



Study on Intelligent Cleaning of Hydro-logical Data in the Main Canal of the Middle Route of the South-to-North Water Diversion Project

Xiaonan Chen^{1,*}, Yilin Wang¹, Qihao Gu¹, Yanguo Jin¹ and Chunqing Duan²

¹China South-to-North Water Diversion Middle Route Corporation Limited, Beijing, 100038, China

²Government Affairs Service Center of Beijing Municipal Bureau of Water Affairs, Beijing, 100071, China

*Corresponding author's e-mail: chenxiaonan@nsbd.cn

Abstract. Real-time hydro-logical data such as water level and flow of the main canal of the Middle Route of the South-to-North Water Diversion Project are the basis for decision-making of water conveyance scheduling. Due to the influences of external disturbance, measurement system error and other factors, the ill-conditioned hydro-logical data will cause the calculation distortion of the scheduling models, and even lead to calculation failure. Therefore, cleaning the hydro-logical data is necessary. In this paper, aiming at the logical errors in the upstream and downstream flow data space and the jump of the time series of water level data, the water balance model based on particle swarm optimization and the exponential weighted moving average model are established respectively, and the ill-conditioned water regime data is cleaned horizontally and vertically in space and time. Taking the channel section between the Chuanhuang control gate and the Zhang River control gate as a typical research interval, the flow inversion point is automatically identified. The flow data of 12 control gates and 26 water diversion points involved in the channel section are uniformly corrected to realize the rationality of upstream and downstream logic. At the same time, the Yan River control gate in the research section is selected as the representative. Under the basic stable state of operation within 48 hours, the water level data sequence in front of the gate every 2 hours is analyzed, and the jump data is automatically identified and reasonably corrected. The results show that the model established in this paper can automatically identify the ill-conditioned water data and carry out intelligent cleaning. The processed data can better meet the needs of water transfer scheduling analysis and decision-making, and has the value of popularization and application.

Keywords: Middle Route of the South-to-North Water Diversion Project; data cleaning; water dispatching; particle swarm optimization algorithm; exponential weighted moving average model

© The Author(s) 2025

Y. Qiu et al. (eds.), *Proceedings of the 2024 7th International Conference on Civil Architecture, Hydropower and Engineering Management (CAHEM 2024)*, Advances in Engineering Research 256, https://doi.org/10.2991/978-94-6463-650-5_7

1 Introduction

The South-to-North Water Diversion Project is related to the strategic overall situation, long-term development and people's well-being. Since the Middle Route of the South-to-North Water Diversion Project has been in official operation on December 12, 2014, the cumulative volume of water transferred has exceeded 61 billion m^3 , benefiting more than 150 million people along the route, with remarkable comprehensive benefits. Water transfer scheduling is the core business of the Middle Route of the South-to-North Water Diversion Project, and timely and accurate hydro-logical data is the most important basis for real-time scheduling decisions. In addition, hydro-logical data is also the basis of various model calculations in the development stage of intelligent water conservancy. Sick hydro-logical data will make the model calculation results distorted, and even lead to model calculation failure. Therefore, hydro-logical data cleaning is very necessary. Data cleaning research first appeared in the United States for the correction of insurance numbers^[1]. With the popularization and application of big data technology and the generation and development of various models, the accuracy of data is becoming more and more important. Therefore, data cleaning is getting more and more attention. In recent years, data cleaning in various industries^[2-8] has a great development, and the application is more and more extensive.

Hydro-logical data cleaning is the application and expansion of the field of data cleaning in the water industry^[9]. Hou Feng et al.^[10] used the isolated forest algorithm to identify and remove abnormal water quality data, and then the missing data were interpolated based on AdaBoost algorithm, which was better than RF algorithm with higher accuracy. Xue Ping^[11] made a corresponding study on outlier detection and data filling in water level data cleaning of water transfer projects, and the results showed that the filter + 3σ model is more dominant in outlier detection, while the interpolation algorithm is simple, reasonable and suitable for data filling. Chen Zeng^[12] proposed an adaptive denoising method based on empirical wavelet transform and multi-scale fuzzy entropy for various types of noises in water quality collection data, and compared the denoising effect with wavelet transform, similar modal decomposition, complete similar modal decomposition, and empirical wavelet transform, which was more effective, and proposed a combination of migration learning and long and short-term memory model for the problem of missing data, and the experiments showed that the accuracy of the data filling rate was greatly improved. Zhang Jiahong et al.^[13] constructed a three-phase cleaning model of "data preprocessing - outlier detection - vacancy interpolation" for the problem of dirty data in the big data of Shenzhen Nanshan District's smart water system, and the results showed that the average cleaning rate of dirty data reaches 94%. Fu Gui^[14] fully considered the dynamic environment of hydrological data anomaly identification, and proposed an anomaly identification method for hydrological monitoring data based on the random forest algorithm, which was mainly based on the improvement of the random forest for the anomaly value feature extraction. Subsequently, the semantic similarity measure was used to realize the recognition and clustering of anomalous data, and simulation tests showed good results.

The Middle Route of the South-to-North Water Diversion Project has 64 control gates along the route, dividing the main canal into more than 60 sections, forming a

highly hydraulically synergistic "Connecting reservoir group", and the project has set up water level meters and flow meters at control gates and other places for real-time collection of water level and flow rate and other hydro-logical data. The time series of water condition elements may sometimes produce outliers due to factors such as water transmission through the open channel being interfered with by external winds and waves, as well as jumps in the data collected by its own equipment. In addition, due to measurement system errors and other reasons, upstream and downstream water condition data logic problems may occur. For example, the measured flow at the upstream under steady state is 2% larger, while the measured flow at the downstream is 2% larger. Individually, the accuracy of each flow measurement point meets the requirements, but may produce a logical error such as the downstream flow is greater than the upstream flow, leading to the failure of the subsequent scheduling analysis calculations. In view of the above problems, there are now some research results^[9], and the use of the results have been targeted research trial. The results show that the method of cleaning a single gate flow is very good, but in the long-distance and large-scale gate flow data cleaning process, the cleaning of the applicable conditions is relatively harsh. At the same time, water level data was not considered for cleaning. Therefore, this paper utilizes modern artificial intelligence technology to clean the water level element in its own time series, and process the flow element in the spatial logical relationship, forming a cleaning condition broader "vertical" and "horizontal" combination of water data in the Middle Route of the South-to-North Water Diversion Project of the intelligent cleaning method, so that the data can better meet the needs of water transmission scheduling analysis and decision-making.

2 Overview of Typical Study Canals

The canal section between the Chuanhuang control gate and the Zhang River control gate was selected as a typical study interval, which is located in the north of Xingyang City, Zhengzhou City, Henan Province to Cixian County, Hebei Province. In the actual operation of this section of the project, it was found that the frequency of ill-conditioned data in the water condition data of the control gates other than through the Chuanhuang and Zhang River control gate was high. As the north-south demarcation point of the Yellow River in the main canal and the demarcation point of Henan and Hebei respectively, Chuanhuang control gate and Zhang River control gate are important cross sections, which have long been focusing on the rate determination of their flow, and the credibility of the water condition data is high. Therefore, this study selects the Chuanhuang control gate (Pile No. 483+471) as the starting point, and Zhang River control gate (Pile No. 731+366) as the end point of the study section, and the research area is shown in Figure 1.



Fig. 1. Overview of the study area

3 Ill-conditioned Data Analysis and Cleaning Ideas

3.1 Ill-conditioned Data Analysis

In order to support the scheduling and operation management of the Middle Route of the South-to-North Water Diversion Project, water level meters are installed along the project in front of and behind the control gates, in front of the release gates, flow meters are set up at the control gates and diversion outlets, and openness meters are set up at all the gates^[15], with a total of 668 water level meters, 163 flow meters, and 909 openness meters, which are used to monitor the key water data information, such as the water level, flow rate, gate openings, water temperature, and flow velocity of the whole project line. In addition, the monitoring equipment collects relevant data every second, and the monitoring platform pushes it to the integrated management platform every half hour on the principle of forming a huge and constantly expanding database.

However, from many years of scheduling operation in practice, in the scheduler decision-making use of the integrated management platform, from time to time there will be a water condition of the sick data. Take the current artificial scheduling most concerned about the flow and water level data as an example: from the flow point of view, the following problems will generally occur: A non-trending jump in flow occurs at steady state. Flow values are not abnormal, but upstream and downstream flows are inverted. Analyze the causes, there are mainly the following four categories: Firstly, the flowmeter failure, resulting in erroneous readings, or even unable to read the data.

Secondly, the data transmission process produces packet loss and other factors affecting the display of errors. Thirdly, when the external disturbance is strong, it is easy to cause sudden changes in the measurement data, and the duration of a single anomaly is often short and discretely distributed^[16], which can lead to the generation of random errors. Fourthly, the flow meter itself is normal, but due to the existence of systematic errors, resulting in the presence of upstream and downstream flow data inverted phenomena that do not conform to the physical mechanism. In terms of water level data, the problems that often occur with water level values include missing or incorrect water level values, or sudden changes in data due to external disturbances. The causes are similar to the first three categories of causes of flow problems described above. Therefore, this paper argues that in the Middle Route of the South-to-North Water Diversion Project, the water ill-conditioned data can be mainly divided into two categories: one is the vertically ill-conditioned data of its own spatial and temporal problems, which is mainly characterized by the results of single-point data analysis are visualized as obvious errors such as mutated data and null values. The other is horizontal ill-conditioned data with upstream and downstream logic problems, which are characterized as reasonable from the analysis of single-point data, but cannot be analyzed by the reasonableness of data between upstream and downstream. In view of the strong upstream and downstream correlation of the flow data and the obvious influence of the upstream on the downstream, the flow ill-conditioned data are considered to be cleaned by horizontal cleaning methods. Water level data is considered to be cleaned by vertical cleaning method. The detailed relationship is shown in Figure 2.

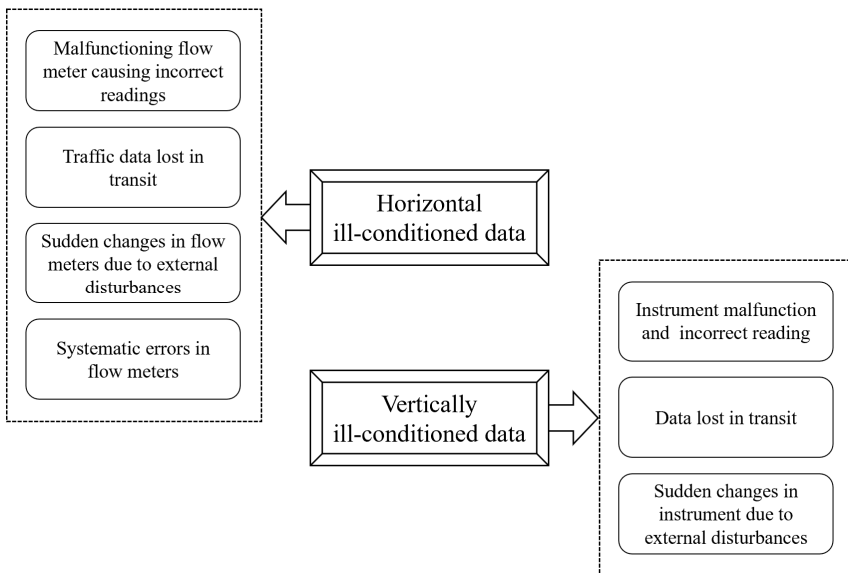


Fig. 2. Ill-conditioned data classification diagram

3.2 Data Cleansing Ideas

In the process of scheduling and operation of the Middle Route of the South-to-North Water Diversion Project, the pre-gate water level and over-gate flow data are the basis for scheduling analysis and command decision-making, and they are also the two most core data. In this paper, we mainly consider the cleaning study of water level and flow data under smooth state.

As for the horizontal ill-conditioned data with upstream and downstream logic problems, it can be seen from Figure 2 above that it mainly exists in the flow ill-conditioned data. Integrate the overflow from the control gates and the diversion flow from the diversion outlets to default to more accurate data from the control gates, especially the flow data from the control gates at the first and last ends. Therefore, with flow balance as a hard constraint and the principle of wide and shallow destruction as an objective function (fitness function), the standard particle swarm algorithm is used to iteratively update the flow data of the diversion outlets that are diverting water during the selected time period (in which the flow rate of the last outlet is not involved in the algorithmic updating). Then, by controlling the error interval of the flow data at the diversion and the rate of water loss along the route, the control gate overflow data are deduced, and the flow rate at the end diversion can be obtained by using the updated upstream control gates to make a difference with the downstream control gates. The detailed process is described below:

Step 1: Build a standard particle swarm algorithm model and set the parameters accordingly.

Step 2: Input based on measured data in the study area.

Step 3: The flow data for the first and last control gates in the study area are considered accurate based on experience, so the flow rate for each control gate is calculated by algorithmically updating the flow rate for the remaining mouths in the study section, except for the last mouth that is diverting water in the selected moment, plus the rate of loss of the canal segment. The last diversion that is diverting water is updated by subtracting the upstream control gate flow and the downstream control gate flow projected for that gate.

Step 4: Determine whether the last diversion that is diverting water in the selected moment of the study segment that is not involved in the algorithm update and constraints satisfies the error interval after the calculation, if not, return to the third step.

Step 5: Output calculated control gate overflow, updated diversion and release gate flows.

For vertically ill-conditioned data that have their own spatio-temporal problems, it can be seen from Figure 2 above that both flow data and water level data will have such problems, but considering the strong upstream and downstream correlation of the flow data itself, the flow ill-conditioned data can be cleaned by utilizing the horizontal ill-conditioned data cleaning method, and it is more effective and in line with the physical mechanism. For the water level data, an exponentially weighted sliding average model is established to automatically identify ill-conditioned data beyond the threshold range, and then corrected. The process is as follows:

Step 1: Establish the exponentially weighted sliding average method model and set the corresponding parameters.

Step 2: Input according to the measured data in the study area.

Step 3: Artificially given error intervals, identify water level data exceeding the error threshold, and clean water level ill-conditioned data exceeding the error intervals.

Step 4: Output the updated water level data that satisfy the error interval.

4 Modeling Process

4.1 Horizontal Ill-conditioned Data Cleaning Model

Standard Particle Swarm Model. Partical Swarm Optimization (PSO) is a swarm intelligence optimization algorithm, originated from the study of bird feeding behavior, proposed by Drs. Eberthart and Kennedy^[17]. PSO algorithm has the shortcomings of easy to fall into the local optimum, and the convergence speed is affected by the inertia weights^[18], but these shortcomings can be avoided by repeated experiments and adjusting the parameters of PSO algorithm^[19]. PSO algorithm has been widely used in reservoir scheduling, water resource allocation, hydraulic modeling parameter research and other aspects and the effect is remarkable^[19-24], the development is more mature.

The PSO algorithm uses constant inertia weights and learning factors with the following velocity and position equations:

$$v_{id}(t+1) = \omega v_{id}(t) + c_1 r_1 (P_{id} - x_{id}(t)) + c_2 r_2 (P_{gd} - x_{id}(t)) \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (2)$$

Where $v_{id}(t)$ denotes the velocity of the first particle in the first dimension of the first generation. $x_{id}(t)$ represents the position of the i -th particle in the d -th dimension and t -th generation. ω is the inertia factor, a constant whose magnitude determines the global search capability. c_1 is the self-learning factor, which determines the local search capability, and c_2 is the social learning factor, which determines the global search capability. The constants r_1 and r_2 which generally take values from 0 to 1 as random numbers. P_{id} stands for the individual optimal solution, and P_{gd} stands for the globally optimal solution.

Objective Function (Fitness Function). The objective function selected in this paper refers to the principle of wide-shallow destruction. The principle of wide-shallow destruction is generally mostly used in the research on water allocation^[25], and its concept is mainly to distribute the water shortage evenly among the water users in order to prevent the large-scale centralized water shortage of each water user under the situation of insufficient incoming water^[26, 27]. In this paper, the objective function is defined as minimizing the ratio of the change in each control gate, diversion gate, and release gate to the sum of the original flow rate, which is given in the following equation:

$$R = \frac{\sum |(Q_f - Q_{fs})| + \sum |(Q_t - Q_{ts})| + \sum |(Q_j - Q_{js})|}{(\sum Q_{fs} + \sum Q_{ts} + \sum Q_{js})} \quad (3)$$

Where R represents the total rate of change. $\sum |Q_f - Q_{fs}|$ represents the sum of the absolute values of the differences between the calculated and measured values at the manifold. $\sum |Q_t - Q_{ts}|$ and $\sum |Q_j - Q_{js}|$ represent the sum of the absolute values of the differences between the calculated and measured values of the release and control gates, respectively. $\sum Q_{fs}$, $\sum Q_{ts}$, and $\sum Q_{js}$ represent the sum of the measured flows at the diversion, the release gate, and the control gate, respectively.

4.1.3. Restrictive Condition.

A. Flow balance constraints

$$Q_{in} - \sum Q_f - \sum Q_t - \sum Q_s = Q_{out} \quad (4)$$

Where Q_{in} denotes inflow to the canal section. $\sum Q_f$ denotes the total diversion flow of all diversion points in the canal section. $\sum Q_t$ denotes the total receding flow of all release gates in the canal section. Q_s denotes the sum of the flow rate of water loss in each control gate section derived from the rate of water loss in the study section. Q_{out} indicates drainage outflow.

B. Error constraints for diversion and release gates

In view of the fact that the flow rates of the diversion and release gates in the selected time period of this paper are not large, it is necessary to set a reasonable error coefficient W to constrain the value of its variation.

$$\begin{cases} W = 0.05 & 0 < Q_f < 1 \\ W = 0.1 & 1 \leq Q_f < 2 \\ W = 0.15 & Q_f \geq 2 \end{cases} \quad (5)$$

C. Velocity and position constraints

$$\begin{cases} v > V_{max} & v = V_{max} \\ v < V_{min} & v = V_{min} \\ x > X_{max} & x = X_{max} \\ x < X_{min} & x = X_{min} \end{cases} \quad (6)$$

4.2 Vertically Ill-conditioned Data Cleaning Models

Vertically ill-conditioned data are modeled using an exponentially weighted sliding average. Exponentially Weighted Moving Averages is one of the basic methods of many algorithms in deep learning, which refers to giving different weights to the observations and calculating the current value based on the last observation and according to the different weights of the historical observations. It is characterized by the fact that the weight of each value in the calculation decreases exponentially over time, and the closer

to the current observation the greater the weight, in other words, the closer to the current observation the greater the impact on the calculation results. The formula is as follows:

$$V_t = \beta^n V_{t-n} + (1 - \beta)(\beta^{n-1} \theta_{t-n+1} + \dots \beta^0 \theta_t)$$

$$\beta = \frac{n-1}{n} \quad (7)$$

Where β represents the attenuation coefficient, which takes values from 0 to 1. θ_t represents the value of the variable V at time t . n Represents the number of historical values.

5 Case Study

5.1 Horizontal Ill-conditioned Data Cleansing

Data Sources. Considering the condition of equilibrium state, it is necessary to select the case that the flow rate of the diversion and release gates in the study section does not vary much on a daily basis, while all the gates south of the Chuanhuang control gate and north of the Zhang River control gate are not operated. The data at 8:00 a.m. on May 10, 2023 was selected as the base data for horizontal ill-conditioned data cleaning, the water level and flow relationship of the control gates is shown in Table 1 below, diversion flows at the diversion and release gates are shown in Table 2, and the flow relationship diagram is shown in Figure 3 below.

Table 1. The flow table of the control gates in the study area

| Number of control gate | Name of control gate | Flow (m ³ /s) |
|------------------------|--|--------------------------|
| 1 | Chuanhuang control gate | 176.71 |
| 2 | Ji River control gate | 174.91 |
| 3 | Yan River control gate | 169.71 |
| 4 | Kuichengzhai River control gate | 171.57 |
| 5 | Yu River control gate | 169.19 |
| 6 | Branch of Huangshui River control gate | 170.99 |
| 7 | Mengfen River control gate | 166.34 |
| 8 | Xiangquan River control gate | 159.30 |
| 9 | Qi River control gate | 156.89 |
| 10 | Tang River control gate | 152.07 |
| 11 | Anyang River control gate | 152.88 |
| 12 | Zhang River control gate | 150.33 |

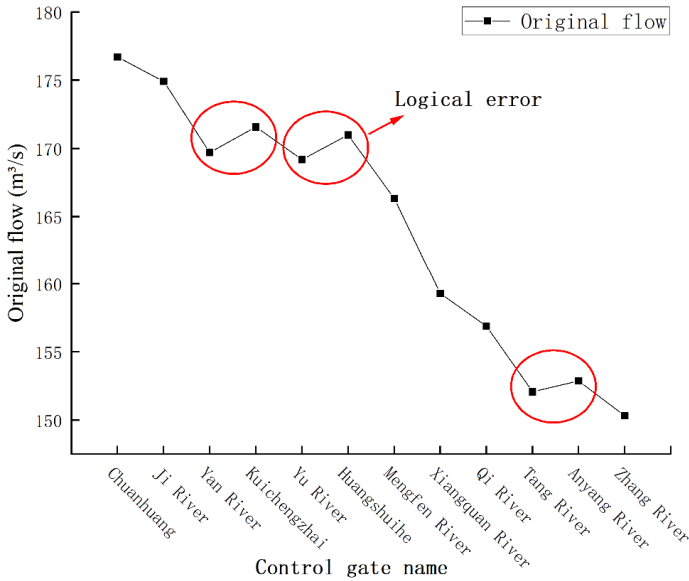


Fig. 3. The flow chart of the control gates in the study area

Table 2. The flow table of the diversion outlets and release gates

| Name | Flow(m³/s) | Name | Flow(m³/s) |
|---------------------------------|------------|------------------------------------|------------|
| Beileng diversion outlet | 0.22 | Laodaojin diversion outlet | 4.17 |
| Beishijian diversion outlet | 0.6 | Wensimen diversion outlet | 0.85 |
| Fucheng diversion outlet | 1.36 | Xiangquan River release gate | 5 |
| Yan River release gate | 0 | Yuanzhuang diversion outlet | 0.5 |
| Li River release gate | 0 | Sanlitun diversion outlet | 6.82 |
| Sulin diversion outlet | 1.34 | Qi River release gate | 0 |
| Kuichengzhai River release gate | 0 | Liuzhuang of Hebi diversion outlet | 0.61 |
| Baizhuang diversion outlet | 0 | Dongzhuang diversion outlet | 1.02 |
| Guotun diversion outlet | 0.32 | Tang River release gate | 0 |

| Name | Flow(m ³ /s) | Name | Flow(m ³ /s) |
|---|-------------------------|------------------------------|-------------------------|
| Yu River release gate | 0 | Xiaoying diversion outlet | 1.33 |
| Branch of Huangshui River release gate | 0 | Nanliusi diversion outlet | 1.92 |
| Lugu diversion outlet | 0.44 | Anyang River release gate | 0 |
| Mengfen River release gate | 0 | Zhang River release gate | 0 |

Calculation of Loss Ratio for Study Section. The loss rate conditions and loss rates for the study section were calculated based on the available water body data as well as the diversions, and the equations and results are as follows:

$$l_z = \frac{w_{sta} - w_{end} - w_{fen}}{w_{sta}} \quad (8)$$

Where l_z represents total loss ratio for the study section. w_{sta} represents the water body at the first moment. w_{end} represents the last moment of the water body. w_{fen} represents the total amount of water diverted and withdrawn during the calculation period. The calculated loss rates for the study segments are shown in Table 3.

Table 3. The results table of loss rate in the study area

| Operating period | Calculation of loss ratio(%) |
|----------------------|------------------------------|
| 2014.12-2015.10 | 5.70 |
| 2015.11-2016.10 | 2.50 |
| 2016.11-2017.10 | 1.53 |
| 2017.11-2018.10 | 0.44 |
| 2018.11-2019.10 | 0.39 |
| 2019.11-2020.10 | 1.31 |
| 2020.11-2021.10 | 2.44 |
| 2021.11-2022.10 | 0.05 |
| 2022.11-2023.05 | 1.29 |
| Average value | 1.74 |

The total length of the study section is 247.90 km, with 12 control gates, and the loss rate is distributed through the distance average, calculated by the following formula:

$$l_n = 1 - (1 - l_z)^{\left(\frac{d_n}{d_z}\right)} \quad (9)$$

Where l_n represents the rate of loss in the canal section between the n-th control gate and the n+1st control gate in the study section ($n=11$). d_n represents the distance of the

canal section between the n -th control gate and the $n+1$ st control gate in the study section ($n=11$). d_z represents the total length of the study segment. The results of the loss rates for the canal sections between the control gates in the study section are shown in Table 4.

Table 4. The results table of loss rate of canals between control gates in the study area

| Number of canal | Name of canal | Flow(m ³ /s) |
|-----------------|---|-------------------------|
| 1 | Chuanhuang - Ji River | 0.13 |
| 2 | Ji River - Yan River | 0.20 |
| 3 | Yan River - Kuichengzhai River | 0.15 |
| 4 | Kuichengzhai River - Yu River | 0.10 |
| 5 | Yu River - Branch of Huangshui River | 0.19 |
| 6 | Branch of Huangshui River - Mengfen River | 0.13 |
| 7 | Mengfen River - Xiangquan River | 0.17 |
| 8 | Xiangquan River - Qi River | 0.21 |
| 9 | Qi River - Tang River | 0.17 |
| 10 | Tang River - Anyang River | 0.20 |
| 11 | Anyang River - Zhang River | 0.10 |

Data Cleaning Results and Analysis. In the flow data of the study section at 8:00 a.m. on May 10, 2023, there were three locations where there were apparent lateral logic errors of downstream flow being greater than upstream flow. Based on the flow equilibrium, the iterative calculations were performed by the standard particle swarm model, which was adjusted by several repetitive experiments, with the iterative parameters listed in Table 5 below, and the results of the calculations listed in Table 6 below:

Table 5. The table of iteration parameters

| Iteration parameters | Value | Number of iterations |
|----------------------|--------------------------------|----------------------|
| ω | 0.8 | 200 |
| $c_1=c_2$ | 2 | 500 |
| $r_1 = r_2$ | Random numbers between 0 and 1 | 1200 |

Table 6. The results table of PSO algorithm

| Number of iterations | Optimal fitness value | Algebra of optimal fitness values |
|----------------------|-----------------------|-----------------------------------|
| 200 | 0.014906 | 118 |
| 500 | 0.014837 | 200 |
| 1200 | 0.014794 | 954 |

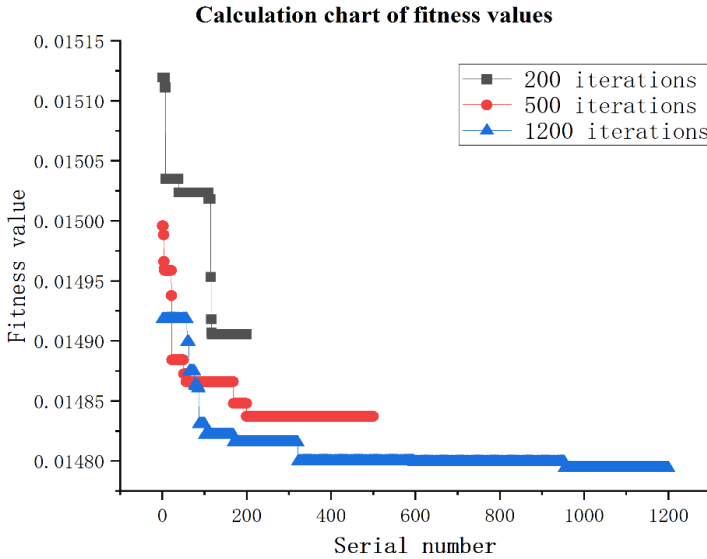


Fig. 4. Fitness value diagram

Figure 4 can clearly see that when iteration 1200 times, the minimum adaptation value is reached at the 954th generation, the convergence effect is good, and it is considered that it has not fallen into premature. In view of the characteristics of the objective function, the optimal solution is taken to be the one with the smallest value of the fitness degree, and the comparison relationship between the overflow and diversion flow of the control gates in the study section and the original flow after updating is shown in Table 7 and Table 8, and Figure 5.

Table 7. The flow table of diversion outlets and release gates in the study area after cleaning

| Name | Flow (m ³ /s) | Percentage of original flow (%) |
|------------------------------------|--------------------------|---------------------------------|
| Beileng diversion outlet | 0.231 | 105 |
| Beishijian diversion outlet | 0.63 | 105 |
| Fucheng diversion outlet | 1.496 | 110 |
| Sanlitun diversion outlet | 1.215 | 90.7 |
| Guotun diversion outlet | 0.336 | 105 |
| Lugu diversion outlet | 0.418 | 95 |
| Laodaojin diversion outlet | 3.545 | 85 |
| Wensimen diversion outlet | 0.808 | 95 |
| Xiangquan River release gate | 4.25 | 85 |
| Yuanzhuang diversion outlet | 0.475 | 95 |
| Sanlitun diversion outlet | 5.797 | 85 |
| Liuzhuang of Hebi diversion outlet | 0.56 | 95 |

| Name | Flow (m ³ /s) | Percentage of original flow (%) |
|-----------------------------|--------------------------|---------------------------------|
| Dongzhuang diversion outlet | 0.918 | 90 |
| Xiaoying diversion outlet | 1.197 | 90 |
| Nanliusi diversion outlet | 1.877 | 97.8 |

Table 8. The flow table of controlling gates in the study area

| Control gate number | Control gate name | Flow (m ³ /s) |
|---------------------|--|--------------------------|
| 1 | Chuanhuang control gate | 176.71 |
| 2 | Ji River control gate | 176.25 |
| 3 | Yan River control gate | 173.77 |
| 4 | Kuichengzhai River control gate | 172.30 |
| 5 | Yu River control gate | 171.80 |
| 6 | Branch of Huangshui River control gate | 171.47 |
| 7 | Mengfen River control gate | 170.84 |
| 8 | Xiangquan River control gate | 161.94 |
| 9 | Qi River control gate | 155.32 |
| 10 | Tang River control gate | 153.56 |
| 11 | Anyang River control gate | 150.48 |
| 12 | Zhang River control gate | 150.33 |

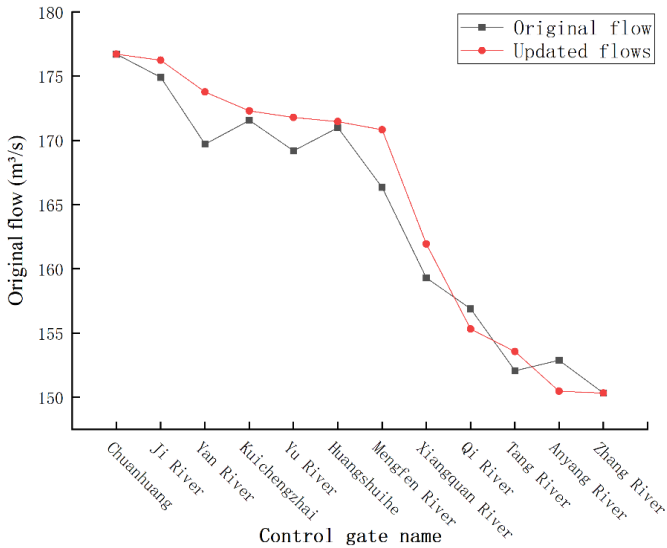


Fig. 5. The comparison flow chart of the control gates in the study area

Verified by example, the standard particle swarm algorithm has obvious effect on lateral ill-conditioned data cleaning. The original actual measurement from the Yan River control gate to the Kuichengzhai control gate, from the Yu River control gate to the Branch of Huangshui River control gate, from the Tang River control gate to An-yang River control gate of the three groups of data with lateral logic error characteristics of the data have been corrected by the cleaning. The updated flows are all within the flow meter error range of the original measurements, the diversion flow corrections are all within the control range, and the study section conforms to the flow balance principle, with good cleaning results.

5.2 Vertically Ill-conditioned Data Cleansing

Data Sources. Considering the equilibrium state, the water level in front of the gate of the sectional gate is relatively smooth or in a slow-change state. In the daily scheduling process, the scheduler generally believes that in the smooth phase of scheduling, the difference in water level between adjacent moments of the same sectional gate should not be greater than 0.03 m, or else it needs to be corrected. Therefore, in the smooth state of the same control gate adjacent moments of large changes (water level difference > 0.03 m) in front of the control gate water level, the need for vertically ill-conditioned data cleaning.

The water level data in front of the Yan River control gate from 8:00 p.m. on July 21, 2023 to 8:00 p.m. on July 23, 2023 were selected as the basis for vertically cleaning. Given that at that time, the South-to-North Water Diversion Central Line Project had entered the stage of high-flow water transfer, with large flow and high water level along the whole line, the Ji River control gates, Yan River control gates, and Kuichengzhai River control gates have been lifted off the water surface and withdrawn from the dispatch, and have been less affected by the dispatch.

Analyze and Calculate. The exponentially weighted sliding average model is applied to identify and clean the water level in front of the Yan River control gate. From Figure 6, the first ill-conditioned data appeared in $n=11$ (July 24, 04:00), therefore, the value of the parameter in equation (7) is $\beta_1 = 0.91$. The second ill-conditioned data appeared in $n=22$ (July 25, 02:00), therefore, the value of the parameter in equation (7) is $\beta_2 = 0.95$. Substituting the parameters in equation (7), the results are calculated as follows in Figure 6.

The results show that after the cleaning of the exponentially weighted sliding average model, the sick water level data are well corrected, the water level change value of the adjacent moment returns to the normal range. In addition, the correlation with the adjacent (previous moment) water level value is large, which is in line with the water transmission scheduling law, and the cleaning effect is good.

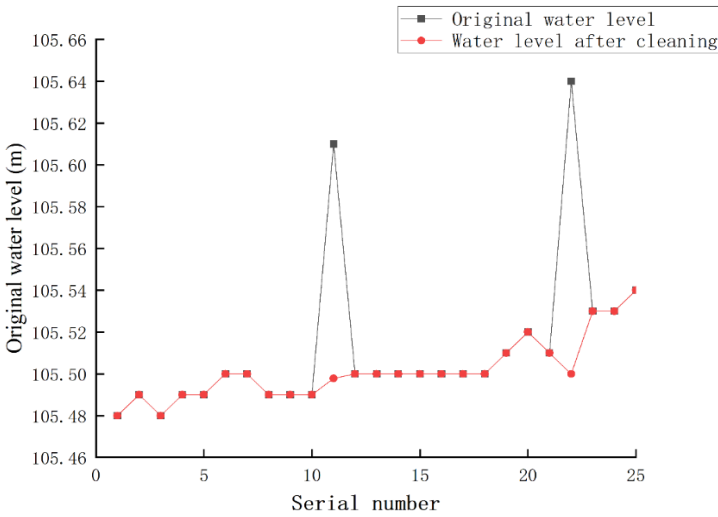


Fig. 6. Comparison diagram of water level of Yan River control gate after cleaning

6 Conclusion

Since the opening and operation of the Middle Route of the South-to-North Water Diversion Project, a large amount of water data has been accumulated. However, with the growth of operation time, equipment and facilities aging and changes in climatic conditions along the trunk canal and other factors, water data in the collection, transmission, display and other aspects of the inevitable error, resulting in ill-conditioned data. In this paper, the causes of water condition data are summarized as the existence of their own spatial and temporal problems of vertical ill-conditioned data and the existence of upstream and downstream logical problems of horizontal ill-conditioned data.

The main purpose of this study is to optimize the ill-conditioned data of flow and water level data, which is most commonly used in water condition data, to a reasonable range, so that the data can better meet the needs of analysis and decision-making in water transmission scheduling, so as to further provide part of the data basis for the establishment of the model in the digital twin of the later central water transmission scheduling in the aspects of hydraulic parameter inversion, hydraulic simulation simulation, and real-time regulation of hydraulic power. Therefore, for the horizontal ill-conditioned data, the standard particle swarm algorithm (PSO), which is widely used, highly mature, relatively easy to model and compute, and effective, was selected and modeled with the principle of wide-shallow destruction as the objective function, and the corrected values of the control gates, diversion gates, and recession gates were taken as the minimum, which is also more more logical, and the results show that: the flow data of the control gates with horizontal logic errors are well corrected, and the updated flow is within the flow meter error of the original measurement value. Meanwhile, the

corrected values of flow at the diversion and release gates are within the control range, the study section conforms to the principle of flow balance, the cleaning effect is good, and the ill-conditioned data return to the range of reasonable intervals. For the existence of their own spatial and temporal problems of vertically ill-conditioned data, a classically effective exponentially weighted sliding average model was selected and the results showed that: The ill-conditioned water level data are all back to the normal range and show a large correlation with the water level value of the previous moment, which is in line with the law of water transmission and scheduling, indicating that the cleaning effect is good. At the same time, under the condition of having long series of historical monitoring data, the data cleaning method proposed in this paper still has some value for popularization and application in similar large-scale water transfer projects and basin water allocation. However, it may encounter the problem of poor results due to the inconsistency of monitoring sources, which can be solved by further filtering the data.

Acknowledgments

This work was financially supported by Water Conservancy Youth Science and Technology Talent Funding Project. The supports are gratefully acknowledged.

References

1. H Galhard, D Florescu, D Shasha and E Simon (2000) An Extensible Framework for Data Cleaning. In: 2000 Proceedings 16th International Conference. San Diego, CA, USA, pp. 312-312.
2. Y J Mei, Y Li, W F Zhou, Y X Guo, W Deng and X B Qiao (2023) Dynamic data cleaning method of abnormal and missing data in a distribution network based on machine learning. *Power System Protection and Control*, 51 (07): 158-69.
3. L Li, Y Liang, N Lin, J Yan, H Meng and Y Q Liu (2023) Data cleaning method considering temporal and spatial correlation for measured wind speed of wind turbines. *Acta Energiæ Solaris Sinica*, 1-9.
4. J Y Wang, Y J Chen, Y Yuan, C Chen and G R Wang (2023) Efficient Data Cleaning Framework for K-Nearest Neighbor Learning Models. *Journal of Frontiers of Computer Science and Technology*, 17(09): 2241-2251.
5. C Y Li (2022) Research on time series data cleaning method based on correlation. Shenyang Aerospace University.
6. L L Yang and H J Hu (2023) Research on comprehensive energy data cleaning based on improved GMM algorithm. *Electronic Measurement Technology*, 46(04):78-83.
7. S Q Li, H W He, P F Zhao and S Cheng (2022) Data cleaning and restoring method for vehicle battery big data platform. *Applied Energy*. 320(4):119292.
8. C L Guo, S F Ji, Y Lin, H S Huang and L L Wang (2022) Method of cleaning TCM data with Aho_Corasick algorithm. *Computer Era*, 03:77-80.
9. W T Wei, Y G Jin, Z Zhang, X H Lei, P Xue and Y L Wang (2022) Application of inverted data cleaning for flow monitoring stations in the middle route of the South-to-North Water Transfer Project. *South-to-North Water Transfers and Water Science & Technology*, 20(06): 1158-1167.

10. F Hou, P Li, H T Pang, Y Tian, F M Chen, Q Q Tian and L L Qian (2023) Water quality monitoring data cleaning method based on AdaBoost algorithm. *Mechanical & Electrical Technique of Hydropower Station*, 46(05):109-111+126.
11. P Xue (2022) Research on Water Level Data Cleaning and Prediction Model of Water Transfer Project. University of Jinan.
12. Z Chen (2023) Research on Water Quality Time Series Data Cleaning and Early Warning in Qiantang River Basin. Dissertation Submitted to Hangzhou Dianzi University.
13. J H Zhang and X H Chen (2021) Building and application of intelligent water system and big data cleaning model in Nanshan district. *Technical Supervision in Water Resources*, 12:32-35+121.
14. Gui Fu (2022) Research on Abnormal Identification of Hydrological Monitoring Data Based on Improved Random Forest Algorithm. *Water Conservancy Science and Technology and Economy*, 28(08):76-80.
15. X N Chen, Y G Jin, X Y Xu and W He (2023) Thinking on smart water dispatching in the South-to-North Water Diversion Middle Route Project. *Journal of Hohai University (Natural Sciences)*, 51(05):46-55.
16. Y Li, X J Shen, Y F Zhang and Y Wang (2023) Cleaning Method of Wind Speed Outliers for Wind Turbines Based on Velocity and Correlation Constraints. *J. Transactions of China Electrotechnical Society*, 38(07): 1793-1807.
17. J Kennedy and R Eberhart (1995) Particle swarm optimization. In: *International Conference on Neural Networks*, Perth, WA, Australia, pp. 1942-1948 vol.4
18. C Y Wu, F L Wang and L Ma (2010) An Improved Particle Swarm Optimization Algorithm. *Control Engineering of China*, 17(03): 359-362.
19. X W Liu, H Wang, X H Lei, W H Liao, M N Wang, W P Wang and P P Zhang (2018) Influence of parameter settings in PSO Algorithm on simulation results of Xin'anjiang model. *South-to-North Water Transfers and Water Science & Technology*, 16 (01): 69-74+208.
20. B L Du (2022) Research on Optimal Allocation of Water Resources in Dali County Based on Simulated Annealing Particle Swarm Optimization Algorithm. *Xi'an University of Technology*.
21. B Y Jia, S Q Wu, Z W Fan, Z K Ma, C Xie and G Q Liu (2018) Application of particle swarm optimization in parameter calibration of channel hydrodynamic model. *South-to-North Water Transfers and Water Science & Technology*, 16(03):143-148.
22. T S Li, R Huang, Z P Sun, S S Guo, K Yi, Y Han and J Chen (2020) Optimizing Water Distribution in Canal Networks Using Multi-objective Particle Swarm Optimization Method. *Journal of Irrigation and Drainage*, 39 (09): 95-100+25.
23. J J Song, H L Zhao and Y Z Jiang (2015) Application of particle swarm optimization in the optimal water allocation of Miyun Reservoir. *South-to-North Water Transfers and Water Science & Technology*, 13(02):378-381.
24. Y H Zhang, P J Han, C Ma, Y H Tao, T Li, Z L Ding, L W Han, X Q Zhang and X L Zhang (2023) Research on back analysis of dynamic parameters of earth-rock dam based on improved particle swarm optimization algorithm. *Water Resources and Hydropower Engineering*, 54 (06): 110-123.
25. J H Li (2006) Study on Water Resources Allocation Model Based on Rule. China Institute of Water Resources and Hydropower Research.
26. X H Ma and Q T Zuo (2007) Regional Scale Initial Water Rights Allocation Model and Application Research. In: *Proceedings of the Fifth China Water Forum*. Beijing, 681-385.
27. R S Wei, Z Q Yan, Z H Zhou, Y Z Jiang and K Wang (2023) Optimal method of reservoir drought-limited water level based on principle of wide and shallow damage. *Water Resources Protection*, 39 (04): 152-8+66.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

