

A New Water Level Prediction Model for Open-channel Water Transfer Projects with Pumping Stations Based on KNN-ENKF

Peiru Yan¹, Zhao Zhang², Xiaohui Lei³, Cheng Sui⁴, and Zhiyue Yao¹

¹Shandong Agricultural University, Tai'an, China

²China Institute of Water Resources and Hydropower Research, Beijing, China

³Hebei University of Engineering, Handan, China

⁴Shandong Dayu Water Construction Group CO.LTD, Jinan, China

DOI: <https://doi.org/10.14796/JWMM.C583>

ABSTRACT

It is difficult to accurately predict the water levels of an open-channel water transfer project with pumping stations due to various uncertainties. To solve this problem, this paper proposes a new water level prediction model based on K Nearest Neighbor Algorithm-Ensemble Kalman Filter (KNN-ENKF). A hydrodynamic model is used to construct the water regime database, and the KNN is used to search for the data that best matches the current water regime state, which is taken as the initial value of the model to ensure the prediction accuracy in the initial stage. Then, ENKF is used to real-time correct the water regime state and reduce error accumulation. A case study is conducted with the Songzhuang-Bushang section of the Jiaodong Water Transfer Project, China. It is found that the KNN-ENKF model can significantly improve the water level prediction accuracy with an average error of less than 0.04 m.

1. INTRODUCTION

Water shortage has become a serious problem in many countries and regions largely because of rapid population growth and economic development (Elhenawy et al. 2023; Raveesh et al. 2021), and a common solution to this problem is to build open-channel water transfer projects with pumping stations (Han et al. 2020; Yan et al. 2022). For such a

Yan, P., Z. Zhang, X. Lei, C. Sui, and Z. Yao. 2026. "A New Water Level Prediction Model for Open-channel Water Transfer Projects with Pumping Stations Based on KNN-ENKF." *Journal of Water Management Modeling* 34: C583. <https://doi.org/10.14796/JWMM.C583> www.chijournal.org ISSN: 2292-6062 Yan et al. 2026



project, the water level can be an important indicator for evaluating the operation conditions of the channel, and therefore it is essential to accurately predict water level changes to minimize energy consumption and operating costs of pumping stations. In general, the water level prediction models are built based on either data-driven or hydrodynamic models. A data-driven model is intended to explore the relationship between the data without considering the physical mechanism. With the rapid development of artificial intelligence technology, a multitude of data-driven models have been applied to water level prediction (Ren et al. 2020; Jang et al. 2022; Yan et al. 2023). However, these data-driven models often require high data quality, and the prediction accuracy will be low in the presence of outliers in the data or in the case of a small amount of data. The hydrodynamic models based on physical mechanisms would provide a more accurate prediction using a smaller amount of data through calibration. The traditional Saint-Venant equations can be used to simulate channels rather than water transfer projects that involve numerous structures such as gates, pumping stations, and inverted siphons. To meet engineering requirements, these structures are modularized as internal boundaries and coupled with the Saint-Venant equations, enabling the continuous simulation of water transfer projects (Zhang et al. 2007; Lu et al. 2018). To improve model accuracy, some methods such as genetic algorithms and Kalman filters are used to calibrate model parameters (Tang et al. 2010; Lei et al. 2019). Studies have shown that calibrated models can significantly enhance prediction accuracy. However, many uncertainties are not considered in these models. This is an important problem because many water transfer projects are built in remote suburbs and are greatly affected by many unknown uncertainties (Yan et al. 2025).

Data assimilation is an effective means of reducing uncertainties, and therefore it has the potential to improve the prediction accuracy of the hydrologic model (Bourgin et al. 2014; Ouellet-Proulx et al. 2017; Kim et al. 2021; Liu et al. 2016). Current data assimilation methods include Kalman filter, particle filter, and variational assimilation (Van Wesemael et al. 2019; Noh et al. 2014; Gan et al. 2022). The Kalman filter and its variant algorithms are widely used because of their simplicity and convenience, particularly the Ensemble Kalman Filter (EnKF). Numerous studies have demonstrated that integrating hydrodynamic models with EnKF effectively improves the accuracy of water level prediction (Barthélémy et al. 2017; Yu et al. 2017; Cooper et al. 2018; Lee et al. 2019). Attempts have also been made to combine data-driven models with data assimilation (Fu et al. 2024). However, although data assimilation is widely used for rivers, it is rarely used for open-channel water transfer projects with pumping stations. The presence of hydraulic structures such as pumping stations is likely to affect flow propagation and consequently the performance of data assimilation. As such, the ensemble Kalman filter is used in this study for water level

prediction of an open-channel water transfer project with pumping stations, and the influences of observation location and data type on the assimilation are investigated.

As the ensemble Kalman filter is a sequential assimilation technique, some time is needed to obtain the ideal state when the initial deviation is large, which will reduce the prediction accuracy of the model in the early stage. At present, a steady-state start or a hot start is used in some software (e.g., MIKE 11). The steady-state start is to calculate the initial water level and discharge at each calculation node using steady-state assumptions based on the water level and discharge data provided at the boundary points. However, it is difficult for the water flow to reach a steady state in practical engineering applications, thus leading to a discrepancy between the initial state of the model and the actual situation, and consequently reducing the prediction accuracy. The hot start is to use the model calculation for a period, and the stable water level and discharge are taken as the initial condition. However, it does not always guarantee that the initial condition is consistent with the actual situation, and it will also increase the calculation time and reduce the calculation efficiency. The interpolation method is also used to set the initial condition (Zhu et al. 2021) and the water level and discharge along the route are interpolated according to observed data. This method is simple and convenient, but it is highly dependent on data. When there are only a few observation points or the project is rather complex with many buildings, it would be difficult to accurately estimate the initial condition of each section. In addition, as the pumping station is essentially a barrier to the flow, the initial error along the route will accumulate and eventually increase the prediction error of the water level in front of the pumping station. At present, little research has been done to address the initial condition of the model. To fill this gap, this study proposes a new method for setting the initial condition. The water regime database is built through the model, and the K -nearest neighbours (KNN) algorithm is used to search for the data that best matches the current situation using the current observation data as the initial condition, which can improve the accuracy of the initial condition, and then the prediction accuracy of the model.

The main innovations of this study are as follows:

1. A new method is proposed to set the initial condition of the model.
2. A new water level prediction model is proposed based on KNN-ENKF for open-channel water transfer projects with pumping stations.
3. The impact of observation position and data type on assimilation is discussed.

The paper is organized as follows: Section 2 introduces the basic principles and evaluation indicators of the model; Section 3 describes the study area; Section 4 presents the main results; and Section 5 presents the conclusions.

2. METHODS

In this study, a new water level prediction model is proposed based on KNN-ENKF for open-channel water transfer projects with pumping stations. To this end, a calibrated hydrodynamic model is first used to calculate the water level and discharge processes along the route under different operating conditions, and then a model database is constructed. Prior to prediction, KNN is used to match the water level and discharge in the database based on the current monitoring station. Then, a hydrodynamic model is used to predict water level changes. Once the latest observation data is obtained, ENKF assimilation is performed to correct the current state to reduce model error and improve model accuracy. The flowchart of the model is shown in Figure 1.

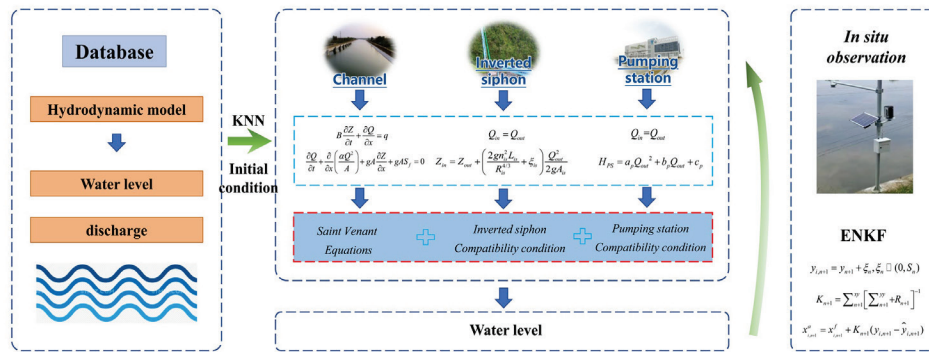


Figure 1 A KNN-ENKF-based water level prediction model for open-channel water transfer projects with pumping stations.

2.1 One-dimensional hydrodynamic model

Open channel

The governing equations (Equations 1 and 2) for the one-dimensional unsteady flow in an open channel were first proposed by Saint Venant in 1871, including continuity and momentum equations (Yan et al. 2021; Kong et al. 2023).

$$B \frac{\partial Z}{\partial t} + \frac{\partial Q}{\partial x} = q \quad (1)$$

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x} \left(\frac{\alpha Q^2}{A} \right) + gA \frac{\partial Z}{\partial x} + gAS_f = 0 \quad (2)$$

Where:

B = channel surface width (m),

Z = water level (m),

t = time (s),

Q = discharge (m³/s),

x = distance along the channel (m),

q = lateral inflow (m²/s),

α = momentum correction coefficient,

A = wetted cross-sectional area (m²),

g = acceleration of gravity (m/s²), and

S_f = friction slope, which can be expressed using Equation 3, as follows:

$$S_f = \frac{n^2 Q |Q|}{A^2 R^{4/3}} \quad (3)$$

Where:

n = Manning's roughness coefficient, and

R = hydraulic radius of the channel (m).

Pumping station

The continuity equation and discharge-head characteristic curve of the pumping station are used as the control equations. The continuity equation not considering the water loss in the pumping station can be written as follows (Equation 4):

$$Q_{in} = Q_{out} \quad (4)$$

Where:

Q_{in} = discharge at the inlet of the pumping station (m³/s), and

Q_{out} = discharge at the outlet of the pumping station (m³/s).

The discharge-head characteristic curve can be written as follows (Equation 5):

$$H_{PS} = Z_{ps,out} - Z_{ps,in} = a_p Q_{out}^2 + b_p Q_{out} + c_p \quad (5)$$

Where:

- H_{ps} = head of the pumping station (m),
- $Z_{ps.in}$ = water level at the inlet of the pumping station (m),
- $Z_{ps.out}$ = water level at the outlet of the pumping station (m), and
- a_p, b_p and c_p = parameters of the characteristic curve.

Inverted siphon

The inverted siphon differs from the open channel in terms of hydraulic characteristics. For instance, the Saint Venant equations are not applicable to an inverted siphon and must be modified, and the computational time step of the open channel is much longer than the propagation time of flow in the inverted siphon. Therefore, the inverted siphon can be treated as head loss (Yan et al. 2021). The continuity and energy equations are used as the governing equations, which are solved simultaneously using the Saint Venant equations. The control equations of the inverted siphon are described in Equations 6 and 7, as follows:

$$Q_{in} = Q_{out} \quad (6)$$

$$Z_{in} = Z_{out} + \left(\frac{2g n_{is}^2 L_{is}}{R_{is}^{4/3}} + \alpha_{is} \right) \frac{Q_{out}^2}{2g A_{is}} \quad (7)$$

Where:

- Q_{in} = discharge at the inlet of the inverted siphon (m³/s),
- Q_{out} = discharge at the outlet of the inverted siphon (m³/s),
- Z_{in} = water level at the inlet of the inverted siphon (m),
- Z_{out} = water level at the outlet of the inverted siphon (m),
- n_{is} = roughness of the inverted siphon,
- L_{is} = length of the inverted siphon (m),
- R_{is} = hydraulic radius (m),
- α_{is} = local loss coefficient of the inverted siphon, and
- A_{is} = wetted cross-sectional area of the inverted siphon (m²), respectively.

Model solution

The Preissmann four-point implicit difference method is used to discretize the Saint Venant equations due to its advantages of fast convergence, high efficiency, and good stability (Lyn and Goodwin 1987). The Taylor-series expansion method (Gottardi and Venutelli 2008) is used in Equation 5, and then the second-order and higher terms can be omitted. The

linearized control equation is combined with the linearized Saint Venant equations, and the Chase method is used to solve the problem (Wang et al. 2015).

2.2 K-nearest neighbour algorithm

The KNN is a non-parametric method developed by Evelyn Fix and Joseph Hodges in 1951 and further improved by Thomas Cover. This algorithm is widely used for prediction problems because it is very simple to implement. The main idea of this algorithm is to calculate the distance between each sample and the target sample based on the feature vector and then sort based on the calculated distance. The nearest k samples are weighted to obtain the final prediction results. The size of the measurement distance reflects the degree of similarity between the feature vector and the history vector, and it can search for data that is most like the current state from the database. In this study, the Euclidean distance is used to measure the distance size, which is described as follows.

In the n -dimensional space, the formula for the Euclidean distance do between two points can be expressed using Equation 8:

$$d_0 = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (8)$$

Where:

x_i = i -th component of the feature vector, and

y_i = i -th component of the corresponding vector in the database.

2.3 Ensemble Kalman filter

The Ensemble Kalman filter algorithm was put forward by Evensen in 1994. Now, it is widely used in ocean, atmosphere, hydrology, groundwater, and other fields because of its simplicity and efficiency.

The assimilation process of the ensemble Kalman filter is as follows:

Prediction

According to the state equation, the ensemble state is predicted, and its average value is calculated using Equations 9 and 10:

$$x_{i,n+1}^f = f(x_{i,n}^a, \theta, u_p) + \varepsilon_{i,n} \quad \varepsilon_{i,n} \sim (0, G_n) \quad (9)$$

$$\bar{x}_{n+1}^f = \frac{1}{m} \sum_{i=1}^m x_{i,n+1}^f \quad (10)$$

Where:

- f = model operator,
- θ = model parameter,
- u_p = forced input of the model,
- $x_{i,n+1}^f$ = predicted value of the state variable for the i -th set at time $n+1$,
- $x_{i,n}^a$ = analysis value of the state variable for the i -th set at time n ,
- $\varepsilon_{i,n}$ = model error of the i -th set at time n , which follows the normal distribution with a mean of 0 and a variance of G_n ,
- \bar{x}_{n+1}^f = average value of the predicted set state at time $n+1$, and
- m = number of sets, respectively.

The observation values of the predicted states of the set and their average values are calculated using Equations 11 and 12:

$$\hat{y}_{i,n+1} = H(x_{i,n+1}^f) \quad (11)$$

$$\bar{y}_{n+1} = \frac{1}{m} \sum_{i=1}^m \hat{y}_{i,n+1} \quad (12)$$

Where:

- $\hat{y}_{i,n+1}$ = observation value of the predicted state,
- \bar{y}_{n+1} = average of the observation values, and
- H = observation operator.

Update

The error covariance matrix of the observed prediction value Σ_{n+1}^{yy} , the cross-covariance matrix between the state prediction value Σ_{n+1}^{xy} , and the observed prediction value are calculated using Equations 13 and 14. Since the true value is unknown, the mean value of ensemble prediction is taken as the true value in the ensemble Kalman filter.

$$\Sigma_{n+1}^{yy} = \frac{1}{m-1} (\hat{y}_{i,n+1} - \bar{y}_{n+1})(\hat{y}_{i,n+1} - \bar{y}_{n+1})^T \quad (13)$$

$$\Sigma_{n+1}^{xy} = \frac{1}{m-1} \sum_{i=1}^m (x_{i,n+1}^f - \bar{x}_{n+1}^f) (\hat{y}_{i,n+1} - \bar{y}_{n+1})^T \quad (14)$$

Then, the Kalman gain matrix K_{n+1} and the corrected state analysis values $x_{i,n+1}^a$ are calculated using Equations 15–17:

$$K_{n+1} = \Sigma_{n+1}^{xy} [\Sigma_{n+1}^{yy} + R_{n+1}]^{-1} \quad (15)$$

$$y_{i,n+1} = y_{n+1} + \xi_n, \xi_n \sim (0, S_n) \quad (16)$$

$$x_{i,n+1}^a = x_{i,n+1}^f + K_{n+1}(y_{i,n+1} - \hat{y}_{i,n+1}) \quad (17)$$

Where:

S_n = observing the variance of noise,

ξ_n = observation noise, which follows the normal distribution with a mean of 0, and a variance of S_n ,

$y_{i,n+1}$ = observation value after adding noise disturbance for the i -th time at $n+1$,

$()^T$ = transposition, and

R_{n+1} = covariance of measurement noise at $n+1$

2.4 Evaluation indicators

To verify the superiority of the models proposed in the paper, the prediction ability of each model is evaluated by using Root-mean-square error (RMSE), Mean absolute error (MAE) and Nash efficiency coefficient (NSE) in Equations 18–20, respectively:

$$RMSE = \sqrt{\frac{1}{m} \sum_{t=1}^m (h(t) - \hat{h}(t))^2} \quad (18)$$

$$MAE = \frac{1}{m} \sum_{t=1}^m |h(t) - \hat{h}(t)| \quad (19)$$

$$NSE = 1 - \frac{\sum_{t=1}^m (h(t) - \hat{h}(t))^2}{\sum_{t=1}^m (h(t) - \bar{h}(t))^2} \quad (20)$$

Where:

M = number of model test samples,

$h(t)$ = measured value (m),

$\hat{h}(t)$ = simulated value (m), and

$\bar{h}(t)$ = average of the measured values (m).

3. STUDY AREA

The Jiaodong Water Transfer Project is located in Shandong Province, China, and it has supplied 2.515 billion m³ water to the Jiaodong area and thus plays a critical role in its social and economic development. The Songzhuang-Bushang section of the project is selected as the study area, which is 90.54 km long, has 2 pumping stations, and 20 inverted siphons (Figure 2). The bottom width of the channel is 3.50–8.20 m, the side slope

coefficient is 1–2, and the bottom slope is 0.05–0.10%. The parameters of the channel and inverted siphon are shown in Table 1 and Table 2, respectively. The design discharge of pumping station P1 is 20.7 m³/s. This station has 6 pump units (4 in use and 2 as back-up) with 4 main pumps (1400HD-9) and 2 regulating pumps (1000HDS-9). The design discharge of pumping station P2 is 19.7 m³/s. This station also has 6 pump units (4 in use and 2 as back-up) with 4 main pumps (1400HD-14) and 2 regulating pumps (1000HDS-12).

Table 1 Characteristic parameters of the channel.

Channel	Start stake number (m)	End stake number (m)	Inlet bottom elevation (m)	Outlet bottom elevation (m)	Bottom width (m)	Slope coefficient	Manning's roughness coefficient
L1	-0+023	35+845	6.5	2.1	4.5–8.2	2	0.016
L2	35+845	57+705	9.3	6.59	3.5–6.5	2	0.018
L3	57+705	90+540	19.65	14.29	4.5–6.4	1.5–2	0.013

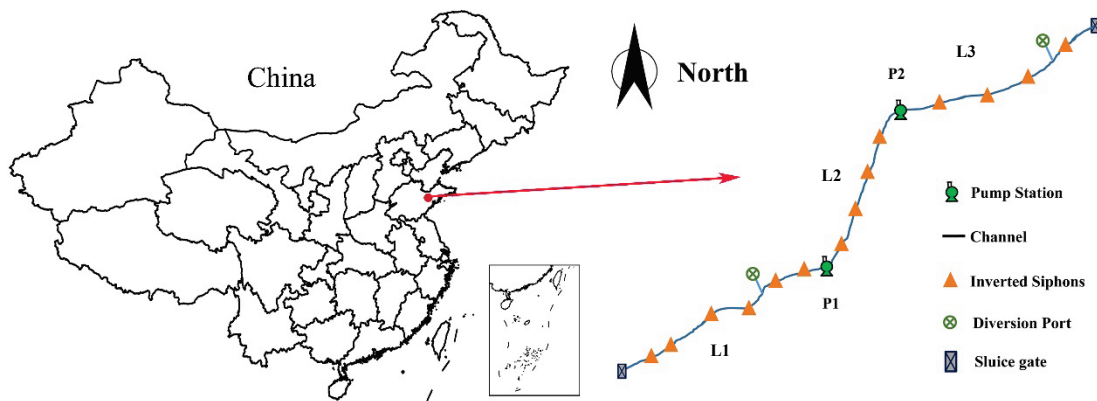


Figure 2 Layout of the Jiadong water transfer project.

Table 2 Characteristic parameters of the inverted siphon.

Inverted siphon	Start stake number (m)	Number of holes	Hole size (width × height)	Length (km)
1	4+465	2	2.5×2.5	0.094
2	5+025	2	3.0×3.0	0.421
3	12+850	2	3.0×3.0	0.0855
4	18+271	2	3.0×3.0	0.161
5	25+056	2	3.0×3.0	0.235
6	30+161	2	3.0×3.0	0.447
7	36+560	2	3.0×3.0	0.0832
8	37+399	2	3.0×3.0	0.0752
9	41+488	2	3.5×3.5	0.0919
10	41+775	2	3.0×3.0	0.0812
11	47+171	2	3.0×.0	0.224
12	50+801	2	3.0×.5	0.116
13	52+995	2	3.0×2.5	0.156
14	54+940	2	3.0×2.5	0.0892
15	59+870	2	3.0×3.0	0.065
16	69+500	2	3.0×3.0	0.694
17	79+711	2	3.0×2.5	0.177
18	86+219	2	3.0×3.0	0.4
19	87+897	2	3.0×2.5	0.0752
20	89+272	2	3.0×2.5	0.0952

A discharge monitoring station is arranged at the channel inlet, and a water level monitoring station is arranged at the channel outlet. Four water level monitoring stations are also arranged in the forebays and afterbays of the two pumping stations, while two discharge monitoring stations are installed inside the two pumping stations. All data are recorded at a time interval of 2 hours.

4. RESULTS AND DISCUSSION

The model parameters were calibrated using the measured water level and discharge data for the period from April 9 to April 19, 2019. The model was validated using measured data for the period from April 20 to April 30, 2019. The upstream boundary of the model was specified as the inflow discharge at the inlet, while the downstream boundary was specified as the water level at the outlet. Three methods were adopted to set the initial

values, which were combined with or without data assimilation, and then resulted in a total of six distinct schemes. Their performances were compared. The detailed configurations of each scheme are presented in Table 3, and the comparison results are presented in Figure 3.

Table 3 Scheme settings.

Scheme	Description
KNN-OPEN	KNN is used to find the water level and discharge data closest to the current state in the database as the initial value of the model. No data assimilation is performed for prediction.
IN-OPEN	The initial value of the model is obtained by interpolation based on the water level and discharge data of the current observation point. No data assimilation is performed for prediction.
CF-OPEN	The constant flow process is calculated based on the current upstream discharge and downstream water level of each channel section, which is used as the initial value of the model. No data assimilation is performed for prediction.
KNN-ENKF	KNN is used to find the water level and discharge data closest to the current state in the database as the initial value of the model. ENKF is used for data assimilation every two hours.
IN-ENKF	The initial value of the model is obtained by interpolation based on the water level and discharge data of the current observation point. ENKF is used for data assimilation every two hours.
CF-ENKF	The constant flow process is calculated based on the current upstream discharge and downstream water level of each channel section, which is used as the initial value of the model. ENKF is used for data assimilation every two hours.

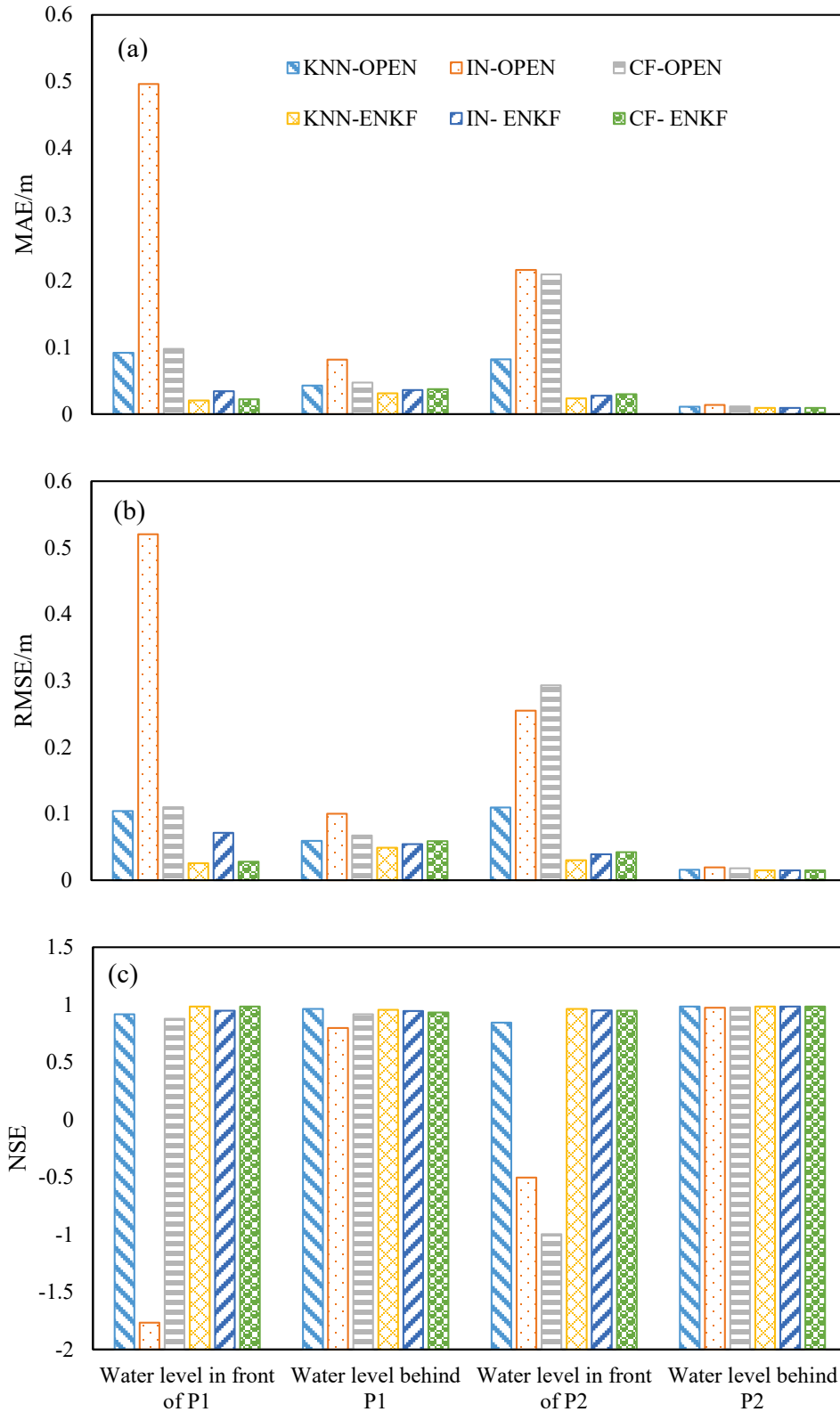


Figure 3 Comparison of prediction results among six schemes during the validation period: (a) MAE, (b) RMSE, and (c) NSE.

4.1 Impact of initial value setting on the prediction results

Figure 3 reveals that KNN-OPEN yields the smallest prediction error, especially for the water level in front of the pumping station. Compared to IN-OPEN and CF-OPEN, the MAE obtained by KNN-OPEN for the predicted water level in front of P1 (0.0923 m) is reduced by 81.39% and 5.53%; the RMSE is reduced by 80.00% and 4.93%; and the NSE is increased by 2.67 and 0.0, respectively. Similarly, the MAE obtained by KNN-OPEN for the predicted water level in front of P2 is reduced by 61.92% and 60.72%; the RMSE is reduced by 57.16% and 71.80%; and the NSE is increased by 1.24 and 1.72, respectively. This indicates that the initial values obtained by the KNN algorithm can best match the current actual state, thereby improving the accuracy of the model. It is necessary to obtain accurate initial values to ensure the accuracy of the hydrodynamic model. As the initial values are spatially heterogeneous due to the undulating terrain of the channel and numerous buildings along the project, the initial values obtained by simple interpolation are most different from the actual situation, hence the worst prediction accuracy. The flow in the channel is not steady during actual operation, resulting in a potential difference between the steady initial value setting and the actual situation and consequently low prediction accuracy of water levels. Thus, the KNN algorithm is used in this study to improve the accuracy of the initial values. He et al. (2022) successfully predicted short-term wind power using the KNN algorithm. Liang et al. (2018) believed that if a complete database is available, the KNN algorithm can better adapt to uncertain, time-varying, and nonlinear short-term prediction problems. In this study, the KNN algorithm is used to predict water regime data at other locations along the route based on the water regime data of main observation stations. If there is a complete water regime database, accurate initial values can be obtained.

KNN-ENKF yields the lowest MAE and RMSE and the highest NSE, especially for the water level in front of the pumping station. Compared to IN-ENKF and CF-ENKF, the MAE for the predicted water level in front of P1 obtained by KNN-ENKF is reduced by 23.19% and 0.63%, and the RMSE is reduced by 46.42% and 0.80%; while the MAE for the predicted water level in front of P2 is reduced by 12.54% and 13.52%, and the RMSE is reduced by 12.91% and 15.14%, respectively.

It is concluded that the initial condition set by the KNN method is closest to the actual situation. The impact of different parameters, including feature vector and the number of nearest neighbours K , on the initial condition of the model is further discussed.

Effect of feature vector on model prediction

The feature vector is important for the KNN algorithm, and the selection of an appropriate feature vector is dependent on the actual situation. In this study, 9 feature vectors are set (Table 4), and their impacts on the prediction results are compared, as shown in Figure 4.

Table 4 Scheme settings.

Scheme	Feature vector
T1	$h1_t, h2_t$
T2	$h1_{t-1}, h1_t, h2_{t-1}, h2_t$
T3	$h1_{t-2}, h1_{t-1}, h1_t, h2_{t-2}, h2_{t-1}, h2_t$
T4	$q1_t, q2_t$
T5	$q1_{t-1}, q1_t, q2_{t-1}, q2_t$
T6	$q1_{t-2}, q1_{t-1}, q1_t, q2_{t-2}, q2_{t-1}, q2_t$
T7	$h1_t, h2_t, q1_t, q2_t$
T8	$h1_{t-1}, h1_t, h2_{t-1}, h2_t, q1_{t-1}, q1_t, q2_{t-1}, q2_t$
T9	$h1_{t-2}, h1_{t-1}, h1_t, h2_{t-2}, h2_{t-1}, h2_t, q1_{t-2}, q1_{t-1}, q1_t, q2_{t-2}, q2_{t-1}, q2_t$

NOTE: $h1$ represents the upstream water level, $h2$ represents the downstream water level, $q1$ represents the upstream discharge, $q2$ represents the downstream discharge, subscript t represents the current moment, $t-1$ represents the previous moment, $t-2$ represents the previous two moments

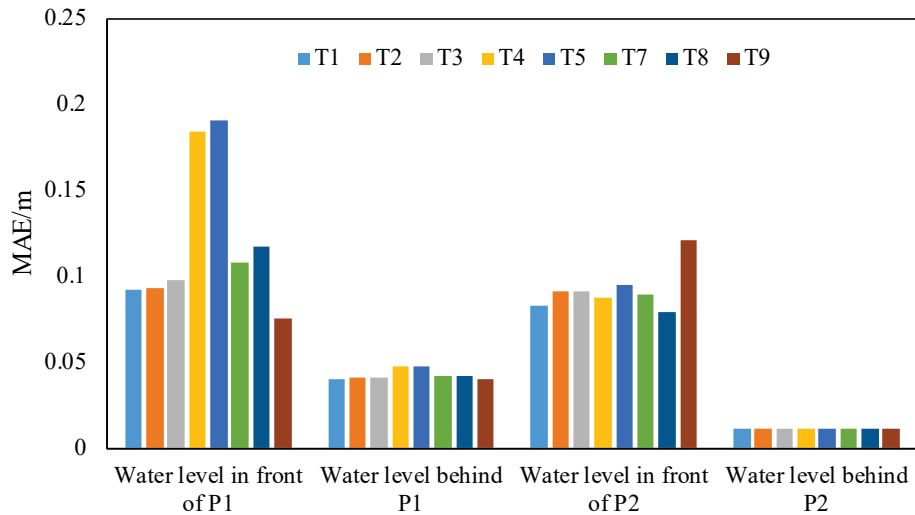


Figure 4 Comparison of prediction errors under different schemes.

As the model collapsed during calculation due to the low initial water level under the T6 scheme, the results are not displayed in Figure 4. It can be seen from Figure 4 that the use of discharge data as the feature vector yields the worst prediction results, and the MAE for the predicted water level in front of P1 is as high as 0.1843 m and 0.1907 m for T4 and T5, respectively; while the use of only water level data as the feature vector yields the best

prediction results, and the average error for the water level can be controlled within 0.1 m, and the prediction error of T1 is lower than that of T2 and T3. However, the results obtained by using both water level and discharge data are rather unstable. In T9, the MAE is lowest for the water level in front of P1, and highest for that in front of P2. In T8, the MAE is lowest for the water level in front of P2, and the water level in front of P1 is as high as 0.1175 m. This indicates that the water level data is an important feature vector that affects the selection of the initial values, which may be because the initial water level has a greater impact on the prediction compared to the initial discharge.

Experiments were conducted to verify this hypothesis, in which the results of T1 were taken as the actual values, and the initial water level and discharge were increased by 10%, 20%, and 30%, respectively. Note that the initial water level and discharge are increased rather than decreased to prevent the model from collapse due to low water level. The results were calculated and compared with the results of T1, as shown in Figure 5.

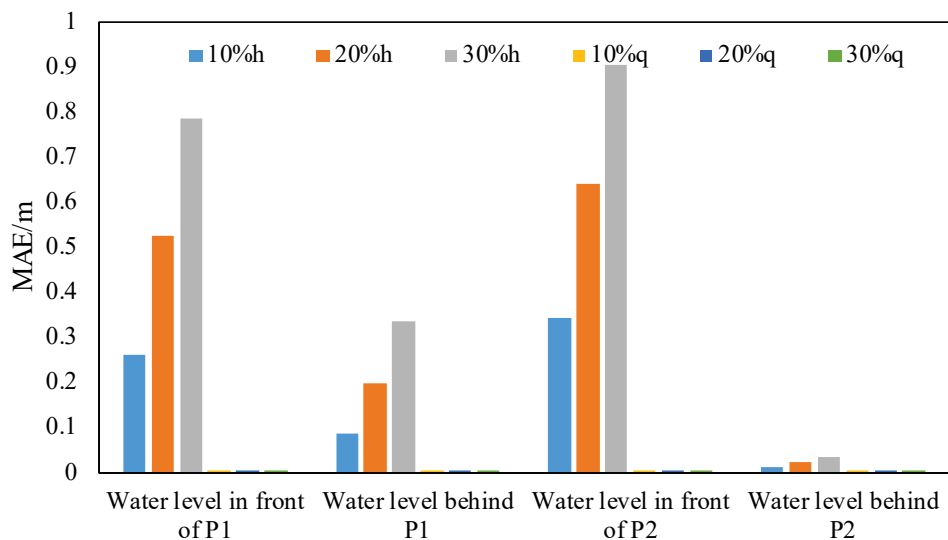


Figure 5 Comparison of prediction errors under different schemes.

It is found that increasing the initial water level leads to a linear increase in the error, while increasing the initial discharge leads to almost no changes in the error. Thus, the hypothesis is supported that the initial water level has a profound impact on the prediction results because of its effect on the storage capacity of the channel, but the impact of the initial discharge is negligible. In an open-channel water transfer project, the pumping station can block the water flow, and its pumping discharge will not change substantially. As there is always a deviation in the storage capacity of the channel, the initial water level will have a significant impact on the prediction results. Therefore, only the upstream and downstream water levels at the current time are selected as the feature vectors, which can

not only reduce the computational complexity but can also improve the accuracy of the initial values of the model.

Impact of K -value on model prediction

In general, there are no strict rules for the selection of K -values. To determine an appropriate K -value, four K -values (1, 3, 5, and 10) were set in this section, and the corresponding simulation results are shown in Figure 6. It is found that there is almost no difference in MAE among the four schemes, indicating that K -value has no significant influence on the selection of the initial value. For the sake of computational efficiency, the K -value is taken to be 1 in this study.

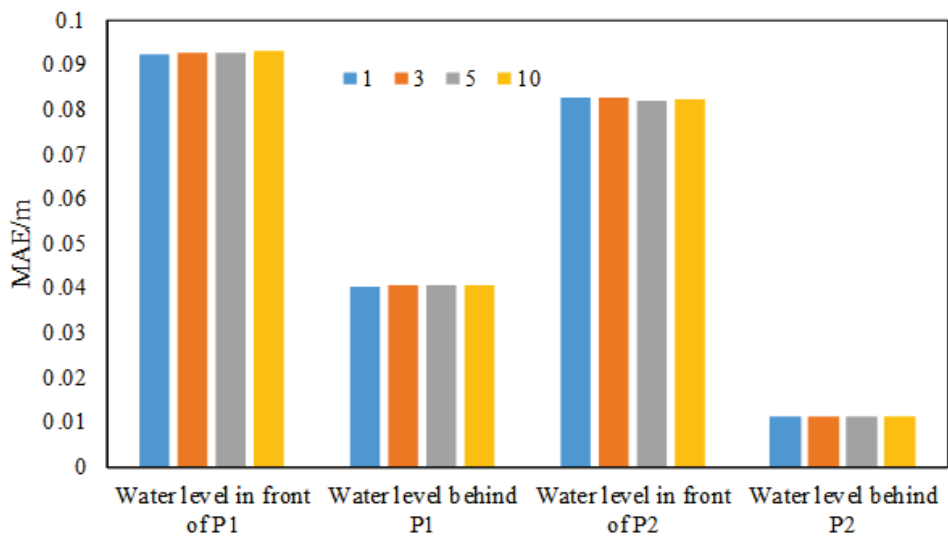


Figure 6 Comparison of prediction errors under different schemes.

4.2 Impact of data assimilation on model prediction

It can be seen from Figure 3 that the MAE and RMSE for the water level in front of P1 are decreased by 65.55–91.65% and 64.07–86.59% after data assimilation; and that behind P1 are decreased by 10.61–51.33% and 9.78–48.60% after data assimilation, respectively. The error for the water level before and after P2 is also decreased. This is because data assimilation can use the latest data to correct the model state at the current time and reduce the accumulation of errors and thus improve the prediction accuracy at the next time. Similar results have also been reported for the prediction of water regimes of rivers. Gu and Lai (2021) used the ensemble Kalman filter method to build the water regime data assimilation model. It was found that the prediction error of the model was reduced by 40% after data assimilation for the Taihu Lake basin. Huang et al. (2017) found that using snow water equivalent data for assimilation effectively improved the prediction accuracy of

seasonal runoff. Patil and Ramsankaran (2018) proposed an assimilation strategy coupling the soil moisture analysis relationship with the ensemble Kalman filter, which could significantly improve the prediction accuracy of runoff.

To better understand the impact of ensemble Kalman filter on the prediction results, the assimilation results are analyzed in the following sections.

Impact of ensemble size on the assimilation results

The ensemble size is an important parameter for the accuracy and efficiency of the ensemble Kalman filter model. In this section, five ensemble sizes (10, 30, 50, 100, and 200) were set and the MAEs for the water level were calculated to determine the most appropriate ensemble size, as shown in Figure 7.

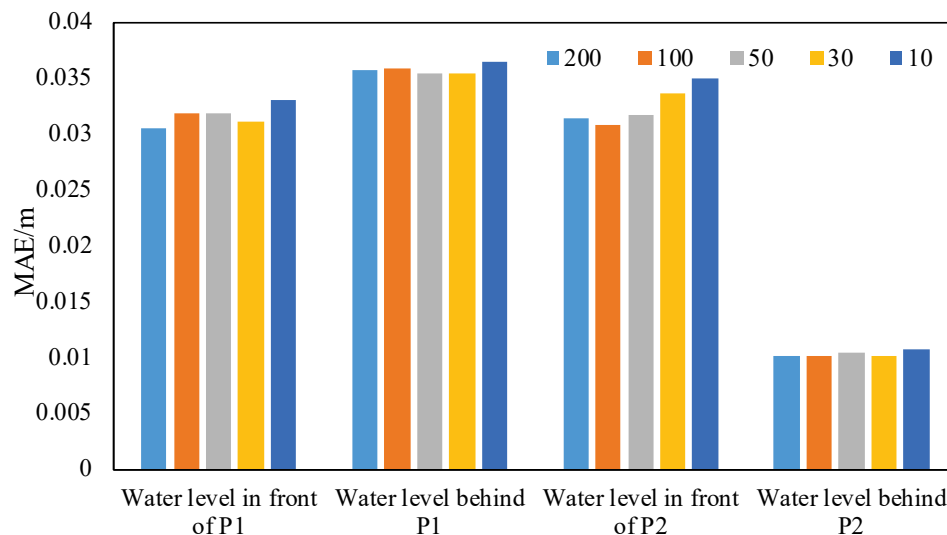


Figure 7 Comparison of prediction errors under different schemes.

It is seen that the maximum error for the water level is obtained at an ensemble size of 10. However, no significant difference is observed in the error at ensemble sizes greater than 50. This is because when the ensemble size is small, the covariance of model prediction errors could not be estimated accurately.

Impact of observation data on the assimilation results

The observation data plays a crucial role in assimilation models, but the impact of its type and monitoring station on the assimilation results for an open-channel water transfer project with pumping stations is still unclear. In this section, six schemes were set to investigate the impact of observation data on assimilation, as shown in Table 5.

Table 5 Scheme settings.

Scheme	Observation data
A1	Water level in front of and behind P1 and P2 discharge of P1 and P2
A2	Water level in front of and behind P1 and P2
A3	Water level in front of P1 and P2
A4	Water level behind P1 and P2
A5	Discharge of P1 and P2
A6	No data assimilation

Figure 8 shows that the prediction accuracy of the water level in front of P1 and P2 in A1–A3 is significantly higher than that in A4–A6, indicating that if the water level data in front of the pumping station is included, the prediction error can be significantly reduced. Similarly, the prediction error for the water level behind P1 and P2 in A1, A2, and A4 is lower than that in A3, A5, and A6. Thus, the inclusion of the water level data behind the pumping station into the assimilation data can significantly reduce the prediction error.

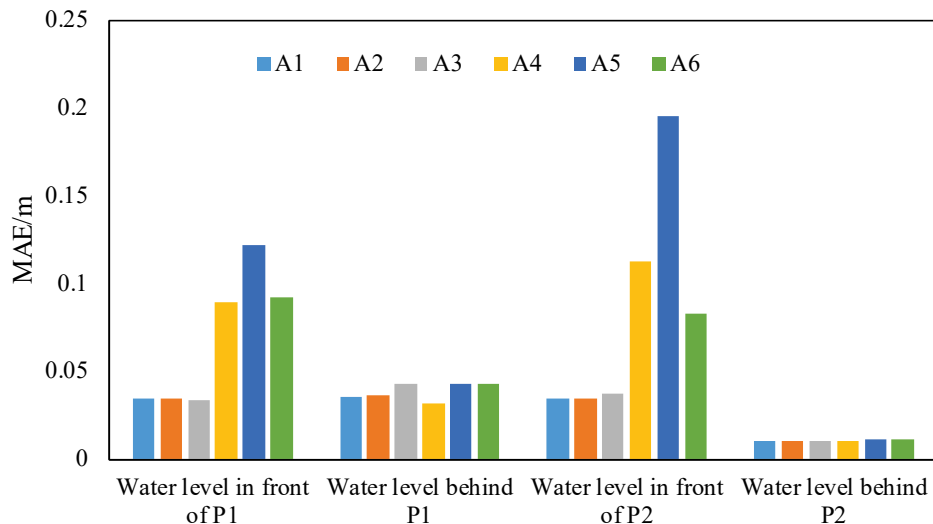


Figure 8 Comparison of prediction errors under different schemes.

The highest prediction error for the water level is observed in A5. Thus, using the discharge data for assimilation will reduce the prediction accuracy of water level, especially the water level in front of the pumping station, which may be attributed to the large monitoring error of discharge and the weak correlation between water level and discharge. The A3 scheme can only improve the accuracy of the water level in front of the pumping station, while the A4 scheme can only improve the accuracy of the water level behind the pumping station. This indicates that although the observation stations before and after the pumping station are very close in distance, the data from one station could not be corrected by the data from the other station. This is because unlike rivers, the pumping station can interrupt the

continuity of the water flow in an open-channel water transfer project, and therefore the water levels before and after the pumping station are not significantly correlated with each other.

Impact of interval time on the assimilation results

The interval time may also play an important role in the assimilation of the ensemble Kalman filter model. In this study, five time intervals (2 h, 4 h, 8 h, 12 h, and 16 h) were set, and the prediction errors under different schemes were calculated, as shown in Figure 9.

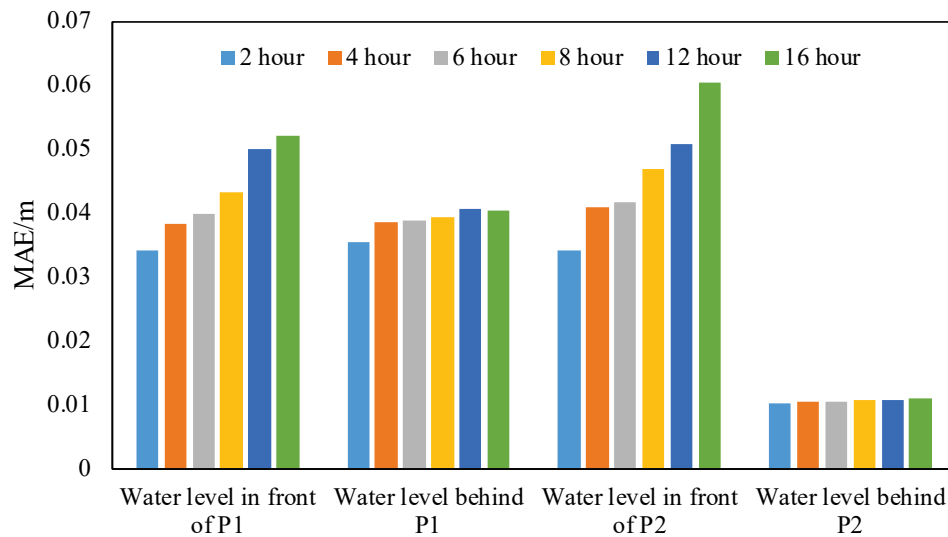


Figure 9 Prediction error of different interval time.

Figure 9 reveals that the prediction error increases as the interval time increases, especially for the water level in front of the pumping station. Compared to the 16-h assimilation, the MAE for the water level in front of P1 and P2 under a 2-h assimilation is decreased by 34.37% and 43.30%, respectively. Similarly, the shorter the time interval is, the longer the calculation time and the lower the calculation efficiency will be. To ensure both calculation accuracy and efficiency, the error threshold can be set according to the need of the project, and assimilation will be done automatically once the predetermined threshold is exceeded. In this section, the water level error threshold is set to 5 cm, and the results after adaptive assimilation are shown in Table 6.

Table 6 Comparison of results using different methods.

Interval time	Water level in front of P1 (m)	Water level behind P1 (m)	Water level in front of P1 (m)	Water level behind P2 (m)	Assimilation number	Calculation time (s)
2 h	0.0342	0.0354	0.0342	0.0102	120	154
4 h	0.0383	0.0387	0.0410	0.0105	60	113
Adaptive	0.0347	0.0370	0.0316	0.0103	77	119

Table 5 shows that the number of assimilation times is 77 with the use of adaptive assimilation strategy, which is 43 less than that under a 2-h assimilation. The computational time is 119 s, which is reduced by 22.73%. However, the MAE for the water level in front of and behind P1 is increased by only 1.44% and 4.32%, and that behind P2 is increased by only 0.97%. The MAE for the water level in front of P1 is increased by 7.60%. Compared to the 4-h assimilation, the assimilation number for the adaptive assimilation strategy is increased by 17 times; the computational time is increased by 5.04%; and the error is reduced by 0.80–22.93%. Notably, the error for the water level in front of P1 and P2 is decreased by 9.40% and 22.93%, respectively. This indicates that the adaptive assimilation strategy can correct the model when there is a significant deviation in the model results, thereby reducing unnecessary assimilation while ensuring the calculation accuracy of the model. It should be noted that this strategy may be suitable not only for predicting water level but also for predicting water quality and temperature. Different assimilation thresholds can be set based on actual engineering needs.

5. CONCLUSIONS

This paper proposes a new prediction model for open-channel water transfer projects with pumping stations based on KNN-ENKF. The KNN algorithm is used to obtain the water regime data along the route that is closest to the current water state as the initial value of the model, which can ensure the prediction accuracy in the initial stage. To reduce the influence of uncertainties on the model, the ensemble Kalman filter model is introduced to correct the model in real time. The main conclusions are as follows.

1. The KNN algorithm can more effectively improve the prediction accuracy of the water level than the interpolation or the steady state start-up method. Compared to IN-OPEN and CF-OPEN, the MAE is reduced by 81.39% and 5.53%, the RMSE is reduced by 80.00% and 4.93%, and the NSE is increased by 2.67 and 0.01, respectively, for the predicted water level in front of P1 obtained by KNN-OPEN; while the MAE is reduced by 61.92% and 60.72%, the RMSE is reduced by 57.16%

and 71.80%, and the NSE is increased by 1.24 and 1.72, respectively, for the predicted water level in front of P2.

2. Data assimilation with different initial values leads to obviously better prediction of the water level, especially the water level in front of the pumping station. After data assimilation, the MAE and RMSE for the water level in front of P1 are decreased by 65.55–91.65% and 64.07–86.59%; and that in front of P2 are decreased by 62.74–83.78% and 63.55–84.13%, respectively.
3. The initial water level has a significant impact on the prediction results, while the impact of the initial discharge can be negligible. When the KNN algorithm is used to obtain the initial value, only upstream and downstream water levels at the current time are selected as the feature vectors. The K -value has little effect on the selection of the initial values. For the sake of computational efficiency, the K -value is taken to be 1.
4. The maximum error for the water level is obtained at an ensemble size of 10, and no significant difference is observed in the error at ensemble sizes greater than 50. To balance prediction accuracy and computational efficiency, the ensemble size of 50 is selected. The presence of a pumping station influences the assimilation effect of two adjacent stations, and data assimilation for one observation station will not improve the prediction accuracy of the other observation station.
5. The prediction errors increase as the interval time increases, especially for the water level in front of the pumping station. The MAE for the water level in front of P1 and P2 under a 2-h assimilation is decreased by 34.37% and 43.30% compared with that under a 16-h assimilation. Compared to a 2-h assimilation, the number of assimilation times is reduced by 43; the model calculation time is reduced by 22.73%; but the error for the water level is increased by 0.97–4.32% with the use of adaptive assimilation strategy.

While the method proposed in this paper can enhance the simulation accuracy, it necessitates the pre-calculation of numerous operating conditions and rich database information. As the project operates over an extended period, model parameters such as roughness may undergo changes, which could lead to a decline in model accuracy. Therefore, both model parameters and states should be corrected simultaneously to further improve the simulation accuracy of the model.

ACKNOWLEDGMENTS

This work was supported by Shandong Provincial Natural Science Foundation (No. ZR2024QE367), the National Natural Science Foundation of China (No. 52209046), and the National Key Research and Development Program of China (No. 2022YFC3204604).

REFERENCES

- Barthélémy, S., S. Ricci, M.C. Rochoux, E. Le Pape, and O. Thual. 2017. “Ensemble-based data assimilation for operational flood forecasting on the merits of state estimation for 1D hydrodynamic forecasting through the example of the “Adour Maritime” river.” *Journal of Hydrology* 552, 210–224. <https://doi.org/10.1016/j.jhydrol.2017.06.017>
- Bourgin, F., M.H. Ramos, G. Thirel, and V. Andréassian. 2014. “Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting.” *Journal of Hydrology* 519 (D): 2775–2784. <https://doi.org/10.1016/j.jhydrol.2014.07.054>
- Cooper, E.S., S.L. Dance, J. Garcia-Pintado, N.K. Nichols, and P.J. Smith. 2018. “Observation impact, domain length and parameter estimation in data assimilation for flood forecasting.” *Environmental Modelling & Software* 104, 199–214. <https://doi.org/10.1016/j.envsoft.2018.03.013>
- Elhenawy, Y., K. Fouad, M. Bassyouni, and T. Majozi. 2023. “Design and performance a novel hybrid membrane distillation humidification–dehumidification system.” *Energy Conversion and Management* 286, 117039. <https://doi.org/10.1016/j.enconman.2023.117039>
- Fu, J-C., M-P. Su, W.C. Liu, W-C. Huang, and H-M. Liu. 2024. “Water Level Forecasting Combining Machine Learning and Ensemble Kalman Filtering in the Danshui River System, Taiwan.” *Water* 16 (23): 3530. <https://doi.org/10.3390/w16233530>
- Gan, Y., Y. Zhang, Y. Liu, C. Kongoli, and C. Grassotti. 2022. “Assimilation of blended in situ-satellite snow water equivalent into the National Water Model for improving hydrologic simulation in two US river basins.” *Science of the Total Environment* 838 (4): 156567. <https://doi.org/10.1016/j.scitotenv.2022.156567>
- Gottardi, G. and M. Venutelli. 2008. “An accurate time integration method for simplified overland flow models.” *Advances in Water Resources* 31 (1): 173–180. <https://doi.org/10.1016/j.advwatres.2007.08.004>

- Gu, L. and X. Lai. 2021. "Influence of field observation on effectiveness of data assimilation using EnKF algorithm for large scale river network." *Journal of Hydroelectric Engineering* 40 (3): 64–75. In Chinese. <https://doi.org/10.11660/slfdx.20210306>
- Han, H., Z. Wang, and B. Liu. 2020. "Tournament incentive mechanisms based on fairness preference in large-scale water diversion projects." *Journal of Cleaner Production* 265, 121861. <https://doi.org/10.1016/j.jclepro.2020.121861>
- He, J., Y. Zhang, and Z. Teng. 2022. "Wind power short-term forecasting system based on K-means and improved KNN algorithm." *Computer Measurement Control* 30 (05): 156–162. <https://doi.org/10.16526/j.cnki.11-4762/tp.2022.05.027>
- Huang, C., A.J. Newman, M.P. Clark, A.W. Wood, and X. Zheng. 2017. "Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States." *Hydrology and Earth System Sciences* 21 (1): 635–650. <https://doi.org/10.5194/hess-21-635-2017>
- Jang, J.H., K.F. Lee, and J.C. Fu. 2022. "Improving River-Stage Forecasting Using Hybrid Models Based on the Combination of Multiple Additive Regression Trees and Runge-Kutta Schemes." *Water Resources Management* 36, 1123–1140. <https://doi.org/10.1007/s11269-022-03077-5>
- Kim, S., H. Shen, S. Noh, D-J. Seo, E. Welles, E. Pelgrim, A. Weerts, E. Lyons, and B. Philips. 2021. "High-resolution modeling and prediction of urban floods using WRF-Hydro and data assimilation." *Journal of Hydrology* 598, 126236. <https://doi.org/10.1016/j.jhydrol.2021.126236>
- Kong, L., Y. Li, H. Tang, S. Yuan, Q. Yang, Q. Ji, Z. Li, and R. Chen. 2023. "Predictive control for the operation of cascade pumping stations in water supply canal systems considering energy consumption and costs." *Applied Energy* 341, 121103. <https://doi.org/10.1016/j.apenergy.2023.121103>
- Lee, H., H. Shen, S.J. Noh, S. Kim, D-J. Seo, and Y. Zhang. 2019. "Improving flood forecasting using conditional bias-penalized ensemble Kalman filter." *Journal of Hydrology* 575, 596–611. <https://doi.org/10.1016/j.jhydrol.2019.05.072>
- Lei, X., Y. Tian, Z. Zhang, L. Wang, X. Xiang, and H. Wang. 2019. "Correction of pumping station parameters in a one-dimensional hydrodynamic model using the Ensemble Kalman filter." *Journal of Hydrology* 568, 108–118. <https://doi.org/10.1016/j.jhydrol.2018.10.062>
- Liang, Y., Z. Mao, W. Zou, and R. Xu. 2018. "Short-Term Traffic Flow Prediction Based on Similar Data Aggregation and KNN with Varying K-value." *Journal of Geo-information*

- Science* 20 (10): 1403–1411. In Chinese.
<https://doi.org/10.12082/dqxxkx.2018.180281>
- Liu, K., Z. Li, C. Yao, J. Chen, K. Zhang, and M. Saifullah. 2016. “Coupling the *k*-nearest neighbor procedure with the Kalman filter for real-time updating of the hydraulic model in flood forecasting.” *International Journal of Sediment Research* 31 (2): 149–158. <https://doi.org/10.1016/j.ijsrc.2016.02.002>
- Lu, L-B., Y. Tian, X-H. Lei, H. Wang, T. Qin, and Z. Zhang. 2018. “Numerical analysis of the hydraulic transient process of the water delivery system of cascade pump stations.” *Water Supply* 18 (5): 1635–1649. <https://doi.org/10.2166/ws.2017.198>
- Lyn, D.A. and P. Goodwin. 1987. “Stability of a General Preisman Scheme.” *Journal of Hydraulic Engineering* 113 (1): 16–28. [https://doi.org/10.1061/\(ASCE\)0733-9429\(1987\)113:1\(16\)](https://doi.org/10.1061/(ASCE)0733-9429(1987)113:1(16))
- Noh, S.J., O. Rakovec, A.H. Weerts, and Y. Tachikawa. 2014. “On Noise Specification in Data Assimilation Schemes for Improved Flood Forecasting Using Distributed Hydrological Models.” *Journal of Hydrology* 519 (D): 2707–2721. <https://doi.org/10.1016/j.jhydrol.2014.07.049>
- Ouellet-Proulx, S., O. Chimi Chiadjeu, M-A. Boucher, and A. St-Hilaire. 2017. “Assimilation of water temperature and discharge data for ensemble water temperature forecasting.” *Journal of Hydrology* 554, 342–359. <https://doi.org/10.1016/j.jhydrol.2017.09.027>
- Patil, A. and R.A.A.J. Ramsankaran. 2018. “Improved streamflow simulations by coupling soil moisture analytical relationship in EnKF based hydrological data assimilation framework.” *Advances in Water Resources* 121, 173–188. <https://doi.org/10.1016/j.advwatres.2018.08.010>
- Raveesh, G., R. Goyal, and S.K. Tyagi. 2021. “Advances in atmospheric water generation technologies.” *Energy Conversion and Management* 239, 114226. <https://doi.org/10.1016/j.enconman.2021.114226>
- Ren, T., X. Liu, J. Niu, X. Lei, and Z. Zhang. 2020. “Real-time water level prediction of cascaded channels based on multilayer perception and recurrent neural network.” *Journal of Hydrology* 585, 124783. <https://doi.org/10.1016/j.jhydrol.2020.124783>
- Tang, H-W., X-K. Xin, W-H. Dai, and Y. Xiao. 2010. “Parameter Identification for Modeling River Network using a Genetic Algorithm.” *Journal of Hydrodynamics* 22, 246–253. [https://doi.org/10.1016/S1001-6058\(09\)60051-2](https://doi.org/10.1016/S1001-6058(09)60051-2)

- Van Wesemael, A., L. Landuyt, H. Lievens, and N.E.C. Verhoest. 2019. “Improving flood inundation forecasts through the assimilation of in situ floodplain water level measurements based on alternative observation network configurations.” *Advances in Water Resources* 130, 229–243. <https://doi.org/10.1016/j.advwatres.2019.05.025>
- Wang, C., J. Yang, and H. Nilsson. 2015. “Simulation of water level fluctuations in a hydraulic system using a coupled liquid-gas model.” *Water* 7 (8): 4446–4476. <https://doi.org/10.3390/w7084446>
- Yan, P., Z. Zhang, Q. Hou, X. Lei, Y. Liu, and H. Wang. 2023. “A novel IBAS-ELM model for prediction of water levels in front of pumping stations.” *Journal of Hydrology* 616, 128810. <https://doi.org/10.1016/j.jhydrol.2022.128810>
- Yan, P., Z. Zhang, X. Lei, Q. Hou, and H. Wang. 2022. “A multi-objective optimal control model of cascade pumping stations considering both cost and safety.” *Journal of Cleaner Production* 345, 131171. <https://doi.org/10.1016/j.jclepro.2022.131171>
- Yan, P., Z. Zhang, X. Lei, Y. Zheng, J. Zhu, H. Wang, and Q. Tan. 2021. “A Simple Method for the Control Time of a Pumping Station to Ensure a Stable Water Level Immediately Upstream of the Pumping Station under a Change of the Discharge in an Open Channel.” *Water* 13 (3): 355. <https://doi.org/10.3390/w13030355>
- Yan, P., Z. Zhang, X. Lei, and P. Xue. 2025. “A new model for rapid identification of unknown water diversion discharge and location.” *Hydrological Sciences Journal* 70 (9): 1454–1463. <https://doi.org/10.1080/02626667.2025.2492231>
- Yu, L., S.K. Tan, and L.H.C. Chua. 2017. “Online Ensemble Modeling for Real Time Water Level Forecasts.” *Water Resources Management* 31, 1105–1119. <https://doi.org/10.1007/s11269-016-1539-8>
- Zhang, C., X.D. Fu, and G.Q. Wang. 2007. “One-dimensional numerical model for unsteady flows in long-route open channel with complex inner boundary conditions.” *South to North Water Transfers and Water Science and Technology* 06, 16–20. In Chinese. <https://doi.org/10.13476/j.cnki.nsbdqk.2007.06.006>
- Zhu, J., Z. Zhang, X. Lei, X. Jing, H. Wang, and P. Yan. 2021. “Ecological scheduling of the middle route of south-to-north water diversion project based on a reinforcement learning model.” *Journal of Hydrology* 596, 126107. <https://doi.org/10.1016/j.jhydrol.2021.126107>