0.24	0.25	0.2	0.2
Testing	R^2	0.982	0.974	0.963	0.961	0.937
	RMSE	0.51	0.55	0.56	0.46	0.36
Scenario 3						
Training	R^2	**0.999**	**0.999**	**0.998**	**0.998**	**0.997**
	RMSE	**0.11**	**0.11**	**0.13**	**0.09**	**0.11**
Testing	R^2	**0.996**	**0.994**	**0.991**	**0.992**	**0.983**
	RMSE	**0.25**	**0.26**	**0.28**	**0.21**	**0.24**

Lowest RMSE and highest R^2 are in bold font.

tributaries, the Yangtze River discharge, the time lag effect and the previous water level, to ensure that the model 'gives the right answers for the right reasons' (Kirchner 2006). Furthermore, efforts were made to increase the overall goodness of fit of each model by calibrating optimal parameters. For example, selection of hidden neurons and learning rate was based on a trial-and-error optimization process in the ANN model. Grid search cross-validation was implemented to determine optimal C and γ in the SVR model. Moreover, the present work established a water level forecasting model in a specific location, thereby reducing the spatial fluctuation in the output, which consequently reduces the uncertainty (Kourgialas *et al.* 2015). However, although the soft computing methods in the present study have desirable forecasting precision, they only consider streamflow conditions. The main uncertainty may be attributed to the influence of meteorological factors (e.g., precipitation and evaporation) and local inflows on lake water level variations. Thus, the employed models may have limitations when applied to simulate water level variations under possible climate change scenarios (Panagoulia 2006). Nevertheless, the proposed model can be used to provide management schemes under streamflow simulation scenarios (e.g., flood control and drought relief for the Poyang Lake region), specifically, through the combined discharge dispatch of upstream reservoirs and the Three Gorges Dam upstream of the Yangtze River to regulate proper water level variations within the lake.

Figure 7 shows the variability of model performance (NSCE value) for 500 parameter sets of RF, ANN, and

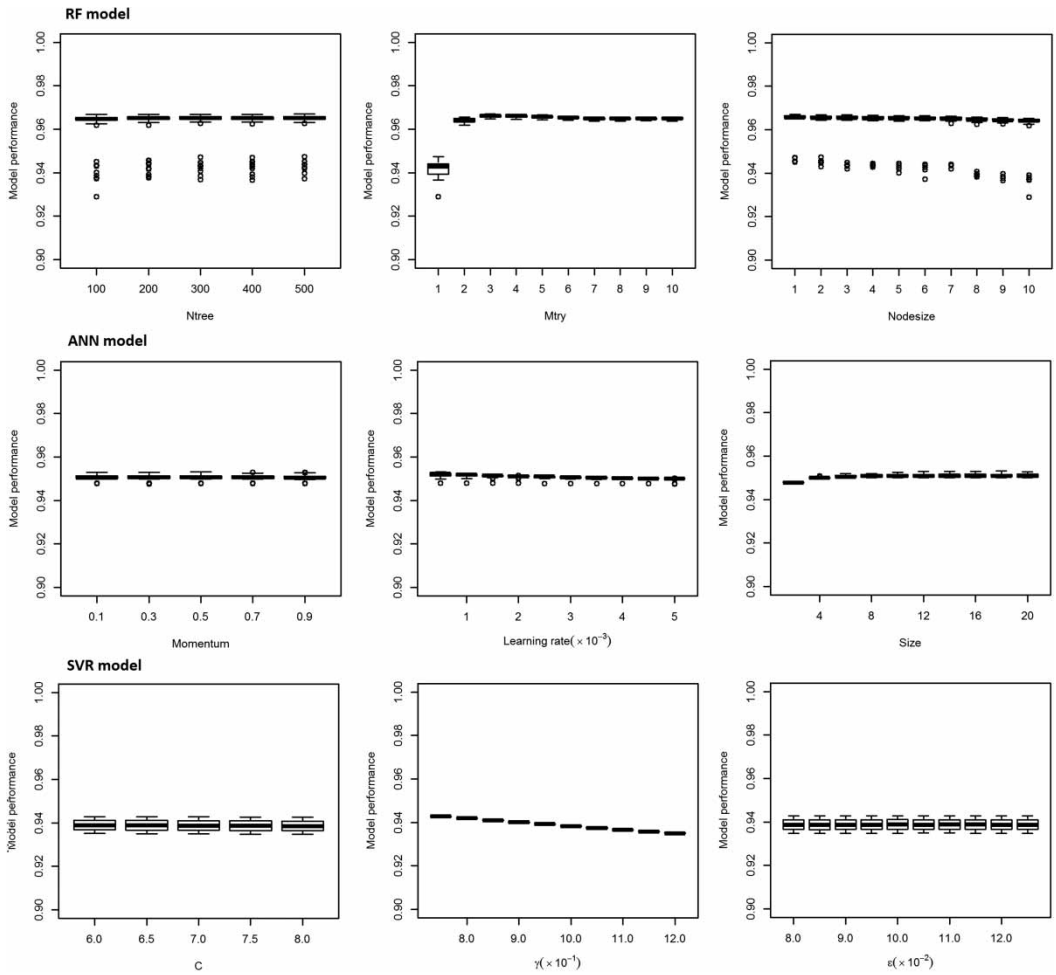

Figure 7 | Variability of model performance (NSCE value) for 500 pairs of parameter sets in the sensitivity analysis.

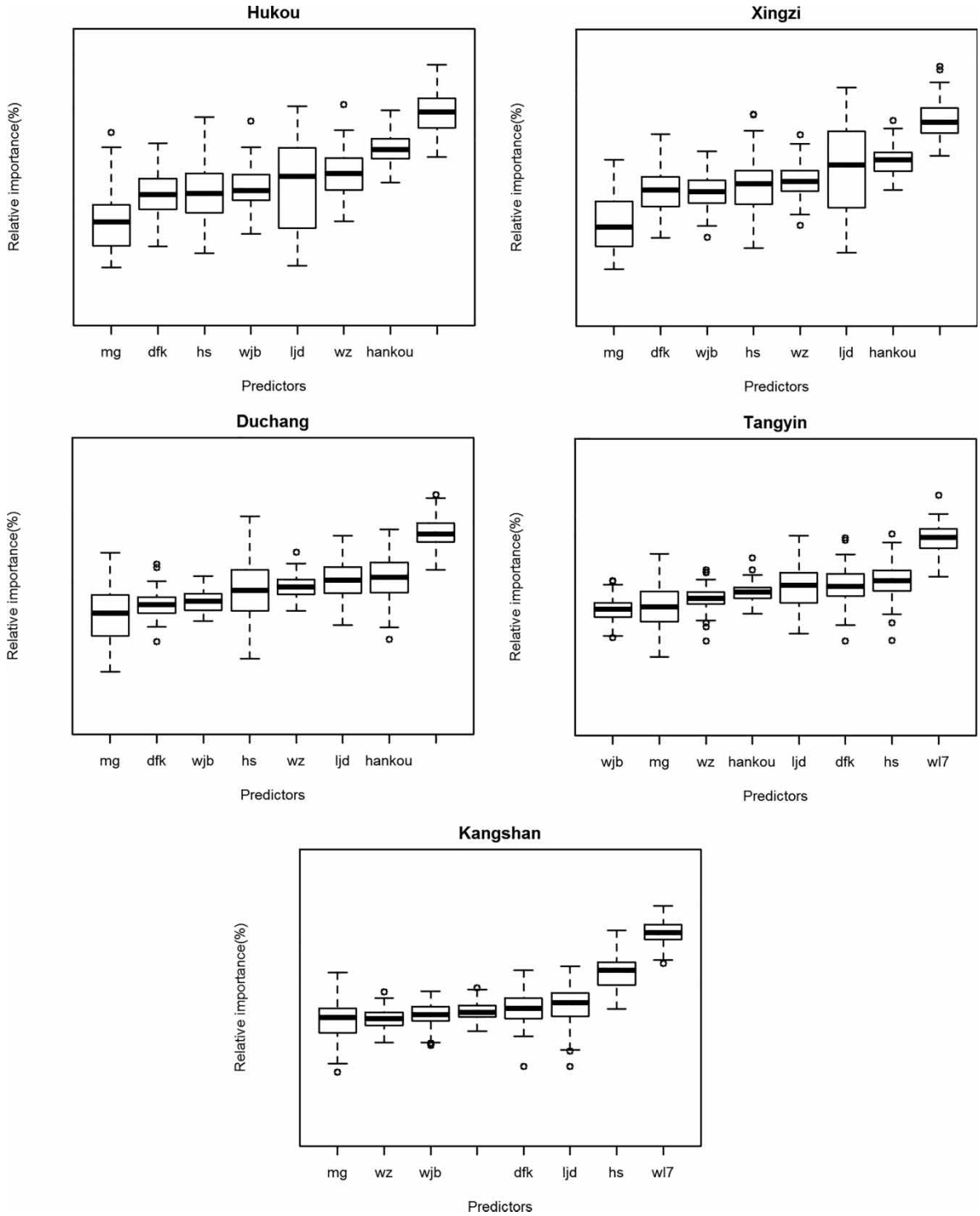

Figure 8 | Relative importance of each predictor as determined from 100 runs of RF models for five water level stations. wz, Waizhou station; mg, Meigang station; wjb, Wanjiabu station (Xiu River); dfk, Dufengkeng station; hankou, Hankou station; ljd, Lijiadu station; hs, Hushan station; wl7, water level of previous 7 days.

SVR model. In this work, it can be seen that the overall sensitivities of the model parameter in each model were low. It is greatly different from the similar analysis of ANN and SVR parameters for groundwater level prediction in a coastal aquifer (Yoon *et al.* 2011), which indicated the models concerned here were fully trained by a large size data set. The model performance showed a relative parallel box plot for each parameter. However, it can be seen that the ANN based on the BPA algorithm is the most stable one. The model uncertainty decreased substantially when the m_{try} of the RF model reached 2, and then become stable. Moreover, Figure 7 also shows a remarkable difference among the prediction performance of the RF, ANN, and SVR models, with a best to worst order of RF, ANN, and SVR ($P < 0.01$).

Relative importance of the predictor variables

In this subsection, models are established for the five hydrological stations within Poyang Lake using the RF model, and five-fold cross-validation under scenario 3. As shown in Figure 8, for each station the previous lake water level (wl7) is the most important predictor for Poyang Lake, with a mean relative importance of 18%. Moreover, the effect of discharge from Hankou station is also considerable (mean relative importance at 12.4%), especially for Hukou and Xingzi water level stations in the Poyang Lake–Yangtze River waterway. This is mainly because the water level variation in the lake is contributed to by both the lake inflow and the Yangtze River, of which the blocking (even intrusion) and pulling effects of the Yangtze River on the outflows from Poyang Lake greatly influence the inter–intra water level fluctuations (Shankman *et al.* 2006; Dai *et al.* 2015). Moreover, the extent of the effect of backflow from the Yangtze River is mainly concentrated in the northern part of Poyang Lake (Cui *et al.* 2009). The remarkable effect of the Yangtze River discharge on the water level of Poyang Lake has also been reported by Li *et al.* (2015) and Jiang & Huang (1997). Additionally, for central and southern water level stations in the lake, important predictors were discharges from Hushan and Dufengkeng stations (Rao River) and discharges from Lijiadu station (Fu River) according to their mean values. This finding implies that these variables are the primary indicators representing the temporal variability of the daily water level. However, the relative importance of stations varies during different runs, especially for Lijiadu station from the Fu River (Figure 8), because the influence of the upstream tributaries and Yangtze River on lake water level variation may change during different seasons. For example, the water levels at Duchang and Xingzi are predominated by the Yangtze River discharge when the lake level is higher than 14.5 m, but the water level begins to be affected by catchment inflow when the lake level is lower than 14.5 m (Ye *et al.* 2014).

CONCLUSIONS

This study aimed to determine the most efficient model by comparing RF with SVR, ANN, and LM, and incorporating the time lag effect as well as previous hydrological status for forecasting the water level within lake stations.

Results demonstrated that for daily water level forecasting, the RF model can obtain more reliable and accurate forecasting results than ANN, SVR, and LM in terms of RMSE and R^2. The best forecasting performance was obtained by incorporating input data with 4-day time lag of discharge from catchment tributaries and the Yangtze River, as well as the water level over the previous week, with RMSEs of 0.25, 026, 0.28, 0.21, and 0.24 m for Hukou, Xingzi, Duchang, Tangyin, and Kangshan, respectively.

In addition, variable importance analysis was implemented for each water level station using the most accurate RF model and scenario 3. Results indicated that the previous water level was the most efficient predictor for water level forecasting. Moreover, the discharge from the Yangtze River also has a fundamental effect on water level variations.

Nevertheless, meteorological factors are not included in this study, thereby unavoidably introducing uncertainty to real-time water level forecasting. Future work should fully consider the complex hydrological and hydrodynamic processes of Poyang Lake.

ACKNOWLEDGEMENTS

This work was financially supported by the National Basic Research Program of China (973 Program) (Grant 2012CB417006) and the National Scientific Foundation of China (Grant 41271500), (Grant 41571107). Special thanks to Prof. Dr Qi Zhang of the Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences for the valuable suggestions.

REFERENCES

Altunkaynak, A. 2007 Forecasting surface water level fluctuations of Lake Van by artificial neural networks. *Water Resour. Manage.* **21** (2), 399–408.

Alvisi, S., Mascellani, G., Franchini, M. & Bardossy, A. 2006 Water level forecasting through fuzzy logic and artificial neural network approaches. *Hydrol. Earth Syst. Sci.* **10** (1), 1–17.

Aqil, M., Kita, I., Yano, A. & Nishiyama, S. 2007 Analysis and prediction of flow from local source in a river basin using a Neuro-fuzzy modeling tool. *J. Environ. Manage.* **85** (1), 215–223.

Bao, Y., Xiong, T. & Hu, Z. 2014 Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing* **129**, 482–493.

Belmans, C., Wesseling, J. & Feddes, R. 1983 Simulation model of the water balance of a cropped soil: SWATRE. *J. Hydrol.* **63** (3), 271–286.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Callegari, M., Mazzoli, P., de Gregorio, L., Notarnicola, C., Pasolli, L., Petitta, M. & Pistocchi, A. 2015 Seasonal river discharge forecasting using support vector regression: a case study in the Italian Alps. *Water* **7** (5), 2494–2515.

Chau, K. & Wu, C. 2010 A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *J. Hydroinform.* **12** (4), 458–473.

Chen, X., Chau, K. & Busari, A. 2015 A comparative study of population-based optimization algorithms for downstream river flow forecasting by a hybrid neural network model. *Eng. Appl. Artif. Intel.* **46**, 258–268.

Cui, L. J., Wu, G. F. & Liu, Y. L. 2009 Monitoring the impact of backflow and dredging on water clarity using MODIS images of Poyang Lake, China. *Hydrol. Process.* **23** (2), 342–350.

Dai, X., Wan, R. & Yang, G. 2015 Non-stationary water-level fluctuation in China's Poyang Lake and its interactions with Yangtze River. *J. Geogr. Sci.* **25** (3), 274–288.

Daliakopoulos, I. N., Coulibaly, P. & Tsanis, I. K. 2005 Groundwater level forecasting using artificial neural networks. *J. Hydrol.* **309** (1), 229–240.

Dawson, C. & Wilby, R. 2001 Hydrological modelling using artificial neural networks. *Prog. Phys. Geog.* **25** (1), 80–108.

El-Shafie, A. & Noureldin, A. 2010 Generalized versus non-generalized neural network model for multi-lead inflow forecasting at Aswan High Dam. *Hydrol. Earth Syst. Sci. Discuss.* **7** (5), 7957–7993.

Feng, C. X. J., Yu, Z. G. S., Kingi, U. & Baig, M. P. 2005 Threefold vs. fivefold cross validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data. *J. Manuf. Syst.* **24** (2), 93–107.

Francke, T., López-Tarazón, J. & Schroder, B. 2008 Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. *Hydrol. Process.* **22** (25), 4892–4904.

Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. 2010 Variable selection using random forests. *Pattern Recogn. Lett.* **31** (14), 2225–2236.

Gholami, V., Chau, K., Fadaee, F., Torkaman, J. & Ghaffari, A. 2015 Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. *J. Hydrol.* **529**, 1060–1069.

Ghorbani, M. A., Khatibi, R., Aytek, A., Makarynskyy, O. & Shiri, J. 2010 Sea water level forecasting using genetic programming and comparing the performance with artificial neural networks. *Comput. Geosci.* **36** (5), 620–627.

Guo, H., Hu, Q. & Jiang, T. 2008 Annual and seasonal streamflow responses to climate and land-cover changes in the Poyang Lake basin, China. *J. Hydrol.* **355** (1), 106–122.

Guo, H., Hu, Q. I., Zhang, Q. I. & Wang, Y. 2012 Annual variations in climatic and hydrological processes and related flood and drought occurrences in the Poyang Lake Basin. *Acta Geographica Sinica* **67** (5), 699–709 (in Chinese).

Hipni, A., El-shafie, A., Najah, A., Karim, O. A., Hussain, A. & Mukhlisin, M. 2013 Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). *Water Resour. Manage.* **27** (10), 3803–3823.

Hsu, K. L., Gupta, H. V. & Sorooshian, S. 1995 Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **31** (10), 2517–2530.

Hu, W., Zhai, S., Zhu, Z. & Han, H. 2008 Impacts of the Yangtze River water transfer on the restoration of Lake Taihu. *Ecol. Eng.* **34** (1), 30–49.

Huang, C.-L., Chen, M.-C. & Wang, C.-J. 2007 Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.* **33** (4), 847–856.

Huang, J., Gao, J. & Zhang, Y. 2015 Combination of artificial neural network and clustering techniques for predicting phytoplankton biomass of Lake Poyang, China. *Limnology* **16** (3), 179–191.

Jiang, J. & Huang, Q. 1997 A study of the impact of the three Gorges Project on the water-level of Poyang Lake. *J. Natural Resour.* **12** (3), 219–224 (in Chinese).

Kecman, V. 2001 *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models (Complex Adaptive Systems)*. MIT Press, Cambridge, MA, USA.

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

Kirchner, J. W. 2006 Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resour. Res.* **42** (3).

Kourgialas, N. N., Dokou, Z. & Karatzas, G. P. 2015 Statistical analysis and ANN modeling for predicting hydrological extremes under climate change scenarios: the example of a small Mediterranean agro-watershed. *J. Environ. Manage.* **154**, 86–101.

Kumar, D. N. & Maity, R. 2008 Bayesian dynamic modelling for nonstationary hydroclimatic time series forecasting along with uncertainty quantification. *Hydrol. Process.* **22** (17), 3488–3499.

Lai, X., Jiang, J., Liang, Q. & Huang, Q. 2013 Large-scale hydrodynamic modeling of the middle Yangtze River Basin with complex river–lake interactions. *J. Hydrol.* **492**, 228–243.

Lan, Y. 2014 Forecasting performance of support vector machine for the Poyang Lake's water level. *Water Sci. Technol.* **70** (9), 1488–1495.

Li, X. & Zhang, Q. 2015 Variation of floods characteristics and their responses to climate and human activities in Poyang Lake, China. *Chinese Geogr. Sci.* **25** (1), 13–25.

Li, Y., Zhang, Q., Yao, J., Werner, A. D. & Li, X. 2013 Hydrodynamic and hydrological modeling of the Poyang Lake catchment system in China. *J. Hydrol. Eng.* **19** (3), 607–616.

Li, Y., Zhang, Q., Werner, A. & Yao, J. 2015 Investigating a complex lake–catchment–river system using artificial neural networks: Poyang Lake (China). *Hydrol. Res.* **46** (6), 912–928.

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.

Malekipirbazari, M. & Aksakalli, V. 2015 Risk assessment in social lending via random forests. *Expert Syst. Appl.* **42** (10), 4621–4631.

Mustafa, M., Isa, M. & Rezaur, R. 2012 Artificial neural networks modeling in water resources engineering: infrastructure and applications. *Int. J. Soc. Human Sci.* **62**, 341–349.

Panagoulia, D. 2006 Artificial neural networks and high and low flows in various climate regimes. *Hydrolog. Sci. J.* **51** (4), 563–587.

Polikar, R., Zhang, C. & Ma, Y. 2012 *Ensemble Machine Learning: Methods and Applications*. Springer, Heidelberg, Germany.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. & Chica-Rivas, M. 2015 Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews* **71**, 804–818.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. 1985 *Learning Internal Representations by Error Propagation*. Technical report. DTIC Document.

Shankman, D., Keim, B. D. & Song, J. 2006 Flood frequency in China's Poyang Lake region: trends and teleconnections. *Int. J. Climatol.* **26** (9), 1255–1266.

Smola, A. J. & Schölkopf, B. 2004 A tutorial on support vector regression. *Stat. Comput.* **14** (3), 199–222.

Sulaiman, M., El-Shafie, A., Karim, O. & Basri, H. 2011 Improved water level forecasting performance by using optimal steepness coefficients in an artificial neural network. *Water Resour. Manage.* **25** (10), 2525–2541.

Team, R. C. 2014 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

Trichakis, I. C., Nikolos, I. K. & Karatzas, G. 2011 Artificial neural network (ANN) based modeling for karstic groundwater level simulation. *Water Resour. Manage.* **25** (4), 1143–1152.

Vapnik, V. 2013 *The Nature of Statistical Learning Theory*. Springer Science & Business Media, New York, USA.

Wantzen, K. M., Rothhaupt, K.-O., Mörtl, M., Cantonati, M., László, G. & Fischer, P. 2008 Ecological effects of water-level fluctuations in lakes: an urgent issue. *Hydrobiologia* **613** (1), 1–4.

Wei, C.-C. 2012 Wavelet kernel support vector machines forecasting techniques: case study on water-level predictions during typhoons. *Expert Syst. Appl.* **39** (5), 5189–5199.

Were, K., Bui, D. T., Dick, Ø. B. & Singh, B. R. 2015 A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol. Indic.* **52**, 394–403.

Ye, X., Li, Y., Li, X. & Zhang, Q. 2014 Factors influencing water level changes in China's largest freshwater lake, Poyang Lake, in the past 50 years. *Water Int.* **39** (7), 983–999.

Yoon, H., Jun, S.-C., Hyun, Y., Bae, G.-O. & Lee, K.-K. 2011 A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *J. Hydrol.* **396** (1), 128–138.

First received 17 December 2015; accepted in revised form 13 June 2016. Available online 26 July 2016