water

MDPI

Article

Improving Lake Level Prediction by Embedding Support Vector Regression in a Data Assimilation Framework

Kang Wang [1], Tengfei Hu [2,*], Peipei Zhang [1], Wenqin Huang [1,3], Jingqiao Mao [1], Yifan Xu [2] and Yong Shi [2]

[1] College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China
[2] Nanjing Hydraulic Research Institute, Nanjing 210029, China
[3] Institute of Bio- and Geosciences: Agrosphere (IBG-3), Forschungszentrum Jülich, 52428 Jülich, Germany
[*] Correspondence: tfhu@nhri.cn

Abstract: Data-driven models are widely used in the field of water level prediction due to their generalizability and predictive abilities. In long-series prediction, however, data-driven models degrade rapidly due to the uncertainty and constraints of model data and parameters. To address the problem of inaccurate continuous water level prediction, this study introduced a data assimilation technique, the unscented Kalman filter (UKF), and embedded support vector regression (SVR) into the framework and applied it to Dongting Lake, the second largest freshwater lake in China. The results demonstrated that the assimilation model is significantly better than the non-assimilation model in predicting water levels and is not affected by the characteristics of lake level changes, with the R^2 increasing from 0.975–0.982 to 0.998–0.999 and the RMSE decreasing from 0.436–0.159 m to 0.105–0.042 m. The prediction lead time also increased with the increase of continuous assimilation data. Further analysis of the assimilation model showed that when there was an assimilation cycle, the prediction remained stable for successive sets of two or more assimilated data, and the prediction lead time increased with successive assimilated data, from 4–8 days (one successive assimilation data) to 9–12 days (five successive assimilation data). Overall, this study found that the data assimilation framework can improve the prediction ability of data-driven models, with assimilated models having a smaller fluctuation range and higher degree of concentration than non-assimilated models. The increase in assimilated data will improve model accuracy as well as the number of days of prediction lead time when an assimilation cycle exists.

Keywords: lake level prediction; data simulation; data-driven model; unscented Kalman filter

Citation: Wang, K.; Hu, T.; Zhang, P.; Huang, W.; Mao, J.; Xu, Y.; Shi, Y. Improving Lake Level Prediction by Embedding Support Vector Regression in a Data Assimilation Framework. *Water* **2022**, *14*, 3718. https://doi.org/10.3390/w14223718

Academic Editor: Leonardo V. Noto

Received: 10 October 2022
Accepted: 15 November 2022
Published: 17 November 2022

1. Introduction

In the last two decades, data-driven techniques have proven to be effective and robust tools for modeling and prediction of various water resource variables [1,2]. These techniques can generally be viewed as universal approximators that identify and generalize the input–output relationship based on the limited information fed to them. Compared to physically based models, data-driven models are relatively pragmatic as they do not require a full understanding of the underlying physics, detailed topographical data and long computational time for calibration and application [3,4]. Data-driven techniques commonly used by the hydrological community include different forms of artificial neural networks (ANNs) and support vector regression (SVR) [5,6]. SVR based on statistical learning theory has been considered a better choice than ANNs for lake level prediction, e.g., as noted in refs. [7,8]. The primary reason is that SVR has an advantage in terms of generalization capability due to the use of structural risk minimization [9]. In addition, SVR has fewer free parameters to estimate than most ANNs.

As suggested by Maier and Dandy [10], different sources of uncertainties should be taken into account in developing data-driven water resource models. In data-driven modeling and forecasting, there are three primary categories of uncertainty: structural, parametric, and data uncertainty. In the case of model structure, an insufficient degree of freedom of the approximator used can lead to potentially large prediction errors [11,12]. Moreover, the time lagged values of input variables must be selected with care. The inclusion of irrelevant or redundant inputs only adds noise to the model and increases the dimension of the problem, while the omission of relevant inputs most likely makes the model unable to describe the system behavior [13]. Since the parameters of data-driven models have no physical interpretation, they need to be estimated from the training samples; this means that different training data sets could probably lead to different parameter estimates [14,15]. In addition, the training samples must be representative of the entire data, in case the approximator is required to extrapolate beyond the range of the training data [16]. Parameter estimation can be performed by applying optimization techniques. However, even though powerful global optimization techniques are within easy reach, they still cannot guarantee globally optimal parameters [17]. In some situations, different parameter sets result in a similar model performance (i.e., the equifinality problem), and it is impossible to identify the best set among them. Furthermore, a data-driven model can be overtrained and thus capture not only the desired input-output mapping but the noise contained in the training data (i.e., the overfitting problem). Apart from model structural and parameter uncertainties, the imperfect data used to calibrate and validate the model can also introduce some degree of uncertainty. First, observations of the system input (forcing) and output are always noise corrupted. Second, the field observation within a limited period of time cannot obtain complete data about the physical system.

The model calibration procedure normally attempts to find the "optimal" estimates of the parameters to fit the model to the data over a long period of time. This procedure, however, attributes all prediction uncertainty to parameter uncertainty but largely ignores the uncertainties associated with model structure and measurements of the system [18]. In such a case, the calibrated model may not have satisfactory performance in the prediction stage [4].

Due to the unpredictability and limitations of data-driven models, their performance in long-term forecasting degrades rapidly in comparison to that of traditional physical models [16]. In meteorological, atmospheric, and hydrological sciences, data assimilation (DA) approaches have been widely used to reduce prediction uncertainty and make optimal predictions [19]. DA refers to the integration of all available information (i.e., dynamic model predictions and observed data) to produce the best estimates of system states or model parameters [20]. Instead of using all the observations simultaneously, sequential DA methods have been developed to recursively update the probability distribution functions (PDFs) of different quantities of interest each time a new observation becomes available [18]. In this manner the model is optimally initialized for a new prediction, meaning that the model's predictive skill can be improved, especially under real-time conditions [21]. Sequential DA adopts Bayes' theorem that the posterior PDFs of various quantities can be inferred by conditioning on the corresponding observations [22]. It represents a general framework, where different sources of uncertainties can all be addressed. A deterministic data-driven model (e.g., SVR lake level model) can be embedded in the sequential DA framework, with model and observation uncertainties described as random noises [4]. Such a framework is an improvement on the conventional calibration procedure that searches for a single set of model parameters, in the sense that the uncertainties in the model and observations can be explicitly accounted for [17].

It is acknowledged that incorporating data-driven models into the DA framework can increase the predictive capacity of the models in applications and greatly delay their rate of deterioration. This paper aims to exploit the potential of combining SVR with an established sequential DA method, the unscented Kalman filter (UKF), in the prediction of lake water level. The hybrid method, SVR+UKF, is applied to the Dongting Lake, the second largest freshwater lake in China. The SVR water level prediction models for each of the three representative stations in Dongting Lake were embedded into the DA framework and applied to the case. The measured water level at the lake hydrological stations were set to be continuously assimilated to the water level prediction models. This was done by predicted water level at one time step using SVR for each of three stations with inputs of remote river discharges and local water levels at the preceding time steps. The lake level observations are assimilated into SVR models using UKF at every time step to study the improvement of the models' performance over SVR without DA. Given that the lake levels might not be continuously available in practice, the impact of different assimilation/non-assimilation cycles on the models' prediction accuracy is also investigated. The investigation can provide useful information about the acceptable lead time of the lake level prediction. The combination of data-driven models with DA approaches for lake water level forecasting has not, to our knowledge, been reported in the literature. Therefore, this study contributes to our understanding of the hydrological uses of this hybrid technique.

2. Materials and Methods

2.1. Study Area and Data Collection

Dongting Lake is located in the middle reaches of the Yangtze River in China, it is the second largest freshwater lake in China and one of the two major river-connected lakes in the Yangtze River Basin. Dongting Lake receives water from four major tributaries, namely the Xiang River, Zi River, Yuan River, and Li River, and the incoming water flows into the Yangtze River at Chenglingji after being impounded by the lake, creating a complicated confluence of lakes and rivers in the middle sections of the Yangtze River. It plays a significant part in Yangtze River flood storage: when the water level of the Jing River is low, the water of Dongting Lake will flow into the Yangtze River, resulting in the "dry effect"; when the water level of the Jing River is high (mainly occurring in the flood season), the outlet of Dongting Lake is supported by the Yangtze River, and the drainage of the lake area is restricted, resulting in the "block effect". Therefore, water level prediction research in Dongting Lake is crucial for preventing extremely low water level in the lake, ensuring the safety of the lake's water supply and enhancing Dongting Lake's ecological environment. This study focuses on the daily water level prediction at three representative hydrological stations in Dongting Lake: Chenglingji, Yingtian, and Xiaohezui. The study area is shown in Figure 1.

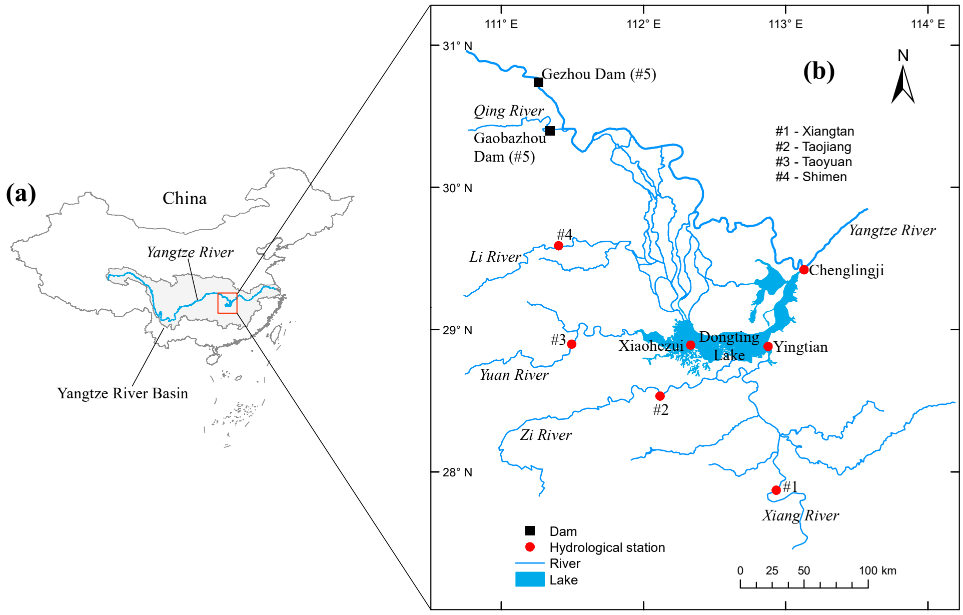

Figure 1. Map of the study area, showing (**a**) its location and (**b**) the Yangtze River–Dongting Lake water system. #1—Xiangtan, #2—Taojiang, #3—Taoyuan, #4—Shimen, #5—Gezhou Dam.

Utilizing physical simulation models as prediction tools requires the calibration of numerous parameters. An alternative way would be a robust data-driven model based on evaluating the data about a system and determining links between the system state variables (input, internal, and output variables) without explicit knowledge of the physical behavior of the system. In order to simulate the water level of a lake connected to the river, historical data including inflow, outflow and pre-lake water level are necessary for the development of a data-based model.

All data-driven models must be trained with acquired data before being evaluated for simulation. The data used for water level prediction in this study are the average daily flows of the main tributaries of Dongting Lake, including the Xiang River, Zi River, Yuan River, and Li River, the average daily flow of the Yangtze River (the sum of the outflow from Gezhou Dam and Gaobazhou Dam), and the daily 8:00 water levels at the five hydrological stations in Dongting Lake. Table 1 describes the data characteristics of the water levels in Dongting Lake, and Figure 2 depicts the time series of river flows. According to the graph, the hydrological data obtained spans four years (2009–2012). Data from 2010 and 2012 (731 observations in total) had higher flood flows and were utilized to train the model, while the data from 2009 and 2011 were used to evaluate the performance of the data assimilation technique; such a data spilt approach enhances the generalization of the model.

Table 1. Statistical characteristics of Dongting Lake water levels.

Station	Dataset [a]	Minimum Value (m)	Maximum Value (m)	Mean Value (m)	Standard Deviation (m)
Chenglingji	Training	20.21	33.40	25.66	3.95
	Testing	20.43	30.86	24.13	2.94
Yingtian	Training	21.21	33.67	26.69	3.66
	Testing	21.32	31.15	25.05	2.75
Xiaohezui	Training	27.89	34.93	29.99	1.68
	Testing	27.91	31.91	29.27	1.02

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

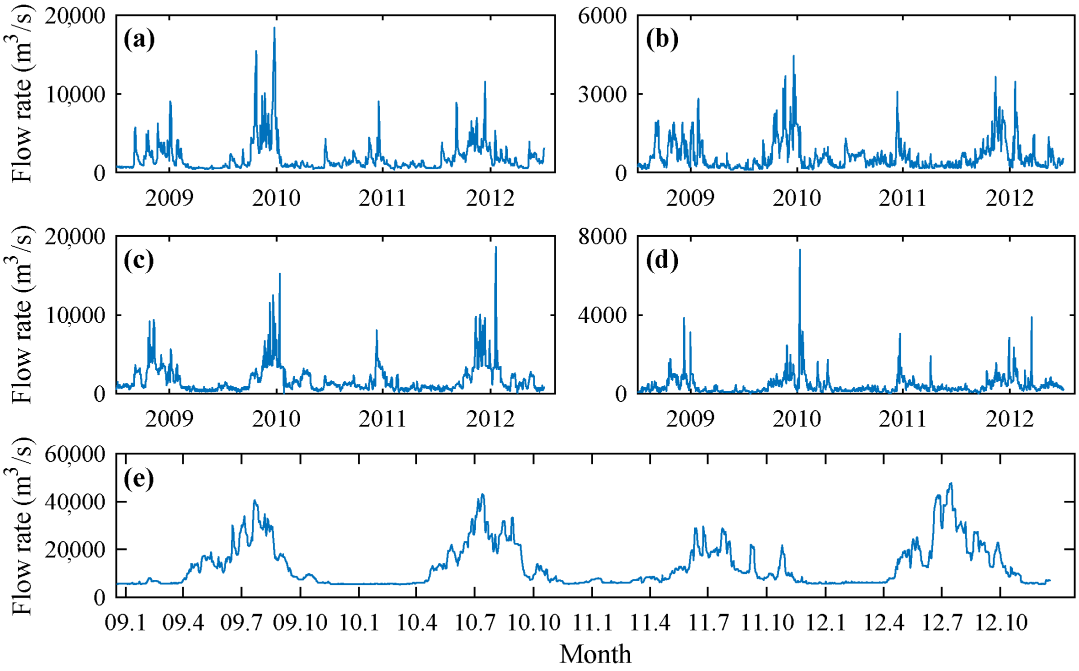

Figure 2. Time series of daily discharges of the (**a**) Xiang River, (**b**) Zi River, (**c**) Yuan River, (**d**) Li River, and (**e**) Yangtze River.

2.2. Problem Formulation

Support vector regression-based water level simulation predictions involve multiple sources of uncertainty, including model structure uncertainty, parameter uncertainty, and data uncertainty. Sources of uncertainty in data-driven modelling and prediction are shown in Table 2. To reduce these uncertainties and achieve data assimilation, a model that evolves over time is required. In discrete time, a state-space model is characterized by a formulation (or representation) where the model's state at time *t* is completely determined by its state at the preceding time *t*–1 and (the effect of) external system forcings.

Table 2. Sources of uncertainties in data-driven modeling and prediction.

Type	Source
Structural uncertainty	Insufficient degree of freedom of the approximator used Unreasonable input variable selection
Parameter uncertainty	Dependency of parameter values on data division Absence of representativeness of training samples Difficulty in finding globally optimal parameters Equifinality problem Overfitting problem
Data uncertainty	Input and output measurement noise Lack of representativeness

To deal with the time-delayed states, the extended state vector is taken, and the dynamic equation of the extended state vector is used in the unscented Kalman filter (UKF). The state of the system (water level) at time *t* depends on the state of the system at one or more previous time steps as well as the present or preceding forcing of the system, where time steps are in days and *t* is the discretized time.

The water level at a hydrological station can be described as follows:

$$\begin{aligned} L_t = f_{\mathrm{SVR}}(&D^1_{t-m_1}, D^1_{t-m_1-1}, \ldots, D^1_{t-n_1}, \ldots, \\ &D^N_{t-m_N}, D^N_{t-m_N-1}, \ldots, D^N_{t-n_N}, L_{t-m_0}, L_{t-m_0-1}, \ldots, L_{t-n_0}) \end{aligned} \tag{1}$$

where $f_{\mathrm{SVR}}(\bullet)$ is a mapping based on support vector regression; L_t is the water level at the site on day t; $D^i_{t\text{–}j}$ (i = 1, . . . , N; j = m_i, . . . , n_i) is the measured flow at river site #i on day $t-j$; $L_{t\text{–}j}$ (j = m_0, . . . , n_0) is the water level at the site on day $t-j$.

$$L_t = f_{\mathrm{SVR}}(L_{t-1}, L_{t-2}, \ldots, L_{t-n_0}, \mathbf{D}_{t-1}) \tag{2}$$

where

$$\mathbf{D}_{t-1} = \left(D^1_{t-m_1}, D^1_{t-m_1-1}, \ldots, D^1_{t-n_1}, \ldots, D^N_{t-m_N}, D^N_{t-m_N-1}, \ldots, D^N_{t-n_N}\right),$$

the corresponding water level dynamics are modelled as

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}, \mathbf{D}_{t-1}) + \mathbf{q}_{t-1} \tag{3}$$

$$\begin{bmatrix} L_t \\ L_{t-1} \\ \vdots \\ L_{t-n_0+1} \end{bmatrix} = \begin{bmatrix} f_{\mathrm{SVR}}(L_{t-1}, L_{t-2}, \ldots, L_{t-n_0}, \mathbf{D}_{t-1}) \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \bullet \begin{bmatrix} L_{t-1} \\ \vdots \\ L_{t-n_0} \end{bmatrix} \end{bmatrix} + \begin{bmatrix} q_{t-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4}$$

the corresponding measurement model is

$$y_t = [1\ 0\ \cdots\ 0]\mathbf{x}_t + r_t \tag{5}$$

where $\mathbf{x}_t \in \Re^n$ is the state of the system at time step t, $\mathbf{q}_{t-1}$ is the process noise at time step t—1, r_t is the measurement noise at time step t.

In the following study, the SVR model is embedded in the unscented Kalman filter (UKF) to improve the efficiency of the assimilation simulation at each step, starting with a brief introduction to the SVR and UKF methods; the flow chart of the study is shown in Figure 3.

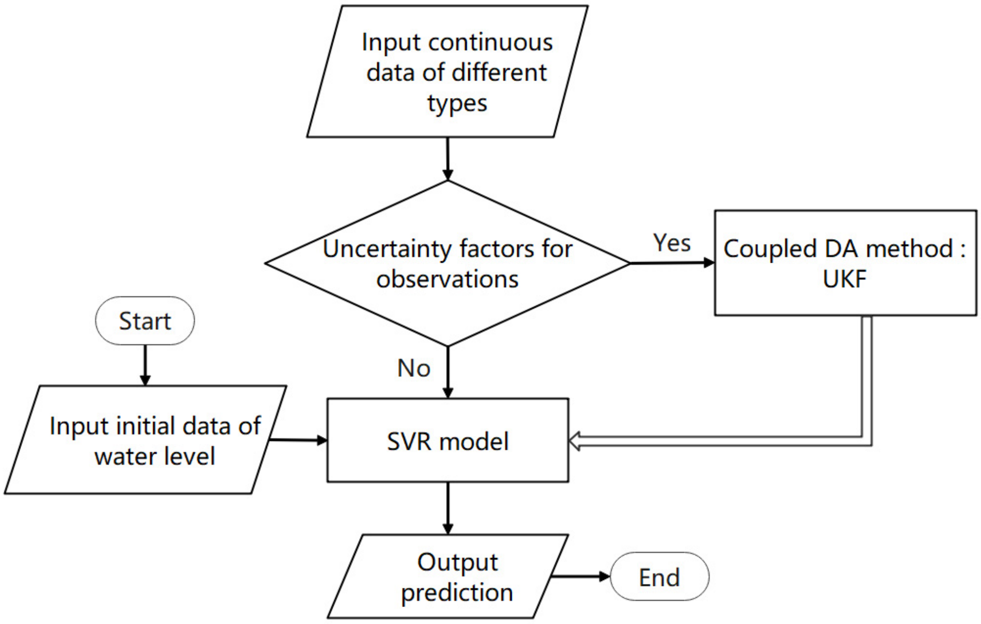
Input continuous data of different types
Uncertainty factors for observations
Yes
Coupled DA method : UKF
Start
No
Input initial data of water level
SVR model
Output prediction
End

Figure 3. Study flow chart.

2.3. Support Vector Regression

Support vector machine regression (SVR) is a model for the application of support vector machines (SVMs) to regression problems [23]. Let vector (x_j, y_j), $j\epsilon\{1, 2, 3, \ldots, N\}$ represent the set of observed input and output data. The aim of a data-based model such as SVR is, in general, to search for a function $f(x)$ as an approximation of the value y_i with minimum risk using only the available independent and identically distributed data. In the SVR algorithm, the estimation function is determined by a small subset of training samples, namely support vectors. In this algorithm, a specific loss function called ε-insensitive loss is also developed to create a sparseness property for SVR. This means that instead of reducing the empirical error across the training data, SVR minimizes a controlled risk function, which states that in order to achieve the least risk, simultaneous management of the model's complexity and the error due to training data is required (principle of the structural risk minimization theory). Using SVR for regression analysis, we need to find a hyperplane; the residual of the data points defining the area within the target is 0, and the distance from the data points (support vectors) outside the region to the boundary is the residual (ζ). Similar to the linear model, we want these residuals (ζ) to be minimal. Therefore, in general, SVR involves identifying the best band region (2ε width) and then regressing the points outside of that region, which was shown in Figure 4

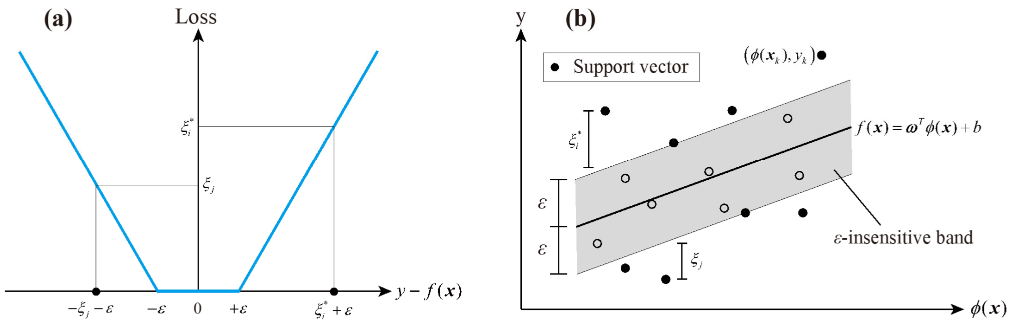

Figure 4. Illustration of (**a**) ε-insensitive loss function and (**b**) support vector regression in the feature space.

2.4. Unscented Kalman Filter

The Kalman filter (KF) is a well-known data assimilation scheme originally introduced by Kalman for linear systems [24]. The classical Kalman filter deals with state estimation for linear processes by updating model predictions using measurements and consists of two steps: prediction and update.

In the prediction step, the KF algorithm uses linear dynamics to predict the state at the current time step based on the state estimate from a previous time. In the update step, the predicted state is refined through weighted combination with observations based on relative errors. This optimal state is then used to advance to the next step, and so on.

In practice, the Kalman filter is a commonly applied procedure for sequential data assimilation and has been successfully used to assimilate observations into existing models. Significant contributions to applications and improvements of KF have been made by many researchers from various backgrounds, e.g., environmental sciences, oceanography, and meteorology [13,18,19]. Variations on the original KF algorithm have made it suitable for use with non-linear problems such as hydrological modelling, leading to the extended Kalman filter (EKF) [22,25].

In the EKF algorithm, local (tangent linear) approximation of the nonlinear state and measurement equations is performed each time data assimilation is conducted. However, the EKF may produce instability or even divergence due to the neglect of the second and higher order derivatives of the model. Evensen introduced the ensemble Kalman filtering (EnKF) algorithm as an alternative to the EKF to address difficulties arising from high-dimensional nonlinear filtering problems [26]. The applicability to nonlinear problems and easy implementation of the EnKF method has led to extensive applications of this DA technique in hydrology, meteorology, and other fields [17,21].

In the EnKF method, the Kalman gain matrix is calculated from the error covariances provided by an ensemble of possible model states that are propagated according to the same deterministic water quality model. EnkF is suboptimal, i.e., it doesn't lead to the optimal state estimate. However, it has the advantage that it is easy to handle large problems and complex models. The EnKF has the following advantages compared to the KF and EKF: (1) implicit propagation of the state error covariance, making it suitable for large-scale problems; (2) does not require model linearization; (3) a limited number of model states are used and convergence is much faster.

There is another extension of the KF that is widely known: the unscented Kalman filter (UKF) [27]. The UKF has been developed to overcome the deficiencies of the linearization in the EKF. It provides a direct and explicit mechanism to transform mean and covariance information and has been previously shown to be a superior tool to EKF in various aspects, especially in strongly nonlinear systems [28].

The unscented transformation (UT) is a method for calculating the statistics of a random variable that undergoes a nonlinear transformation. Like the Taylor series-based approximation, the unscented transform can be used to form a Gaussian approximation to the joint distribution of the random variables **x** and **y**. One advantage of the unscented transform over the Taylor series approximation is that the former is able to better capture the higher order information of the non-linear transformation. In addition, the unscented transformation does not require the calculation of Jacobi or Hessian matrices.

The prediction and update schematic for the UKF is shown in Figure 5, where the prediction steps calculate the predicted state means and covariances:

$$\mathbf{X}_{k-1} = [\mathbf{m}_{k-1} \ \cdots \ \mathbf{m}_{k-1}] + \sqrt{c}\left[0 \ \sqrt{\mathbf{P}_{k-1}} - \sqrt{\mathbf{P}_{k-1}}\right] \tag{6}$$

$$\hat{\mathbf{X}}_k = \mathbf{f}(\mathbf{X}_{k-1}, k-1) \tag{7}$$

$$\mathbf{m}_k^- = \hat{\mathbf{X}}_k \mathbf{w}_m \tag{8}$$

$$\mathbf{P}_k^- = \hat{\mathbf{X}}_k \mathbf{W}\left[\hat{\mathbf{X}}_k\right]^T + \mathbf{Q}_{k-1} \tag{9}$$

where $\mathbf{X}_k$ is the matrix formed by the sigma points, $\mathbf{m}_k$ and $\mathbf{P}_k$ are the mean and covariance of the state estimated at time step k (after obtaining the measurement information); $\mathbf{w}_m$ and $\mathbf{W}$ are process vectors and matrices, respectively; $\mathbf{Q}_{k-1}$ is the process noise at time step k–1.

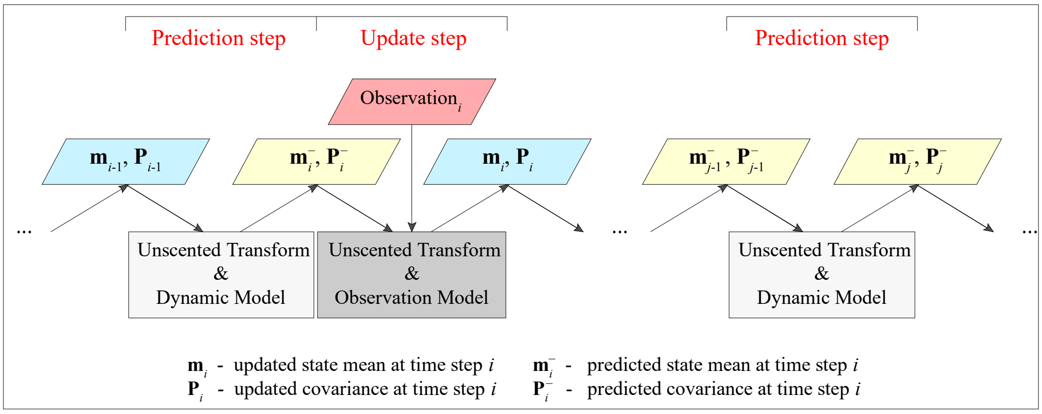
Prediction step
Update step
Prediction step
Observation$_i$
$\mathbf{m}_{i\text{-}1}$, $\mathbf{P}_{i\text{-}1}$
$\mathbf{m}_i^-$, $\mathbf{P}_i^-$
$\mathbf{m}_i$, $\mathbf{P}_i$
$\mathbf{m}_{j\text{-}1}^-$, $\mathbf{P}_{j\text{-}1}^-$
$\mathbf{m}_j^-$, $\mathbf{P}_j^-$
...
Unscented Transform & Dynamic Model
Unscented Transform & Observation Model
...
Unscented Transform & Dynamic Model
...
$\mathbf{m}_i$ - updated state mean at time step i
$\mathbf{P}_i$ - updated covariance at time step i
$\mathbf{m}_i^-$ - predicted state mean at time step i
$\mathbf{P}_i^-$ - predicted covariance at time step i

Figure 5. Schematic diagram of the unscented Kalman filter.

The UKF addresses nonlinear problems by using a deterministic sampling approach. The state distribution is approximated by Gaussian Random Variables (GRVs) but is now represented using a minimal set of carefully chosen sample points. These sample points completely capture the true mean and covariance of the GRV and, when propagated through the true nonlinear system, capture the posterior mean and covariance accurately to the third order (Taylor series expansion) for any nonlinearity. The EKF, in contrast, only achieves first-order accuracy. Remarkably, the computational complexity of the UKF is the same order as that of the EKF. The EKF can be viewed as providing "first-order" approximations to the optimal terms. These approximations, however, can introduce large errors in the true posterior mean and covariance of the transformed (Gaussian) random variable, which may lead to sub-optimal performance and sometimes divergence of the filter.

Comparing the EnKF and UKF, it has been found that the former has certain limitations in practice: (1) the EnKF requires the existence of a Jacobi or Hessian matrix, which may not be satisfied in practical problems; (2) in many cases, the Jacobi or Hessian matrix is very difficult to compute and is error-prone. Due to these shortcomings of the EnKF, the UKF has weaker assumptions than the EnKF and achieves better theoretical performance. Based on the above considerations, this study uses the UKF for the assimilation prediction of water level data in Dongting Lake.

3. Results and Discussion

In this study, the UKF was employed as the underlying model, and the Dongting Lake water level data from 2010 and 2012 was chosen as the training set for simultaneous parameter optimization. The performance tests with the SVR embedded in the data assimilation framework therefore used data from the test period, i.e., from 2009 and 2011. The water level prediction models for each of the three representative stations in Dongting Lake were embedded into the data assimilation framework and applied to both years. Additionally, the measured water levels at the lake stations were configured to be continuously assimilated into the water level prediction models.

3.1. Data Assimilation Model Prediction Results

To reflect the improved performance of the Dongting Lake water level prediction model under the data assimilation framework, the predicted output of the data assimilation model (labeled SVR+UKF) was compared with the continuous water level prediction results (labeled SVR) using the model predictions replacing the next one-step measured values for a continuous period of one year. The comparisons for the three representative stations in Lake Dongting are shown in Figures 6–8.

From Figures 6–8, the results of the support vector regression's continuous prediction of water level at each station in Dongting Lake for a continuous period of one year correspond well with the measured values. Based on the extent of divergence from the projected water level readings, the Xiaohezui model has the best prediction ability, followed by the Chenglingji model, while the Yingtian station has the largest prediction error. The performance of the non-assimilative model for the continuous prediction of Dongting Lake water levels fluctuated throughout the year. Generally, the accuracy of the non-assimilative model is better in the low-flow period and the rising and receding water level periods, while the model accuracy is slightly worse in the high-flow period. As can be seen from the scatter plots in Figures 6–8, the predictions from the non-assimilation model are rarely systematically over- or underestimated and deviate 'stochastically' from the measured values across the range of water levels.

The assimilation model achieves better prediction accuracy than the non-assimilation model, and the performance improvement is significant. The water levels predicted by the assimilation model are close to the 100% coincidence line, with minimal deviation. Among the assimilation models for various stations, the Xiaohezui model has the most accurate predictions, followed by the Chenglingji model, and the Yingtian model has the least accurate predictions. The above model performance ranking findings are comparable to the non-assimilation model ranking results, indicating that the higher the accuracy of the dynamic models in the data assimilation framework, the more accurate the output of the data assimilation predictions. The prediction accuracy of all the assimilation models did not vary significantly within the year, indicating that the characteristics of lake level changes during the prediction period had a relatively weak effect on the performance of the assimilation models, which could provide stable predictions of the Dongting Lake water level at different times of the year.

Multiple performance indicators were chosen to assess the accuracy of the assimilated and non-assimilated Dongting Lake water level prediction models, including the coefficient of determination R^2, the root mean square error (RMSE), and the mean relative error (MRE):

$$R^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{11}$$

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{12}$$

where y_i is the observed value; $\hat{y}_i$ is the predicted value; $\overline{y}$ is the average of the observed values; and n is the number of observation records.

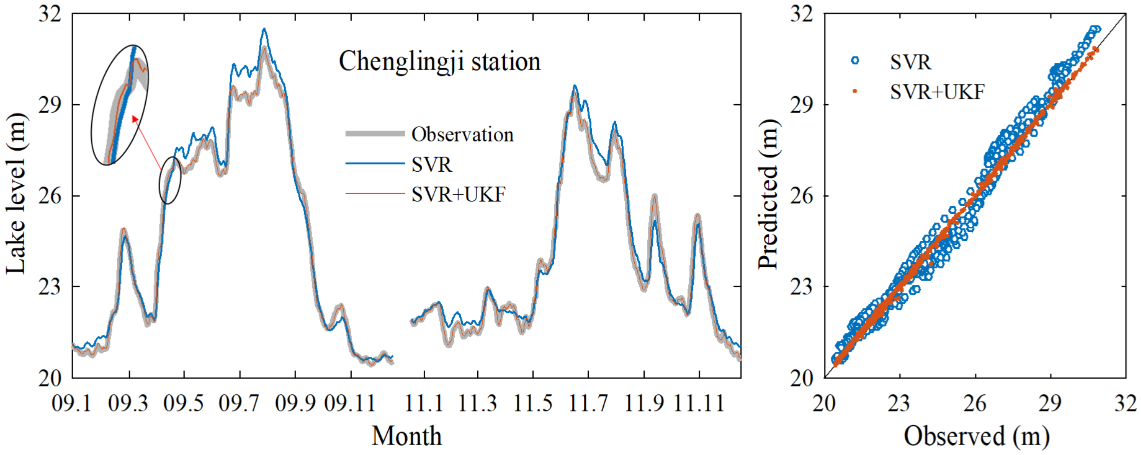

Figure 6. Comparisons between Chenglingji water levels predicted with SVR and SVR+UKF.

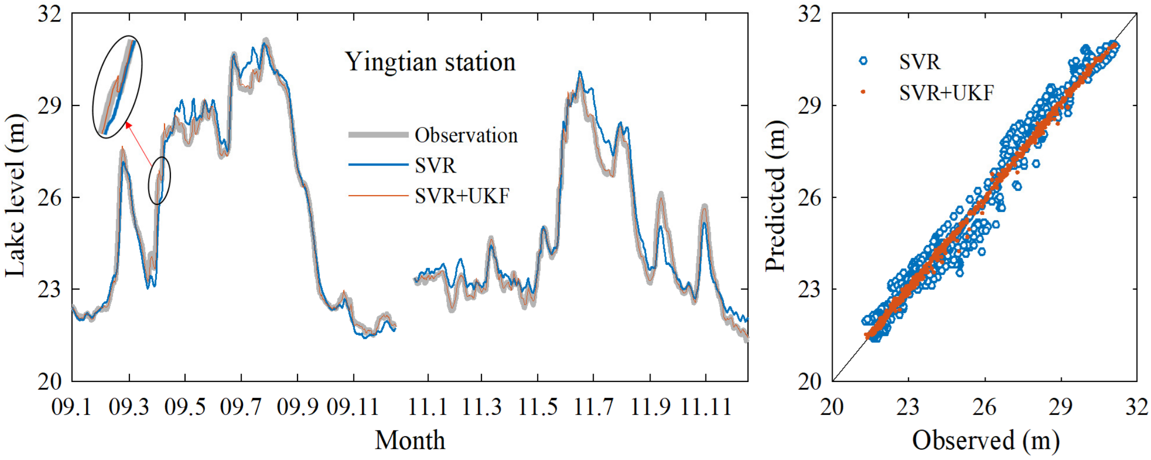

Figure 7. Comparisons between Yingtian water levels predicted with SVR and SVR+UKF.

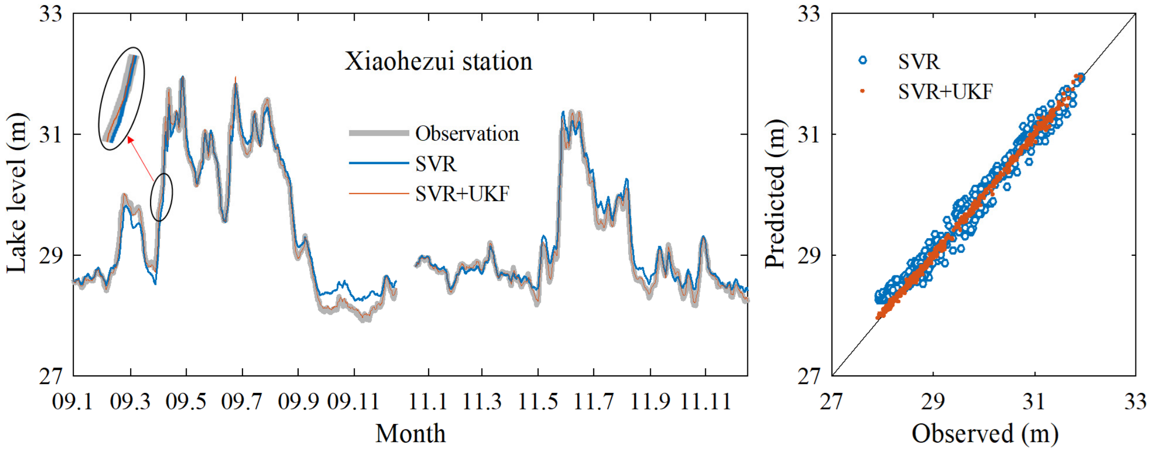

Figure 8. Comparisons between Xiaohezui water levels predicted with SVR and SVR+UKF.

To ensure that all variables receive equal weights during the training process, it is necessary to normalize the raw data. All the data will be normalized between 0 and 1, so that all the parameters could be assigned equal weights. One popular approach is to assume that the errors are zero-mean white noise sequences with a normal (i.e., Gaussian) probability distribution. In addition, it is typically assumed that the model error and observation error are uncorrelated in order to obtain optimal estimates.

Table 3 presents the performance assessment results of the assimilated and non-assimilated Dongting Lake water level prediction models. It can be seen that the prediction accuracy of the non-assimilated model is in an acceptable range, with the coefficient of determination R^2 varying from 0.975 (Yingtian) to 0.982 (Chenglingji) at different stations, the RMSE varying from 0.159 m (Xiaohezui) to 0.436 m (Yingtian), and the MRE in the range of 0.004 (Xiaohezui) to 0.013 (Yingtian). Among the three representative stations in Dongting Lake, Xiaohezui has the highest prediction accuracy, followed by Chenglingji, and Yingtian has the lowest, which is comparable to the error distribution of predictions obtained using the SVR model alone.

Table 3. Performance comparisons between SVR and SVR+UKF.

Station	SVR			SVR+UKF		
	R^2	RMSE (m)	MRE	R^2	RMSE (m)	MRE
Chenglingji	0.982	0.395	0.012	0.999	0.068	0.002
Yingtian	0.975	0.436	0.013	0.999	0.105	0.003
Xiaohezui	0.976	0.159	0.004	0.998	0.042	0.001

The model prediction performance was significantly improved when the actual water level measurements at each station were continuously assimilated into the Dongting Lake water level prediction model using the UKF. According to Table 3, the coefficient of determination R^2 varies from 0.998 (Xiaohezui) to 0.999 (Chenglingji and Yingtian), the RMSE varies from 0.042 m (Xiaohezui) to 0.105 m (Yingtian), and the MRE varies from 0.001 (Xiaohezui) to 0.003 (Yingtian). The performance ranking results of the assimilation models for the different sites according to both the RMSE and MRE metrics are consistent with the previous results.

As the above comparison between the two models was carried out for the long-term time series of water levels, the models also needed to be evaluated in the short-term series; in addition, the difference in peak values was an important indicator for model validation. Therefore, input data from the high water level period (June to August) was selected to re-run the model and compare predicted and measured values. The comparisons for the three representative stations in Lake Dongting are shown in Figures 9–11.

The performance evaluation results of the assimilated and non-assimilated Dongting Lake water level prediction models for the peak water level period are presented in Table 4. The prediction accuracy of the non-assimilated model is within an acceptable range, with the coefficient of determination R^2 ranging from 0.949 (Chenglingji) to 0.984 (Xiaohezui) at different stations, the RMSE ranging from 0.103 (Xiaohezui) to 0.487 (Chenglingji), and the MRE in the range of 0.003 (Xiaohezui) to 0.017 (Chenglingji). Among the three representative stations in Dongting Lake, Xiaohezui has the best prediction accuracy, Yingtian has the second highest, and Xiaohezui has the lowest, which is equivalent to the error distribution of predictions produced using the SVR model alone.

When the actual water levels at each station were continuously incorporated into the Dongting Lake water level prediction model utilizing the UKF, the performance of the model prediction was greatly enhanced. According to Table 4, R^2 ranges from 0.998 (Yingtian) to 0.999 (Chenglingji and Xiaohezui), RMSE ranges from 0.037 (Xiaohezui) to 0.116 (Yingtian), and MRE ranges from 0.001 (Xiaohezui) to 0.004 (Yingtian). It is worth mentioning that the phenomenon that the determination coefficient R^2 of the Xiaohezui model is slightly smaller than that of the models at the remaining two sites is related to the smaller range of water level fluctuations at these two sites.

From the perspective of the error probability (Figure 12), the prediction errors for the data assimilation model fluctuate in a smaller range and are more concentrated than those for the non-assimilation model. For the non-assimilated models at Chenglingji and Yingtian, the prediction error range is approximately −1 m to 1 m, while for the non-assimilated model at Xiaohezui, the error range narrows to −0.5 m to 0.5 m. The prediction error fluctuation range for the assimilated models at the three aforementioned stations is further significantly reduced, with the maximum range for the assimilated model at Yingtian being −0.3 m to 0.3 m.

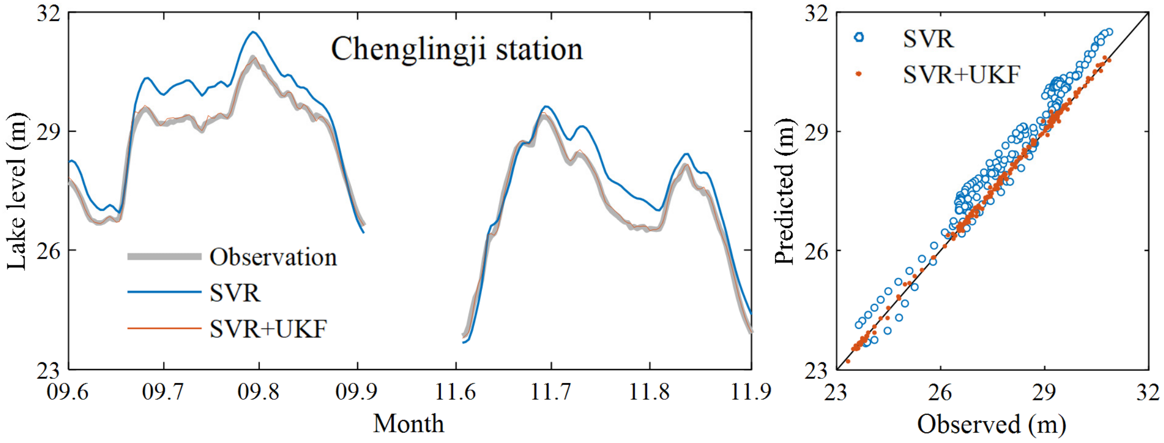

Figure 9. Comparisons between peak water levels in Chenglingji predicted with SVR and SVR+UKF.

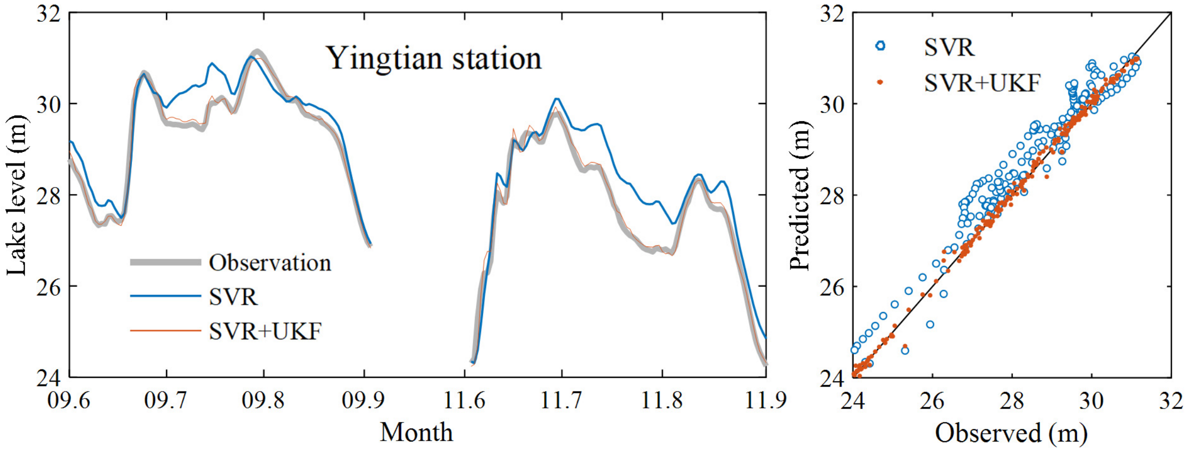

Figure 10. Comparisons between peak water levels in Yingtian predicted with SVR and SVR+UKF.

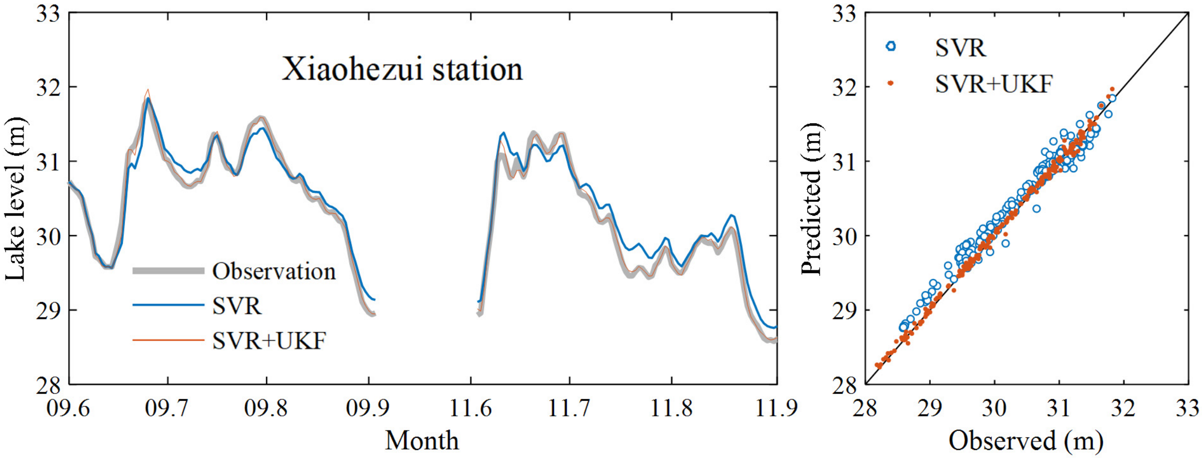

Figure 11. Comparisons between peak water levels in Xiaohezui predicted with SVR and SVR+UKF.

Table 4. Performance comparisons between peak water levels with SVR and SVR+UKF.

Station	SVR			SVR+UKF		
	R^2	RMSE (m)	MRE	R^2	RMSE (m)	MRE
Chenglingji	0.949	0.487	0.017	0.999	0.073	0.002
Yingtian	0.961	0.412	0.016	0.998	0.116	0.004
Xiaohezui	0.984	0.103	0.003	0.999	0.037	0.001

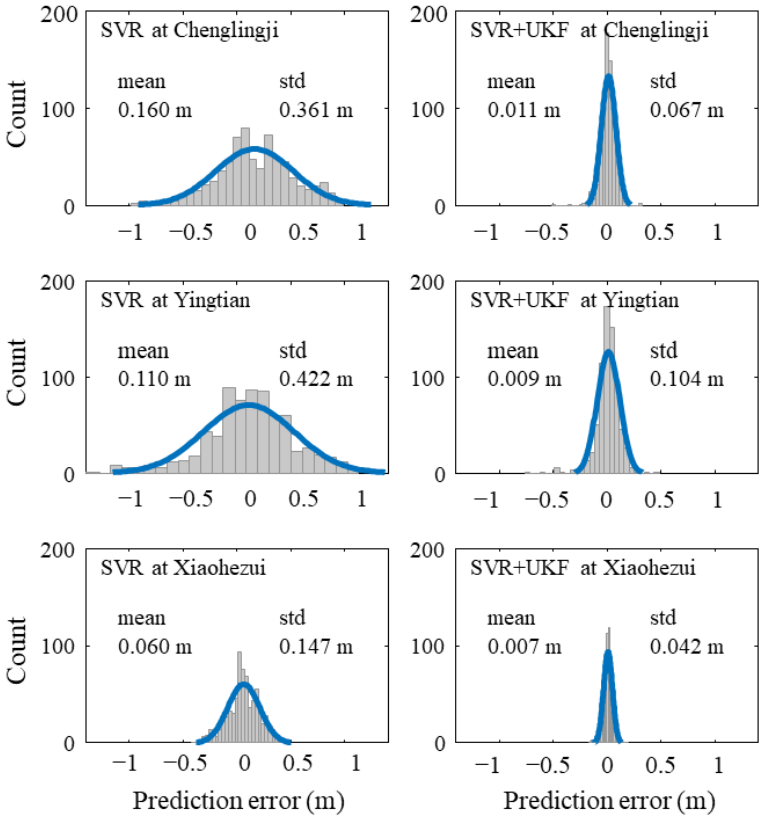

Figure 12. Probabilistic histograms of prediction errors of SVR and SVR+UKF.

The mean values for the non-assimilated models at Chenglingji and Yingtian are 0.160 m and 0.110 m respectively, while the mean value for the Xiaohezui model is smaller at 0.060 m. For the assimilated model, the positive and negative distribution of errors is more even, with the mean value of errors at all sites not greater than 0.024 m. Given the small fluctuation range of errors for the assimilated model, the standard deviation of errors is also smaller. The standard deviation of the error is also smaller than that of the non-assimilated model. The maximum values of the standard deviation of errors for the assimilation and non-assimilation models were 0.104 m and 0.422 m at the three stations, respectively, and the model with the largest standard deviations was the Yingtian model.

3.2. Further Model Testing

The data-driven model requires further analysis of the performance from different perspectives after it has been built due to the rapid deterioration and lack of stability. An important aspect for water level prediction is the ability to respond to missing data. For example, when a water level monitoring instrument fails to transmit data to the model and it can’t be repaired immediately, the lack of data will continue for some time; the model’s performance in this case also represents its ability to maintain stable prediction in extreme circumstances.

We define T_a as the number of data that can be assimilated at one time in a cycle and T_{non} as the number of missing data after one assimilation in a cycle (e.g., 1 for data transmission, 0 for no data transmission, $T_a = 2$ and $T_{non} = 1$ when the data is transmitted in the format 110110110 . . . , and $T_a = 3$ and $T_{non} = 2$ when the format is 111001110011100 . . .). During the application period of the model, predictions are made for each day and the model is specifically pre-warmed up using the T_a/T_{non} cycle.

Figures 13–15 provide the error analysis for the three representative stations with varying values of T_a and T_{non}. The results show consistent patterns and trends: when T_a is certain, the prediction accuracy decreases as T_{non} increases, i.e., as more missing data are transmitted in a cycle, the overall accuracy of the model will continue to decrease; when T_{non} is certain, the prediction accuracy increases as T_a increases, i.e., when the more data are continuously transmitted for assimilation in a cycle, the more the accuracy of the model will increase. Notably, the performance of the model with $T_a = 1$ (green line) is significantly worse than with $T_a \geq 2$, and it fluctuates dramatically, indicating that the periodic transmission of only one datum is unfavorable to the assimilation model and that it is essential to ensure that two or more data are transmitted to the assimilation model.

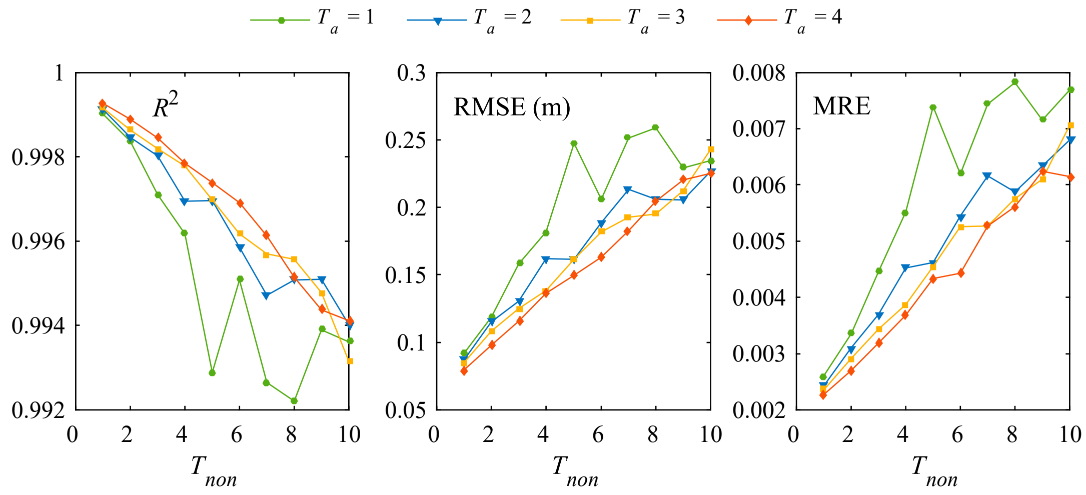

Figure 13. SVR+UKF performance under different assimilation/non-assimilation cycles at Chenglingji.

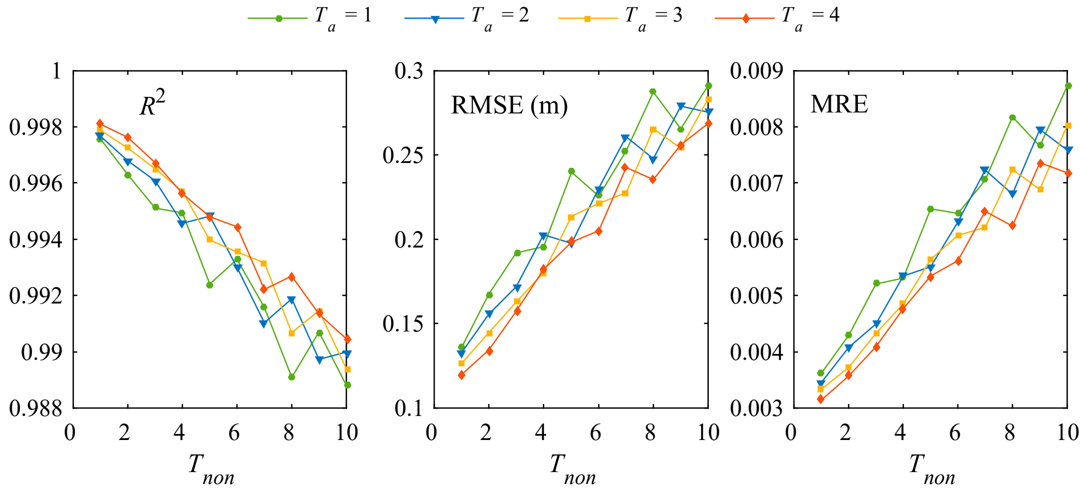

Figure 14. SVR+UKF performance under different assimilation/non-assimilation cycles at Yingtian.

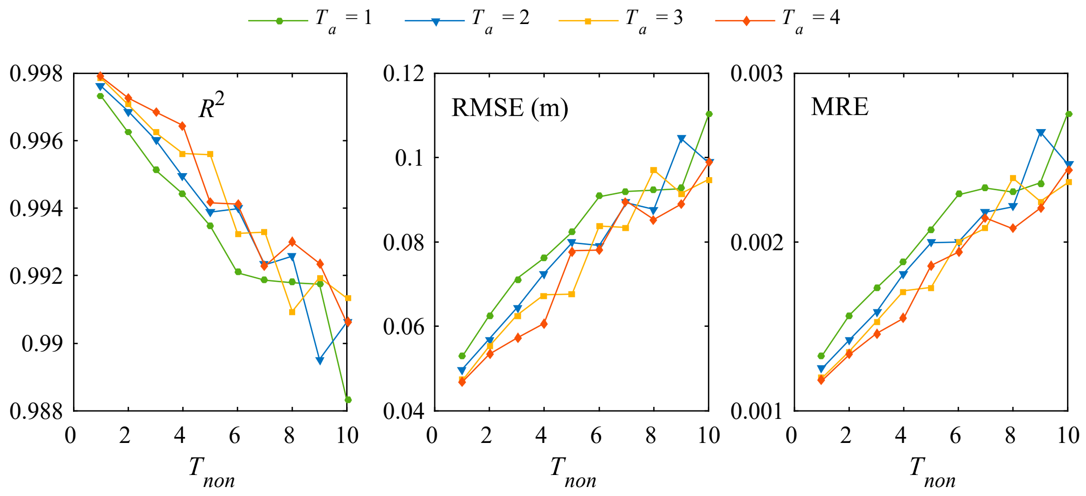

Figure 15. SVR+UKF performance under different assimilation/non-assimilation cycles at Xiaohezui.

Another indicator that is vital for water level prediction is the prediction lead time, which enables management to make decisions in advance. For each day of the model application period described above, T_a data were assimilated before the prediction, and after the prediction was completed for that day, the data were not assimilated and the forecast values were utilized to continue the prediction. For each of the three sites, the error evaluation criteria were set according to their prediction error characteristics, i.e., the upper quartile of prediction error reached 0.1 m/0.2 m/0.2 m before the model was able to run for several time steps (days) with different assimilated data; Figure 16 demonstrates that the time lead time for all three sites exhibited a significant positive correlation with T_a. The Xiaohezui model can achieve a lead time of approximately 10 days, with an upper quartile of prediction error of less than 0.1 m, followed by the Chenglingji model with an upper quartile of prediction error of less than 0.2 m. The lead time for $T_a \geq 2$ is 6–9 days, and the Yingtian model performed poorly, with a lead time of 5–7 days for conditions where the upper quartile of prediction error was less than 0.2 m ($T_a \geq 2$).

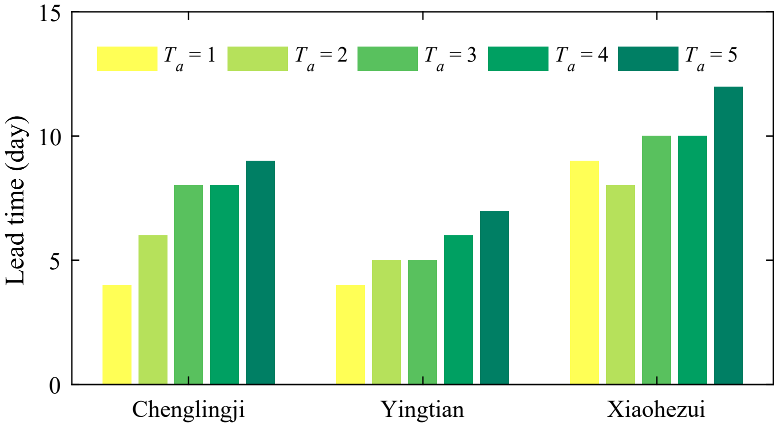

Figure 16. Lead time of SVR+UKF for Dongting Lake water level prediction.

Data-driven approaches are potent tools for fitting links between design variables and prediction targets, employing fewer conditions and fitting them more effectively than conventional physical models. The accuracy of data-driven models is affected by structure, parameters, and data uncertainty, which causes data-driven models to degrade rapidly in comparison to classic physical models when predicting lengthy series [29,30]. Recurrent neural networks (RNNs) and data assimilation (DA), among other data-driven methods, are frequently employed to overcome this challenge. RNNs may formulate weights using past data and have an advantage when learning non-linear sequence characteristics. However, due to the gradient explosion or gradient disappearance problems, RNN cannot have too long a time series, resulting in RNN networks that can only support short-term but not long-term memory [14,31]. It was found that embedding data-driven models into DA techniques can significantly improve predictions and reduce the rate of deterioration [21]. The unscented Kalman filter is one of the DA techniques based on the unscented transform, discarding the traditional approach of linearizing non-linear functions and employing a Kalman linear filtering framework that does not linearly ignore higher order terms in advance, resulting in a more precise computation of non-linear distribution statistics [28].

In this study, the SVR model was added to the DA framework to significantly improve the prediction accuracy. According to the performance evaluation results of the assimilated and non-assimilated Dongting Lake water level prediction models in Table 3, it can be seen that from the non-assimilated model to the assimilated model, the coefficient of determination R^2 improved from 0.982 to 0.999 for Chenglingji, from 0.975 to 0.999 for Yingtian, and from 0.975 to 0.998 for Xiaohezui; Chenglingji's RMSE decreased from 0.395 m to 0.068 m, Yingtian from 0.436 m to 0.105 m, and Xiaohezui from 0.159 m to 0.042 m, all of which significantly improved the model's predictive capability in practical application scenarios. This data assimilation method is equally applicable to other artificial neural network (ANN) methods and can be extended by coupling them.

Although DA+SVR has made significant improvements to prediction model accuracy and prediction days, DA+SVR is still in the realm of data-driven methods, and the underlying model is still SVR, which has applicability for systems with constant (relatively static) laws, with the drawback that, consistent with traditional ANNs, it does not respond well to changing systems, and in realistic situations there may be long periods of non-existence of the expected assimilated data.

4. Conclusions

This study chose to embed the SVR into the UKF model to evaluate its potential for modeling lake water levels, comparing lake level estimates with the model employing only the SVR and drawing the following conclusions: (1) The SVR+UKF model outperformed the SVR model regardless of the assimilated time scale. (2) The prediction performance of the assimilated model is unaffected by the characteristics of lake level variations during the prediction time period and is able to retain its consistency. (3) The assimilation model's prediction error has a narrower fluctuation range and a larger concentration than the non-assimilation model. (4) For assimilation, it is preferable to have at least two consecutive data, which is advantageous for error indicators and prediction lead time. The results show that the data assimilation method provides an alternative for coupling data-driven models that can improve the predictive capability of the model in practical application scenarios.

Author Contributions: Conceptualization, T.H. and K.W.; methodology, T.H.; software, T.H. and P.Z.; validation, W.H.; formal analysis, Y.X. and Y.S.; data curation, K.W. and W.H.; writing—original draft preparation, T.H.; writing—review and editing, K.W. and T.H.; supervision, T.H. and J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (2022YFC3002704, 2021YFC3200301), the National Natural Science Foundation of China (51909168, 52109013), and the China Postdoctoral Science Foundation (2021M701049).

Data Availability Statement: The data used in this study are available on request from the corresponding author.

Acknowledgments: The constructive comments and suggestions of the anonymous reviewers and the editors are gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Govindaraju, R.S. Artificial Neural Networks in Hydrology. I: Preliminary Concepts. *J. Hydrol. Eng.* **2000**, *5*, 115–123.
2. Govindaraju, R.S. Artificial Neural Networks in Hydrology. II: Hydrologic Applications. *J. Hydrol. Eng.* **2000**, *5*, 124–137.
3. Abebe, A.J.; Price, R.K. Information Theory and Neural Networks for Managing Uncertainty in Flood Routing. *J. Comput. Civil. Eng.* **2004**, *18*, 373–380. [CrossRef]
4. van den Boogaard, H.; Mynett, A. Dynamic Neural Networks with Data Assimilation. *Hydrol. Process.* **2004**, *18*, 1959–1966. [CrossRef]
5. Maier, H.R.; Jain, A.; Dandy, G.C.; Sudheer, K.P. 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.* **2010**, *25*, 891–909. [CrossRef]
6. Moayedi, H.; Armaghani, D.J. Optimizing an ANN Model with ICA for Estimating Bearing Capacity of Driven Pile in Cohesionless Soil. *Eng. Comput.* **2018**, *34*, 347–356. [CrossRef]
7. Cimen, M.; Kisi, O. Comparison of Two Different Data-Driven Techniques in Modeling Lake Level Fluctuations in Turkey. *J. Hydrol.* **2009**, *378*, 253–262. [CrossRef]
8. Khan, M.S.; Coulibaly, P. Application of Support Vector Machine in Lake Water Level Prediction. *J. Hydrol. Eng.* **2006**, *11*, 199–205. [CrossRef]
9. Gu, R.C.; McCutcheon, S.; Chen, C.J. Development of Weather-Dependent Flow Require-ments for River Temperature Control. *Environ. Manag.* **1999**, *24*, 529–540. [CrossRef]
10. Maier, H.R.; Dandy, G.C. Neural Networks for the Prediction and Forecasting of Water Resources Variables: A Review of Modelling Issues and Applications. *Environ. Modell. Softw.* **2000**, *15*, 101–124. [CrossRef]
11. Zhang, X.; Liang, F.; Srinivasan, R.; Van Liew, M. Estimating Uncertainty of Streamflow Simulation Using Bayesian Neural Networks. *Water Resour. Res.* **2009**, *45*, W02403. [CrossRef]
12. Zhang, X.; Wang, H.; Peng, A.; Wang, W.; Li, B.; Huang, X. Quantifying the Uncertainties in Data-Driven Models for Reservoir Inflow Prediction. *Water Resour. Manag.* **2020**, *34*, 1479–1493. [CrossRef]
13. Galelli, S.; Humphrey, G.B.; Maier, H.R.; Castelletti, A.; Dandy, G.C.; Gibbs, M.S. An Evaluation Framework for Input Variable Selection Algorithms for Environmental Data-Driven Models. *Environ. Modell. Softw.* **2014**, *62*, 33–51. [CrossRef]
14. Kingston, G.B.; Lambert, M.F.; Maier, H.R. Bayesian Training of Artificial Neural Networks Used for Water Resources Modeling. *Water Resour. Res.* **2005**, *41*, W12409. [CrossRef]
15. Quilty, J.; Adamowski, J.; Boucher, M.-A. A Stochastic Data-Driven Ensemble Forecasting Framework for Water Resources: A Case Study Using Ensemble Members Derived from a Database of Deterministic Wavelet-Based Models. *Water Resour. Res.* **2019**, *55*, 175–202. [CrossRef]
16. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, *157*, 498–513. [CrossRef]
17. Vrugt, J.A.; Diks, C.G.H.; Gupta, H.V.; Bouten, W.; Verstraten, J.M. Improved Treatment of Uncertainty in Hydrologic Modeling: Combining the Strengths of Global Optimization and Data Assimilation. *Water Resour. Res.* **2005**, *41*, W01017. [CrossRef]
18. Liu, Y.; Gupta, H.V. Uncertainty in Hydrologic Modeling: Toward an Integrated Data Assimilation Framework. *Water Resour. Res.* **2007**, *43*, W07401. [CrossRef]
19. Liu, Y.; Weerts, A.H.; Clark, M.; Franssen, H.-J.H.; Kumar, S.; Moradkhani, H.; Seo, D.-J.; Schwanenberg, D.; Smith, P.; van Dijk, A.I.J.M.; et al. Advancing Data Assimilation in Operational Hy-drologic Forecasting: Progresses, Challenges, and Emerging Opportunities. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 3863–3887. [CrossRef]
20. Mao, J.Q.; Lee, J.H.W.; Choi, K.W. The Extended Kalman Filter for Forecast of Algal Bloom Dynamics. *Water Res.* **2009**, *43*, 4214–4224. [CrossRef]
21. Zamani, A.; Azimian, A.; Heemink, A.; Solomatine, D. Non-Linear Wave Data Assimilation with an ANN-Type Wind-Wave Model and Ensemble Kalman Filter (EnKF). *Appl. Math. Model.* **2010**, *34*, 1984–1999. [CrossRef]
22. Gill, M.K.; Kemblowski, M.W.; McKee, M. Soil Moisture Data Assimilation Using Support Vector Machines and Ensemble Kalman Filter. *J. Am. Water Resour. Assoc.* **2007**, *43*, 1004–1015. [CrossRef]
23. Yu, Z.; Liu, D.; Lu, H.; Fu, X.; Xiang, L.; Zhu, Y. A Multi-Layer Soil Moisture Data Assimilation Using Support Vector Machines and Ensemble Particle Filter. *J. Hydrol.* **2012**, *475*, 53–64. [CrossRef]
24. Kalman, R. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **1960**, 35–45. [CrossRef]
25. Lee, J.H.W.; Mao, J.Q.; Choi, K.W. The Extended Kalman Filter for Short Term Prediction of Algal Bloom Dynamics. In *Advances in Water Resources and Hydraulic Engineering*; Zhang, C.K., Tang, H.W., Eds.; Tsinghua Univ Press: Beijing, China, 2009; Volume 1–6, pp. 513–517.
26. Evensen, G. Sequential Data Assimilation with a Nonlinear Quasi-geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics. *J. Geophys. Res. Ocean.* **1994**, *99*, 10143–10162. [CrossRef]
27. Julier, S.J.; Uhlmann, J.K. A New Extension of the Kalman Filter to Nonlinear Systems. In Proceedings of the Signal Processing, Sensor Fusion, and Target Recognition VI, Orlando, FL, USA, 21–24 April 1997; Kadar, I., Ed.; SPIE—International Society Optical Engineering: Bellingham, WA, USA, 1997; Volume 3068, pp. 182–193.
28. Qi, J.; Sun, K.; Wang, J.; Liu, H. Dynamic State Estimation for Multi-Machine Power System by Unscented Kalman Filter with Enhanced Numerical Stability. *IEEE Trans. Smart Grid* **2018**, *9*, 1184–1196. [CrossRef]

29. Ahani, A.; Shourian, M.; Rad, P.R. Performance Assessment of the Linear, Nonlinear and Nonparametric Data Driven Models in River Flow Forecasting. *Water Resour. Manag.* **2018**, *32*, 383–399. [CrossRef]
30. Shu, C.; Ouarda, T.B.M.J. Flood Frequency Analysis at Ungauged Sites Using Artificial Neural Networks in Canonical Correlation Analysis Physiographic Space. *Water Resour. Res.* **2007**, *43*, W07438. [CrossRef]
31. Haque, A.; Rahman, S. Short-Term Electrical Load Forecasting through Heuristic Configuration of Regularized Deep Neural Network. *Appl. Soft. Comput.* **2022**, *122*, 108877. [CrossRef]