Journal of Hydroinformatics

Journal of Hydroinformatics Vol 25 No 3, 1072 doi: 10.2166/hydro.2023.227

Development of a novel outlier index for real-time detection of water level outliers for open-channel water transfer projects

Luyan Zhou[a], Yu Qiao[b], Zhao Zhang[c], Zhongkai Han[d], Xiaohui Lei[c,*], Yufeng Qin[e] and Hao Wang[c]

[a] School of Resources and Civil Engineering, Northeastern University, Shenyang 110819, China
[b] Construction and Administration Bureau of South-to-North Water Division Middle Route Project, Beijing 100038, China
[c] China Institute of Water Resources and Hydropower Research, Beijing 100038, China
[d] Water Resources Research Institute of Shandong Province, Jinan 250013, China
[e] Yucheng City Water Conservancy Bureau, Yuchen 251200, China
*Corresponding author. E-mail: lxh@iwhr.com

ABSTRACT

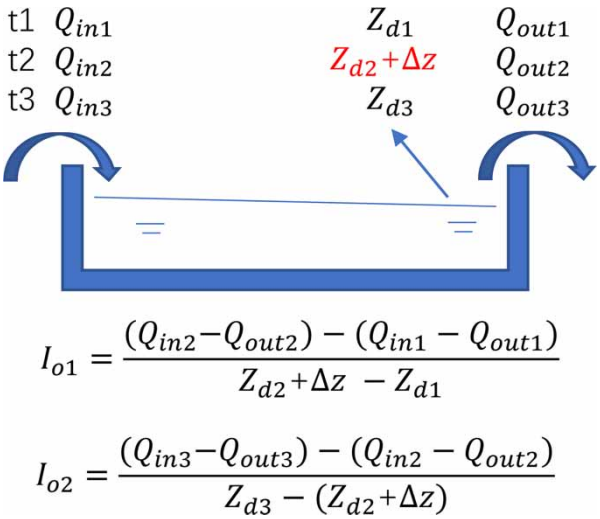
Real-time detection of water level outliers is critical for real-time regulation of gates or pump stations in open-channel water transfer projects. However, this remains a challenging task because of the lack of definition of water level outliers and the imbalance of flow monitoring data. In this study, we define the water level outliers and then propose a highly accurate outlier index for real-time detection of water level outliers based on the water level-flow relationship, and the thresholds for water level outliers are determined based on the order of magnitude of flow and water level differences. A case study is performed with the South-to-North Water Diversion Project of China. A random noise is added to 15 randomly selected non-adjacent monitoring datasets to verify the accuracy of the index, and the noise is increased from 4 to 9 cm at a step of 1 cm. The results show that a total of 159 outliers are detected out of 180 outliers with an accuracy rate of 88.3%.

Key words: open-channel, outlier detection, outlier index, real-time, water level data

HIGHLIGHTS

- An outlier index is proposed based on the water level-flow relationship.
- The definition of water level outliers in open-channel water transfer projects is proposed.
- A case study shows that the proposed method can effectively identify outliers in real time.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Cross-basin water transfer projects are intended to alleviate water shortage and optimize water allocation, and such projects are complex systems characterized by long-distance water transmission with the help of various hydraulic structures such as gates and pump stations. In order to solve the problem of slope damage caused by water pressure difference, the water levels in front of pump stations or gates should be kept as stable as possible (Clemmens *et al.* 2001). Therefore, there is a need for accurate real-time monitoring of water levels in front of pump stations or gates. As the presence of outliers can considerably reduce the quality of the datasets, effective methods are needed to detect these outliers in real time (Boiten 2008; Herschy 2014). Many hydrodynamic models and statistical methods are proposed to offer a potential solution but are not particularly useful in a real-time setting, especially for long-distance open-channel water transfer projects.

The one-dimensional hydrodynamic model based on *Saint-Venant* equations (Yi *et al.* 2017; Zhu *et al.* 2021) can detect outliers in water level datasets by simulating water movement. The *Saint-Venant* equations consist of a continuity equation and a motion equation, where the continuity equation requires the balance between the inflow and outflow of the same channel. However, this is difficult to achieve for long-distance open-channel water transfer projects whose channels are often tens of kilometers long because of abnormal operation of equipment, monitoring errors, rainfall, leakage, and even flow inversion (when the water level is stable, the downstream flow monitoring data points are larger than the upstream ones for a long time), which makes the equation set unsolvable. For this reason, the model is not applicable to real-time detection of outliers in water level datasets.

Statistically, the distribution-based 3-sigma method (Le *et al.* 2013; Hwang *et al.* 2016, 2019; Chen & Liao 2020) and the quantile-based box-plot method (Hubert & Vandervieren 2008; Zhao & Yang 2019) proposed by Tukey (1977) have been widely used to detect outliers in water level datasets. Nevertheless, these two methods do not apply to open-channel water transfer projects because of the masking effect (Rousseeuw & Hubert 2011) and the swamping effect (Rousseeuw & Hubert 2011). The masking effect means that the fitted model cannot detect the deviating observations, and the swamping effect means that some data points are incorrectly identified as outliers due to the presence of another good subset. It is apparent that the former effect may lead to false-negative misdiagnosis (missing of outliers), while the latter effect may lead to false-positive misdiagnosis (the outliers identified are actually not outliers). Such misdiagnoses are attributed to the neglect of the hydraulic correlation between water level changes and flow changes in statistical methods. Because of changes in the water level of open-channel water transfer projects (i.e., continuous short- or long-term changes in adjacent water levels in response to changes in operational requirement or scheduling objective), it is difficult to determine the length and method parameters (e.g., sigma estimation) for detection of outliers. Hence, statistical methods are not applicable to real-time detection of outliers in water level datasets of water transfer projects.

The inconsistency of the definition of water level outliers and the imbalance of flow in water level datasets make real-time detection of water level outliers extremely difficult. Outliers can be defined in different ways (Hawkins 1980; Barnett & Lewis 1994; Chandola *et al.* 2009; Rousseeuw & Hubert 2011; Liu *et al.* 2012), and no consensus is reached for open-channel water transfer projects due to the complexity of water level changes. Very often, datasets are reviewed manually at regular intervals, which is labor-intensive and inefficient and easily causes false or incomplete identification. It is clear that changes in the water levels of a given channel are affected not only by changes in inflow but also by changes in outflow. However, the monitoring of the inflow and outflow of the channel is subject to a variety of uncertainties, resulting in long-term water imbalance in the datasets. Therefore, the key for real-time detection of water level outliers is to process the flow datasets under the influence of uncertainties in an easier and quicker way. Based on the water level-flow relationship, we propose for the first time a novel high-accuracy outlier index for the real-time detection of outliers in water level datasets, which contributes to improving the quality of the water level-flow datasets and provides high-precision data for hydrodynamic simulation, emergency warning, and data mining. Our major contributions are threefold:

1. An outlier index is proposed based on the water level-flow relationship and the thresholds for water level outliers are determined based on the order of magnitude of flow and water level differences.
2. The definition of water level outliers for open-channel water transfer projects is proposed.
3. A case study is presented to demonstrate the effectiveness of the outlier index in the real-time identification of outliers for open-channel water transfer projects.

The paper is organized as follows. Section 2 describes the outlier index and its definition and derivation process; Section 3 presents the study area and the method for outlier detection; Section 4 describes and analyzes the results; Section 5 presents the conclusions.

2. OUTLIER INDEX

2.1. Definition of water level outliers

How the water level outliers are defined is closely related to the accuracy of the detection method. As water levels are generally measured using a water level ruler with a deviation of about 3–5 cm due to the influence of various factors, the measured water levels with a deviation of more than 3 cm from the reading of the ruler are defined as an outlier in this study. However, the manual reading of the water ruler is difficult to obtain, and the simulated value of the water level from the one-dimensional hydrodynamic model (see Section 3.2) is chosen to replace the reading of the water ruler in this paper.

2.2. Classification based on the order of magnitude

The first-order difference (FOD), a special case of the generalized difference method, is used in this study to calculate the changes in the monitoring data at adjacent moments in order to convert a non-smooth dataset into a smooth one (Thanh *et al.* 2022). This method is computationally fast and does not require high computer performance, which ensures real-time (minute level) detection of water level outliers. The FOD value can visually indicate the direction and magnitude of the change of adjacent data points (Shao *et al.* 2020). The calculation formula is as follows:

$$D(V)_n = V(n+1) - V(n) \quad (n \geq 1) \tag{1}$$

where D is the FOD value, V is the type of raw datasets, and n is the sequence.

The water level ruler is generally accurate to the centimeter level. To make the analysis more practical, the raw datasets of water level and flow are processed before the first-order differential calculation, and the values of water level and flow are rounded to two decimals (0.01 m) and one decimal (0.1 m^3/s), respectively. After FOD calculation, $|D(Z)_n|$ and $|D(Q)_n|$ are classified by the order of magnitude, where Z is the water level data point, m; and Q is the flow data point, m^3/s. The classification results are shown in Table 1.

2.3. Derivation of the outlier index

Each channel can be viewed as a small reservoir consisting of an upstream sluice gate (or pump), a downstream sluice gate (or pump), and other hydraulic structures (Figure 1), and its storage capacity follows the water balance principle:

$$V_{t+1} = V_t + (Q^{i-1} + Q_{in} - Q^i - Q_o^i - Q_{loss})\,\Delta t \quad (i \geq 2) \tag{2}$$

where V is the storage of channel $i-1$, m^3; t is the time, s; Q^{i-1} is the inflow at gate $i-1$ (or pump $i-1$), m^3/s; Q_{in} is the inflow of channel $i-1$ from other sources, m^3/s; Q^i is the outflow at gate i (or pump i), m^3/s; Q_o^i is the outflow through all outlets, m^3/s and Q_{loss} is the loss of channel i – 1, m^3/s.

Table 1 | Classification of FOD results of water level and flow

Category	A	B	C	D	E
$\|D(Q)_n\|$		[0.1, 1)	[1, 10)	[10, 100)	[100, 1,000)
$\|D(Z)_n\|$	[0.01, 0.1)	[0.1, 1)	[1, 10)	[10, 100)	[100, 1,000)
Orders of magnitude	-2	-1	0	1	2
Number	0.01	0.1	1	10	100
Scientific notation	10^{-2}	10^{-1}	10^0	10^1	10^2

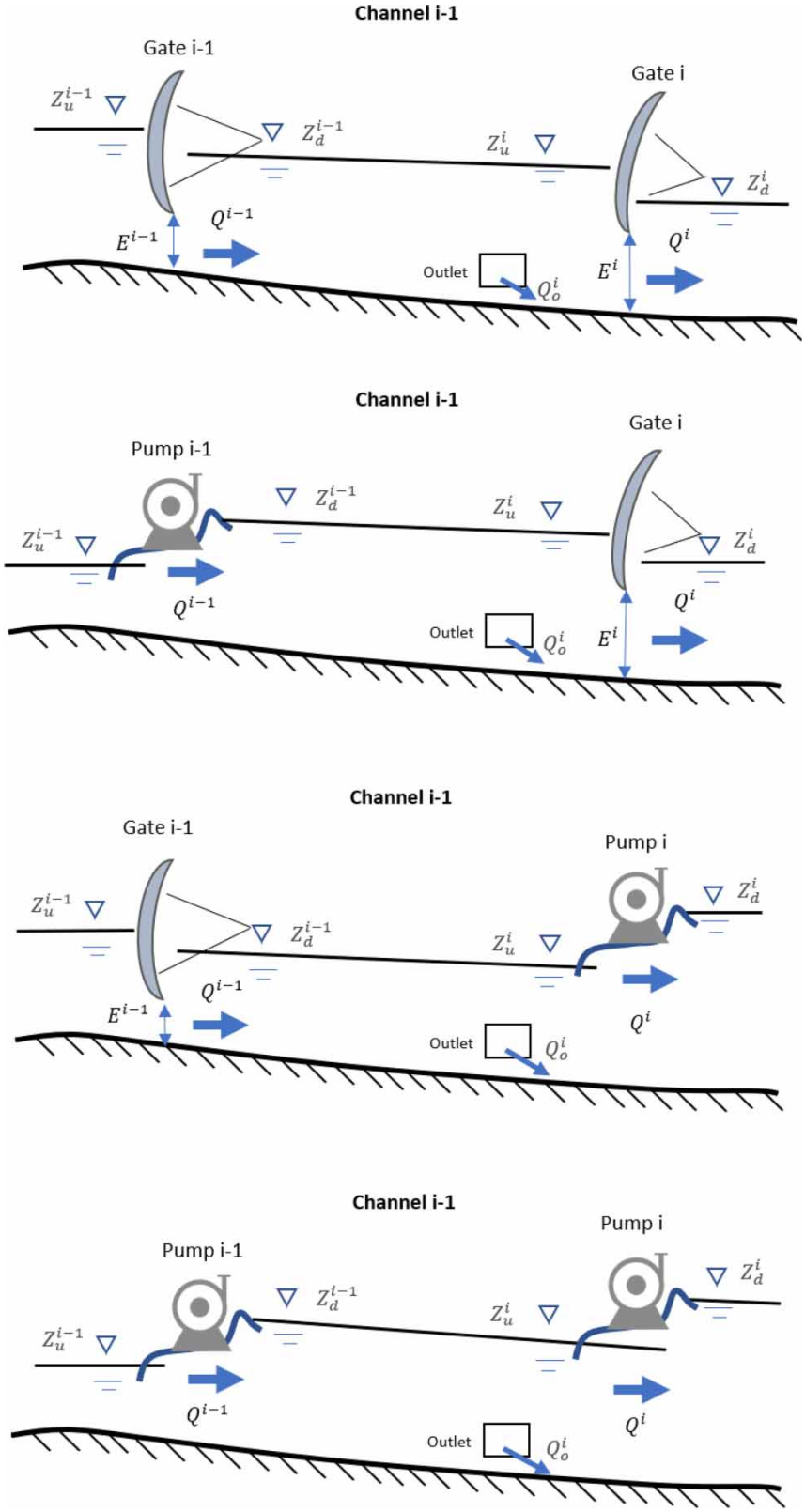


Figure 1 | Schematic of channel structures and hydraulic elements. (1) Upstream and downstream gates. (2) Upstream pump station and downstream gate. (3) Upstream gate and downstream pump station. (4) Upstream and downstream pump stations.

Assume that channel $i-1$ is an ideal body with the same rectangular section. Under ideal conditions (i.e., no head loss and accurate monitoring), the water level change is $\Delta Z = (Q^{i-1} + Q_{\text{in}} - Q^i - Q_{\text{o}}^i - Q_{\text{loss}})\,\Delta t/S_t$. Thus,

$$\frac{S_t}{\Delta t} = \frac{Q^{i-1} + Q_{\text{in}} - Q^i - Q_{\text{o}}^i - Q_{\text{loss}}}{\Delta Z} \quad (\Delta Z \neq 0) \tag{3}$$

where S_t is the water surface area at time t, m^2, and other symbols are the same as above.

In reality, most open channels are prismatic shaped and have irregular sections of different shapes and sizes, and the water surface is sloped due to head loss. Thus, it is difficult to accurately determine the water surface area before and after the water level change. Q_{in} and Q_{loss} are usually not available because they are difficult to monitor. Nevertheless, according to the characteristics of the order of magnitude of the data, the right part of Equation (3) can still be used to detect water level outliers. For channel $i-1$, when $D(Z_u^i)_t$ is not equal to 0, the ratio (I_{o}) of the expected change in water volume to the change in measured water level is obtained:

$$(I_{\text{o}})_{t+1} = \frac{(Q^{i-1} - Q^i - Q_{\text{o}}^i)_{t+1}}{D(Z_u^i)_t} \quad (t \geq 1) \tag{4}$$

where I_{o} is the outlier index, m^2/s; and other symbols are the same as above.

For an open-channel water transfer project, there is an error for the flow difference in Equation (4) due to the unavailability of Q_{in} and Q_{loss}. Even under stable water level conditions, the flow difference fluctuates by an order of magnitude of -1 to 1. To reduce the fluctuation, the difference between inflow and outflow at the current moment can be used as the basis to calculate the flow difference at the next moment. Thus, the calculation of the outlier index should begin when the water level is stable. Assume that the flow difference at the current moment q_t ($q_t = Q_t^{i-1} - Q_t^i - (Q_{\text{o}}^i)_t$) and the inflow and outflow of the open channel at the next moment are changed by a, b, and c, respectively, then

$$\begin{aligned} q_{t+1} &= (Q_t^{i-1} + a) - (Q_t^i + b) - ((Q_{\text{o}}^i)_t + c) - q_t \\ &= a - b - c = D(Q^{i-1})_t - D(Q^i)_t - D(Q_{\text{o}}^i)_t \end{aligned} \tag{5}$$

Hence, the outlier index is calculated as follows:

$$(I_{\text{o}})_{t+1} = \frac{D(Q)_t}{D(Z)_t} = \frac{D(Q^{i-1})_t - D(Q^i)_t - D(Q_{\text{o}}^i)_t}{D(Z_u^i)_t} \quad (t \geq 1) \tag{6}$$

The range of $|I_{\text{o}}|$ is determined depending on the water transfer project. For open-channel water transfer projects, the flow difference is usually several orders of magnitude higher than that of the water level difference. In general, $|D(Z)_t|$ belongs to class A or B. If $|D(Z)_t|$ belongs to class A, then $|D(Q)_t|$ belongs to class B or C. In this case, the corresponding range of $|I_{\text{o}}|$ is 10–90 or 100–900, respectively; similarly, if $|D(Z)_t|$ belongs to class B, then $|D(Q)_t|$ belongs to class C or D, and the corresponding range of $|I_{\text{o}}|$ is 10–99 or 100–999, respectively. When $|I_{\text{o}}|$ is outside the above range, there will be an outlier in $|D(Z)_t|$, but which one, $(Z_u^i)_t$ or $(Z_u^i)_{t+1}$, is the outlier needs to be further determined.

The low-frequency (e.g., hourly) flow data need to be homogenized. For low-frequency flow datasets, the flow data points at adjacent moments may fluctuate substantially due to various causes such as equipment errors and gate or pump regulation. Under stable water level conditions, there would be a small change (<0.05 cm) in the water level at adjacent moments, and then the fluctuation range of the low-frequency flow data at adjacent moments has a large impact on the $|I_{\text{o}}|$ value. Therefore, the data at adjacent moments can be averaged to reduce the fluctuation and improve the identification of outliers.

3. STUDY AREA AND METHODS

3.1. Study area

China is rich in freshwater resources, but the per capita freshwater availability is low (Wang *et al.* 2023) due to the large population and uneven distribution of freshwater resources between north and south (Jing *et al.* 2019; Li *et al.* 2023). The Middle Route of the South-to-North Water Diversion Project (Figure 2) is constructed to divert water from Danjiangkou Reservoir to

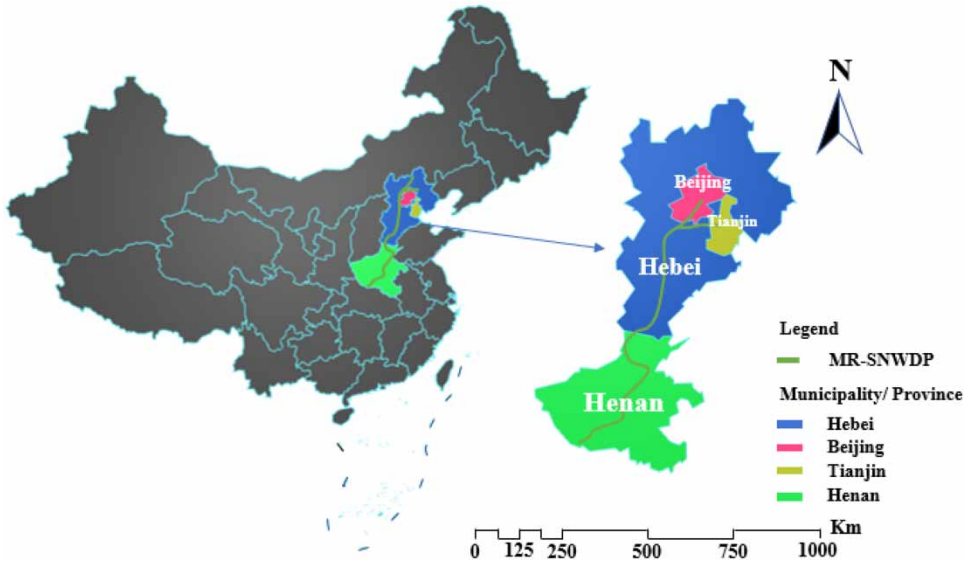


Figure 2 | The map of the Middle Route of the South-to-North Water Diversion Project.

the north in order to alleviate the water shortage in northern China. The operation and scheduling of this project are perhaps the most difficult in the world because (1) there are more than 60 sluice gates (Ren *et al.* 2020) along the 1,277 km-long main route (Cheng *et al.* 2023) with strong coupling between cascaded channels; (2) the regulation capacity of the project is limited because of no water storage facilities along the route; and (3) the control requirements are exceptionally high. For instance, the variation of the water level should not exceed 15 cm within an hour and 30 cm within 24 h (Zhou *et al.* 2022). Thus, real-time detection of outliers is of great significance for the successful water transfer of the project.

This project has been put into operation since 12 December 2014, and now a large amount of water level datasets is available under different climate and operating conditions. Due to security and privacy concerns, we do not have access to minute-level monitoring data. In this study, 15,570 rows of data are collected for each gate for the period 2017–2021 at 2-h time intervals.

The water levels before gate 4, FOD values, and flow differences in channel 3 in 2018 are shown in Figure 3, and the hydrodynamic features of gate 4 with outlets are shown in Table 2.

Figure 3(a) shows that the overall variation of the water level (144.4–144.9 m) before the gate in 2018 is within 0.5 m. It is also noted that the fluctuation of the dataset becomes smoother after FOD calculation (b). In most instances, the changes in water level are mostly less than 0.05 m and symmetrical about 0, and the differences between inflow and outflow of channel 3 are between 0 and 10. This indicates that evaporation and leakage loss during water transfer should not be ignored and that the outlier index formula is reasonable. The 2-h frequency flow data need to be homogenized. In this study, the average of the flow at two adjacent moments is taken as the flow at the latter moment.

3.2. Method

In this section, the simulated results of the 1D hydrodynamic model are used as water ruler data to find the true values of the water level, and a random noise greater than 3 cm is added to some true values. Finally, the outlier index is used for the detection of outliers.

Saint-Venant equations are often used to describe the 1-D flow in open channels:

$$B\frac{\partial Z}{\partial t} + \frac{\partial Q}{\partial x} = q \tag{7}$$

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{\alpha Q^2}{A}\right) + gA\frac{\partial Z}{\partial x} + gAS_f = 0 \tag{8}$$

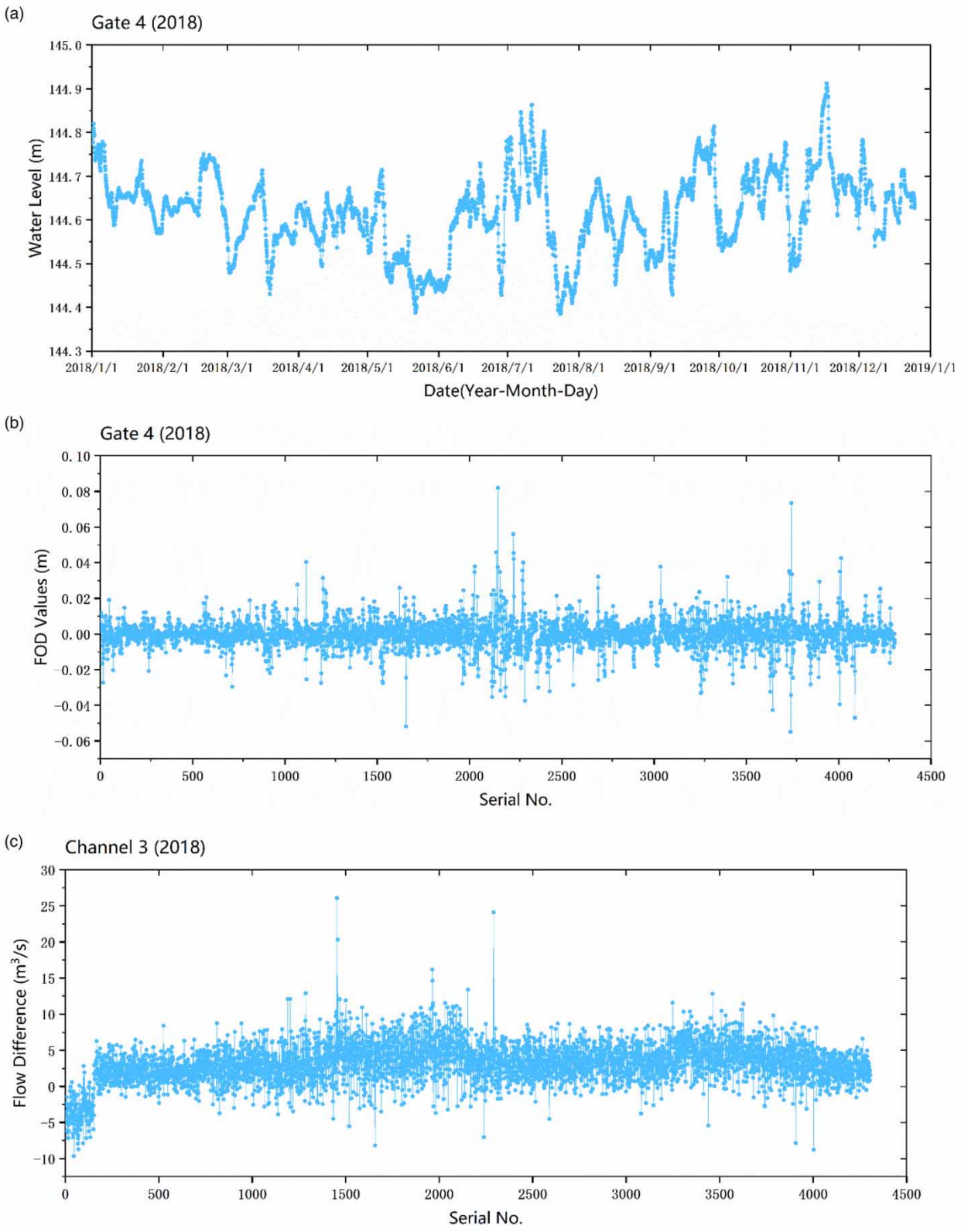


Figure 3 | The water levels before gate 4 (a), FOD values (b), and flow differences in channel 3 (c) in 2018.

Table 2 | The recorded hydrodynamic data of gate 4 in 2018

No.	Date and time	Water level before the gate (m)	Gate-hole 1	Gate-hole 2	Flow (m^3/s)	Outlet 1 (m^3/s)	Outlet 2 (m^3/s)
4	2018-01-01 00:00:00	144.775	1,050	1,080	128.0543	0	0
4	2018-01-01 02:00:00	144.7781	1,050	1,080	125.9459	0	0
4	……	…	…	…	…	…	…
4	2018-07-01 02:00:00	144.7558	1,720	1,720	182.1611	0	0
4	……	…	…	…	…	…	…
4	2018-12-25 14:00:00	144.6307	1,280	1,280	143.2842	0	0
4	2018-12-25 16:00:00	144.6273	1,280	1,280	142.3666	0	0

[1]Null values in the raw datasets are filled with zero.

No.: gates are sequentially numbered from upstream to downstream. Date and time: the date and time of each data record with a time interval of 2 h. Water level before the gate: the values are obtained automatically from the monitoring equipment and are the average of all the gate holes. There are 2 gate holes in gate 4. Outliers in the water level datasets before the gate would be detected. Gate-hole 1, Gate-hole 2: the opening sizes of each gate hole are obtained automatically from the monitoring equipment. Flow: the sum of the flow through each gate hole. The monitoring equipment is usually deployed in front of the gate. Outlet 1, Outlet 2: Outlets are located in channel 3 and the total value indicates the outflow of channel 3.

where B is the channel cross-sectional width, m; Z is the water level, m; t is the time, s; Q is the discharge, m^3/s; x is the distance along the channel, m; q is the lateral inflow, m^3/s; α is the momentum correction coefficient; A is the wetted cross-sectional area, m^2; g is the gravitational acceleration, m/s^2; and S_f is the friction slope, which can be calculated from the following equation:

$$S_f = \frac{n^2 Q|Q|}{A^2 R^{4/3}} \tag{9}$$

where n is the Manning's roughness coefficient, and R is the hydraulic radius of the channel section, m.

In the hydrodynamic model, the water level results of gate 49 for 7 consecutive days (total 85 data points) are used as the upstream boundary condition, and the flow results of gate 50 for the same time series are used as the downstream boundary condition. The model output is the water level data in front of gate 50. The *Preissmann* four-point implicit difference scheme is used to discretize Equations (7) and (8). If the deviation between simulated and gauged data at the corresponding moment is within 3 cm, the 85 monitored data points are considered to be true values.

A random noise greater than 3 cm is added to 15 randomly selected non-adjacent data to construct the water level dataset with outliers. At each time, the noise of the same size is added to 15 data and the size is gradually increased from 4 to 9 cm at a step of 1 cm. After that, the outlier index is calculated.

4. RESULTS AND DISCUSSION

The gauged data of channel 49 from 0:00 on 1 October 2017 to 0:00 on 8 October 2017 is selected for simulation, and the roughness of the channel is 0.0165. The simulated water levels before gate 50 and the differences between simulated and gauged water levels are shown in Figure 4. In order to more accurately describe the deviation between the simulated values of the 1D hydrodynamic model and the gauged values, three decimals (0.001 m) are retained for both values.

Figure 5 shows the differences in the inflow and outflow of channel 49 before and after averaging.

The 15 randomly selected non-adjacent water levels and other data are shown in Table 3. The addition of a noise Δz will change the $D(Z)$ associated with the water level point by the same amount and in the opposite direction. The partial outlier indexes after adding noise are shown in Table 4.

Figure 4 shows that the gauged water level is very stable over the 7 days, and more than half of the data have an FOD value of 0. However, the simulated water level is much less stable, but all the deviations from the corresponding gauged data are within 0.03 m. It can be considered that all the gauged datasets are the true values of the water level. The deviation of the water level at adjacent moments is 0 or 0.01, which means that I_o is very sensitive to the change of $D(Q)$. Thus, it is necessary to homogenize Q. Figure 5 shows that when the water level is in a steady state, the fluctuation of the raw flow difference is relatively large, while the mean-shifting treatment can still retain the overall variation trend while reducing the fluctuation of

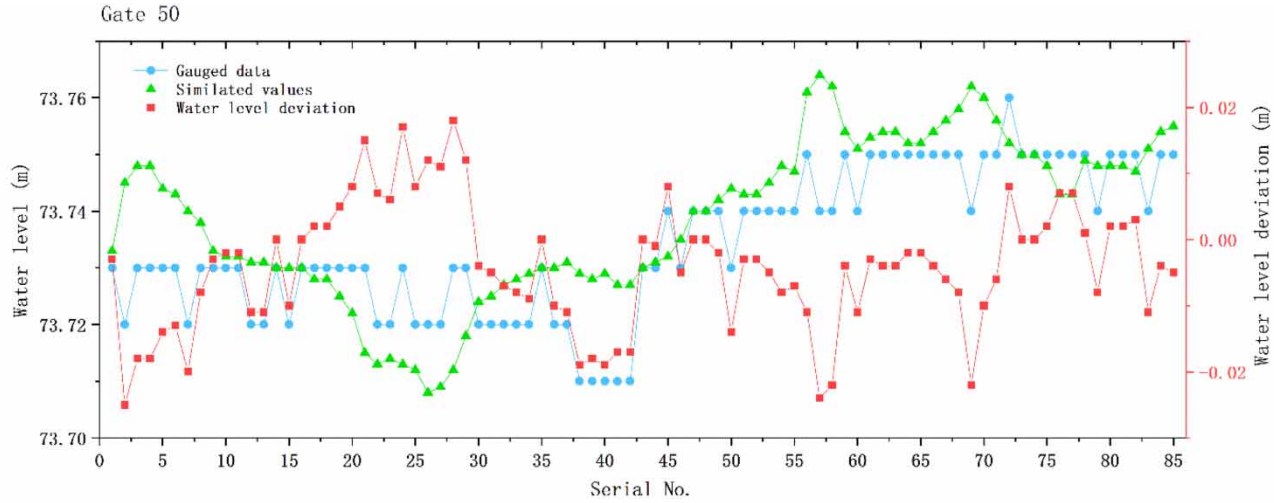


Figure 4 | Simulated and gauged water levels before gate 50 and the differences between them.

the flow difference. Most of the flow difference data lie between −1 and 2, and only three data points (19th, 36th, and 77th) lie between 6 and 8, which are averaged into six points that lie between 3 and 5.

In Table 3, there are 13 numbers with $D(Q)$ values less than or equal to 0.3, which means that I_o is more likely to be lower than 10 if the $D(Z)$ value with the noise is greater than 0.03. Therefore, small amplitude outliers can be easily detected in most data. However, the data like the 5th and 11th data (both $D(Q)$ values are relatively large, or one $D(Q)$ value is 0 and the other $D(Q)$ value is relatively large) are not sensitive to small amplitude outliers. The manual diagnosis is needed to determine whether it is a flow error or a water level error when the outlier index is greater than 90 or 990, or $|D(Z)_t|$ or $|D(Q)_t|$ is 0.

Based on Table 1 and the discussion at the end of Section 2.3, the outliers resulting from the addition of different random noises can be identified according to the outlier index $(I_o)_{t+1}$ or $(I_o)_{t+2}$, as shown in Table 4. As the noise is increased from $\Delta z = 0.04$ m (−0.04 m) to $\Delta z = 0.09$ m (−0.09 m) at a step of 1 cm, a total of 180 outliers are involved. Of these 180 outliers, 159 are detected, 11 (12), 12 (13), 13 (14), 14 (14), 14 (14), and 14 (14), respectively. Since the 2-h dataset is not sufficient enough to characterize the regulation process of gates, and similar to the abnormal 11th data point in Figure 4, the

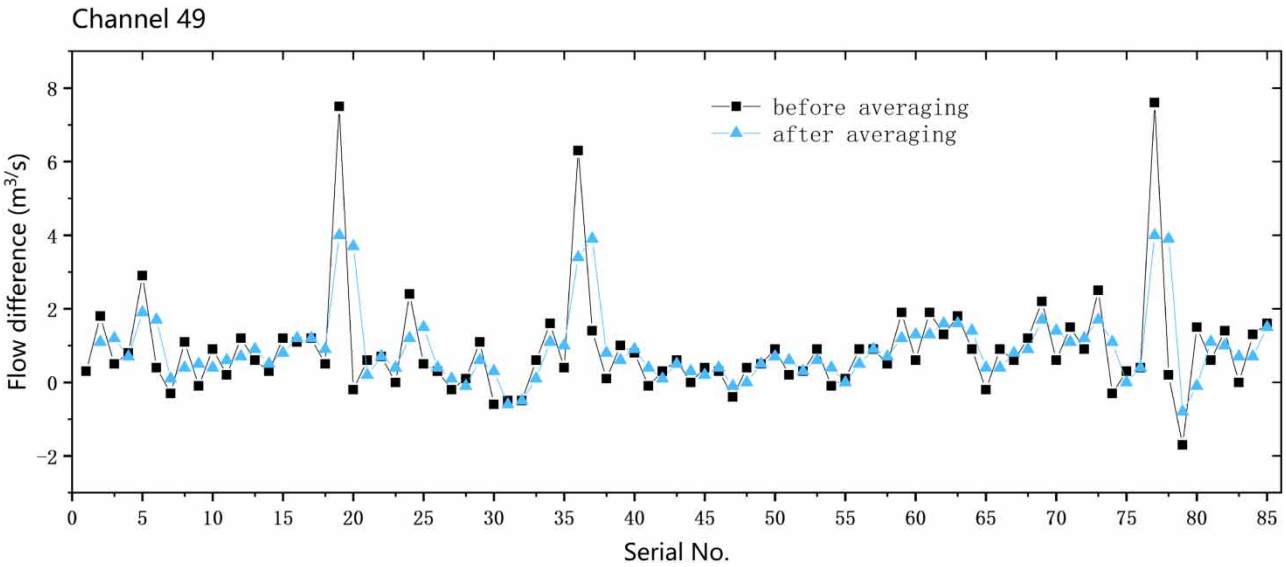


Figure 5 | The differences of inflow and outflow of channel 49 before and after averaging.

Table 3 | Fifteen water level points and other data

No.	Serial No.	$D(Q)_t$ (m³/s)	$D(Z)_t$ (m)	$(I_o)_{t+1}$ (m²/s)	$D(Q)_{t+1}$ (m³/s)	$D(Z)_{t+1}$ (m)	$(I_o)_{t+2}$ (m²/s)
1	7	− 1.6	− 0.01	160	0.3	0.01	30
2	12	0.1	− 0.01	− 10	0.2	0	/
3	18	− 0.3	0	/	3.1	0	/
4	23	− 0.3	0	/	0.8	0.01	80
5	37	0.5	0	/	− 3.1	− 0.01	310
6	44	− 0.2	0	/	− 0.1	0.01	− 10
7	48	0.1	0	/	0.5	0	/
8	51	− 0.1	0.01	− 10	− 0.3	0	/
9	57	0.4	− 0.01	− 40	− 0.2	0	/
10	61	0	0.01	0	0.3	0	/
11	65	− 1	0	/	0	0	/
12	70	− 0.3	0.01	− 30	− 0.3	0	/
13	73	0.5	− 0.01	− 50	− 0.6	0	/
14	77	3.6	0	/	− 0.1	0	/
15	81	1.2	0	/	− 0.1	0	/

manual diagnosis could not be performed and it is temporarily considered as a missed diagnosis. According to Equation (6) and Table 4, when $D(Q)$ is unchanged, the larger the deviation of the abnormal data, the closer the outlier index is to 0, and thus the random noise in Table 4 is representative. The total detection rate of outliers reaches 88.3%, and it increases with the increase of noise.

However, it should be noted that the outlier index may not identify outliers in any one of the following four situations:

(a) $|D(Q)_t| \gg 10*|D(Z)_t|$, and $|D(Q)_{t+1}| \gg 10*|D(Z)_{t+1}|$;
(b) $|D(Q)_t| \gg 10*|D(Z)_t|$, and $|D(Q)_{t+1}| \gg 100*|D(Z)_{t+1}|$;

Table 4 | Water level points with different noises and outlier indexes

		$\Delta z = 0.04$ m		$\Delta z = -0.04$ m		$\Delta z = 0.06$ m		$\Delta z = -0.07$ m		$\Delta z = 0.09$ m	
No.	Serial No.	$(I_o)_{t+1}$ (m²/s)	$(I_o)_{t+2}$ (m²/s)	$(I_o)_{t+1}$ (m²/s)	$(I_o)_{t+2}$ (m²/s)	$(I_o)_{t+1}$ (m²/s)	$(I_o)_{t+2}$ (m²/s)	$(I_o)_{t+1}$ (m²/s)	$(I_o)_{t+2}$ (m²/s)	$(I_o)_{t+1}$ (m²/s)	$(I_o)_{t+2}$ (m²/s)
1	7	− 53	− 10	32	6	− 32	− 6	20	4	− 20	− 4
2	12	3	− 5	− 2	5	2	− 3	− 1	3	1	− 2
3	18	− 8	− 78	8	78	− 5	− 52	4	44	− 3	− 34
4	23	− 8	− 27	8	16	− 5	− 16	4	10	− 3	− 10
5	37	13	62	− 13	− 103	8	44	− 7	− 52	6	31
6	44	− 5	3	5	− 2	− 3	2	3	− 1	− 2	1
7	48	3	− 13	− 3	13	2	− 8	− 1	7	1	− 6
8	51	− 2	8	3	− 8	− 1	5	2	− 4	− 1	3
9	57	13	5	− 8	− 5	8	3	− 5	− 3	5	2
10	61	0	− 8	0	8	0	− 5	0	4	0	− 3
11	65	− 25	0	25	0	− 17	0	14	0	− 11	0
12	70	− 6	8	10	− 8	− 4	5	5	− 4	− 3	3
13	73	17	15	− 10	− 15	10	10	− 6	− 9	6	7
14	77	90	3	− 90	− 3	60	2	− 51	− 1	40	1
15	81	30	3	− 30	− 3	20	2	− 17	− 1	13	1

(c) $|D(Q)_t| \gg 100*|D(Z)_t|$, and $|D(Q)_{t+1}| \gg 10*|D(Z)_{t+1}|$;
(d) $|D(Q)_t| \gg 100*|D(Z)_t|$, and $|D(Q)_{t+1}| \gg 100*|D(Z)_{t+1}|$.

5. CONCLUSIONS

This study may contribute to the detection of water level outliers for open-channel water transfer projects. To our knowledge, this is the first study on the real-time detection of water level outliers for open-channel water transfer projects. The main findings are: (1) an outlier index is proposed based on water balance; (2) the water level outlier is defined for water transfer projects; and (3) a case study with the Middle Route of the South-to-North Water Diversion Project shows that of the 180 outliers with the variation greater than or equal to 0.04 m, 159 outliers can be detected with an accuracy rate of 88.3%.

The limitations of this study warrant further investigation. First, some well-established indicators are excluded due to the lack of high-frequency data. Second, an outlier can be detected in $D(Z)_t$, but which one, $Z(t+1)$ or $Z(t)$, is the outlier needs to be further studied. Finally, the outlier index still has units since water wave motion is not considered, and a dimensionless outlier index should be derived on the basis of momentum conservation in the future.

AUTHOR CONTRIBUTIONS

L.Z., X.L., and H.W. conceptualized the whole article; L.Z. and Z.Z. developed the methodology; Y.Q., L.Z., Z.H., and Z.Z. validated the article; L.Z. and Y.Q. rendered support in data curation; L.Z. wrote the original draft; L.Z. wrote the review and edited the article. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Barnett, V. & Lewis, T. 1994 *Outliers in Statistical Data*. Wiley, New York.
Boiten, W. 2008 *Hydrometry: Ihe Delft Lecture Note Series*. Taylor &Francis, Leiden, The Netherlands.
Chandola, V., Banerjee, A. & Kumar, V. 2009 Anomaly detection: a survey. *Acm Computing Surveys* **41**, 1–58.
Chen, J. & Liao, J. 2020 Monitoring lake level changes in China using multi-altimeter data (2016–2019). *Journal of Hydrology* **590**, 125544.
Cheng, Z., Zhao, Y., Song, T., Cheng, L. & Wang, W. 2023 White elephant or golden goose? An assessment of middle route of the south-to-north water diversion project from the perspective of regional water use efficiency. *Water Resources Management* **37**, 819–834.
Clemmens, A. J., Wahl, T. L., Bos, M. G. & Replogle, J. A. 2001 *Water Measurement with Flumes and Weirs*. International Institute for Land Reclamation and Improvement, Wageningen, The Netherlands.
Hawkins, D. 1980 *Identification of Outliers*. Chapman And Hall, London.
Herschy, R. W. 2014 *Streamflow Measurement*. CRC Press, Boca Raton, FL, USA.
Hubert, M. & Vandervieren, E. 2008 An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis* **52**, 5186–5201.
Hwang, C., Cheng, Y.-S., Han, J., Kao, R., Huang, C.-Y., Wei, S.-H. & Wang, H. 2016 Multi-decadal monitoring of lake level changes in the Qinghai-Tibet plateau by the Topex/Poseidon-Family altimeters: climate implication. *Remote Sensing* **8**, 446.
Hwang, C., Cheng, Y.-S., Yang, W.-H., Zhang, G., Huang, Y.-R., Shen, W.-B. & Pan, Y. 2019 Lake level changes in the Tibetan plateau from Cryosat-2, Saral, Icesat, and Jason-2 Altimeters. *Terrestrial, Atmospheric and Oceanic Sciences* **30**, 33–50.
Jing, L., Zhenxin, B., Cuishan, L., Guoqing, W., Yue, L., Jie, W. & Xiaoxiang, G. 2019 Change law and cause analysis of water resources and water consumption in China in past 20 years. *Journal of Hydro-Science and Engineering* **31**, 31–41.
Le, G., Jingjuan, L. & Guozhuang, S. 2013 Monitoring lake-level changes in the Qinghai–Tibetan plateau using radar altimeter data (2002–2012). *Journal of Applied Remote Sensing* **7** (1), 073470.
Li, Y., Han, Y., Liu, B., Li, H., Du, X., Wang, Q., Wang, X. & Zhu, X. 2023 Construction and application of a refined model for the optimal allocation of water resources – taking Guantao County, China as an example. *Ecological Indicators* **146**, 109929.
Liu, F. T., Ting, K. M. & Zhou, Z.-H. 2012 Isolation-based anomaly detection. *Acm Transactions on Knowledge Discovery from Data* **6**, 1–39.

Ren, T., Liu, X., Niu, J., Lei, X. & Zhang, Z. 2020 Real-time water level prediction of cascaded channels based on multilayer perception and recurrent neural network. *Journal of Hydrology* **585**, 124783.

Rousseeuw, P. J. & Hubert, M. 2011 Robust statistics for outlier detection. *Wires Data Mining and Knowledge Discovery* **1**, 73–79.

Shao, P., Ye, F., Liu, Z., Wang, X., Lu, M. & Mao, Y., 2020 Improving iForest for hydrological time series anomaly detection. In: *Algorithms and Architectures for Parallel Processing* (Qiu, M., ed.). Springer International Publishing, Cham, pp. 170–183.

Thanh, H. V., Binh, D. V., Kantoush, S. A., Nourani, V., Saber, M., Lee, K.-K. & Sumi, T. 2022 Reconstructing daily discharge in a megadelta using machine learning techniques. 58, E2021wr031048.

Tukey, J. W. 1977 *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Wang, Q., Zheng, G., Li, J., Huang, K., Yu, Y. & Qu, S. 2023 Imbalance in the city-level crop water footprint aggravated regional inequality in China. *Science of the Total Environment* **867**, 161577.

Yi, Y., Tang, C., Yang, Z., Zhang, S. & Zhang, C. 2017 A one-dimensional hydrodynamic and water quality model for a water transfer project with multihydraulic structures. *Mathematical Problems in Engineering* **2017**, 2656191.

Zhao, C. & Yang, J. 2019 A robust skewed boxplot for detecting outliers in rainfall observations in real-time flood forecasting. *Advances in Meteorology* **2019**, 1–7.

Zhou, L., Zhang, Z., Zhang, W., An, K., Lei, X. & He, M. 2022 Real-time water level prediction in open channel water transfer projects based on time series similarity. *Water* **14**, 2070.

Zhu, J., Zhang, Z., Lei, X., Jing, X., Wang, H. & Yan, P. 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.

First received 23 December 2022; accepted in revised form 19 April 2023. Available online 4 May 2023