G-RIPS SENDAI 2024

IHI FINAL REPORT

Resilient water management against global warming and for sustainable food supply

Authors:
Mayu Ishikawa[1+]
Takehiro Matsumoto[2]
Kristina Moen[3]
Alaina Stockdill[4]
Wataru Tokonami[5]

Mentors:
Dr. Natsuo Miyatake[*]
Dr. Toshiaki Yachimura[*]
Fumio Hasegawa[**]
Masao Ono[**]

[1] Ochanomizu University
[2] Tohoku University
[3] Colorado State University
[4] University of California Davis
[5] Kyushu University
[+] Project Manager
[*] Academic Mentor, Tohoku University
[**] Industrial Mentor, IHI Corporation

Date of submission: August 8, 2024.
Finalized by industrial mentors: October 16, 2024

Contents

1 Abstract

Water management in canal systems is crucial for a stable water supply [1]. Traditionally, gate operations have been manually controlled by monitoring water levels and relying on the experience and intuition of operators [2]. However, this can be dangerous under severe weather conditions [2]. Furthermore, recent climate changes have made precipitation predictions difficult, resulting in floods and droughts in many regions [3], [4]. To mitigate these issues, automating and optimizing gate operations is essential.

As a first step, this research project will analyze past gate operations using actual observed data on water levels and flow rates. Ultimately, our goal is to achieve a stable water supply through automated gate operations in the future. The data consists of 22 measurement points from a basin with 15 facilities (1 dam, 1 pump, 8 diversion locations, 3 junctions, and 2 spillways) in Shiga prefecture, Japan. Measurements include the flow rate [m^3/s], water level [m], and rainfall [mm], taken every 10 minutes from April 16, 2023, to October 1, 2023.

In this report, we describe data cleaning and processing, gate operation detection, classification of gate operation causes, identification of water shortages, and a concept of automated water management systems. In each process, we conduct a comparative study of between manually detection and mathematical methods, as well as a comparison among representative univariate and multivariate methods, in order to investigate which method is most suitable for the data.

Comparison between several methods indicated that level shift detection can adequately replace manual detection, particularly in identifying gate operations. Furthermore, it is suggested that methods similar to those addressed here show promise for the classification of gate operations. We illustrates our workflow in Figure 3 and summarize our study in a Table (edited by industrial mentors) described in Section 2.3 and Appendix A, respectively.

2 Introduction

2.1 Background

It is predicted that by 2050, more than 40% of the world's population may face severe water shortages [5]. The amount of freshwater available for human use is extremely limited. Freshwater accounts for only 2.5% of the Earth's water [6], with the majority of this freshwater existing as ice or glaciers in regions such as Antarctica and the Arctic [7]. Freshwater in liquid form constitutes about 0.8% of the Earth's total water, and most of this is groundwater. Consequently, only about 0.01% is readily available in rivers and lakes [5]. Currently, we rely on rivers and lakes, and the amount of water available for human use changes depending on precipitation [8]. Therefore, climate change caused by global warming, which results in abnormal weather patterns such as droughts and heavy rains, has a significant impact on the availability of water resources [9].

Global warming affects water resources in two distinct perspectives [10]. The first perspective concerns changes in water demand. As the release of large amounts of greenhouse gases increases atmospheric heat absorption, the temperature will rise [11] and the evaporation from agricultural land will increase [12]. Thus, it is necessary to address the anticipated increase in required water supply.

The second perspective concerns changes in river runoff. Rising temperatures will reduce snowfall and cause it to melt earlier, resulting in changes in irrigation schedules [13]. With the change in river flow, more efficient water use will be required [14]. To adapt to these changes in the demand and supply of water resources, it is considered necessary to review traditional water management practices.

This study analyzes data from an irrigation canal around Lake Biwa, Japan's largest lake located in Shiga Prefecture. This lake supports approximately 17 million people and is one of the most crucial regions for water management in Japan [15]. About 70% of water demand around this area is allocated to irrigation [16]. Thus, making efficient water management in irrigation canals extremely important because irrigation water demand will be uncertain due to anticipated climate change [17].

A water management system consists of a network of irrigation canals and gates that divert water from a river or watershed to meet demand [18]. Figure 1 shows the flow of water along one canal. In irrigation

canals, water flows in the order of primary canal, branch canal, and farmlands. This study focuses on data from one canal, consisting of a dam, canals, and branches that bring water to farmland. Our target canal also contains spillways to handle water overflow and a pump to handle water shortages by pumping water back into the system. Diversion work diverts water flowing along a canal to branches via sluice gates. Branches that bring water to farmland can be further described as "heads" (closest to the main canal) and "tails" (furthest from the main canal). Water flows from the branch head to the branch tail. Junctions combine water flowing along a canal with water from a pump to add water to the system. Of primary interest is to avoid water shortages along the branches with minimal energy use.

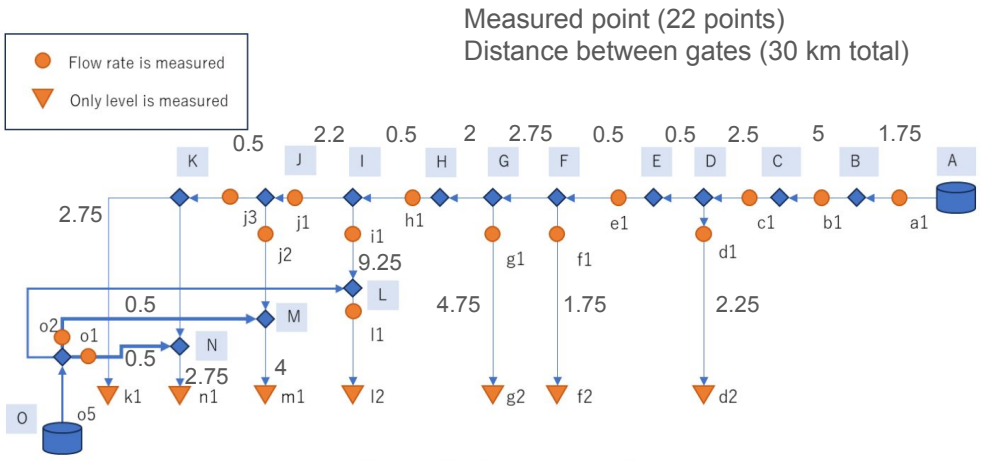


Figure 1: Target canal schematic from [18]. A - O represent facilities, while a1 - o2 are measurement points.

2.2 Preliminaries

2.2.1 Description of Data

The data consists of 22 measurement points taken from the water management system including 15 facilities (1 dam, 1 pump, 8 diversion locations, 3 junctions, and 2 spillways). Figure 1 shows a schematic of the target canal, with the distance in kilometers between facilities provided by industrial mentors for reference.

The values measured by sensors are flow rate, water level, and rainfall (thought not all the measurement points have all 3 values). Flow rate is the water volume at the measurement point and is measured in m^3/s. Water level is measured in m, and rainfall is measured in mm. The measurements are taken every 10 minutes from April 16, 2023, to October 1, 2023 and were taken roughly 24,152 times (though some measurement values are missing).

The canal can be further categorized as primary canals, secondary canals, branch heads, and branch trails. The primary sources of water are a1 (a dam) and o1 and o2 (a pumping station at the furthest downstream point in the primary canal). Only the water level is measured at branch tails. Branch heads and canals may have both flow rate and water level, with flow rate determined by the HQ curve [18] (formally referred to "stage discharge rating curve" [19]). The data does not have information describing gate operation timing and opening degree, demand of the farmland, and water supply from farm ponds and pumps other than o1 and o2.

Figure 2 shows the data of flow rate at c1, water level at d1, and water level at d2. These data are preprocessed to change their scales after replacing negative values to 0. The abrupt drop and subsequent rise at the beginning of June indicate that there was heavy rainfall due to a typhoon. To prevent overflows, all gates were closed and then reopened after the weather condition was improved. In August, the water demands of rice field increased although we had less rain. The stop of water supply means that operators attempted to maintain the water level of the dam.

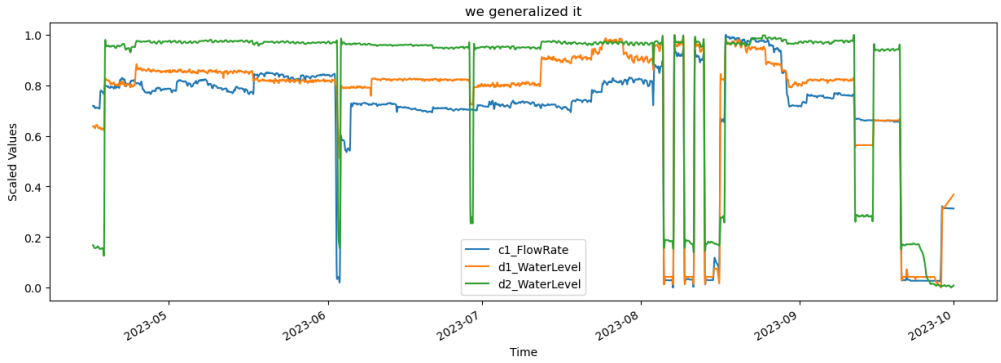


Figure 2: Example of time series data measured by water management system

2.2.2 Assumptions

For simplicity, the following conditions are assumed in this study:

1. Uniform flow
2. Fixed channel geometry (slope and length)
3. No back flows and directed cycles

To satisfy the aforementioned assumptions, particularly the third one, this study mainly focuses on the upstream part of the canal in Figure 1 (facilities A-D), where the waterway network is relatively simple.

2.2.3 Definitions

In this study, several proprietary terms are defined as follows:

Definition 2.1. Gate Operation: A sudden change in the measured quantities (water level and flow rate) that is not considered to be caused by changes in the upstream measured quantities.

Definition 2.2. Classification of Gate Operations: The process of categorizing detected gate operations from data based on the reason for the operation. Gate operations are classified into three reasons: Water Shortage, Surplus, and Constant.

Definition 2.3. Water Shortage: A decrease in the measured quantities that persists for several days before the gate operation and is considered the reason for the gate operation.

Definition 2.4. Surplus: An increase in the measured quantities before the gate operation that is considered the reason for the gate operation.

Definition 2.5. Constant: This term refers to a gate operation where the measured quantities before the operation are almost constant, and the operation is likely due to a scheduled action.

2.3 Objectives

The overall goal of this project is to develop automated methods to optimize water gate operations to minimize water shortages and reduce power usage. Before developing ways to improve the system, it is necessary to understand the workers' current gate operations and identify potential areas of improvement. As the first step towards developing optimal water management, we have two primary objectives:

1. Identify the timing and location of past gate operations based on past flow rate and water level measurements.
2. Classify the cause of past gate operations, i.e. determine whether gates were operated in response to water surplus or shortage.

Furthermore, if time permits, we will work toward another stretch goals.

3. Suggest methods to automate gate operations that eliminate water shortages, reduce power costs (pump usage and gate operation), and reduce water surpluses (flood control).

Figure 3 illustrates the workflow associated with each objective of our study.

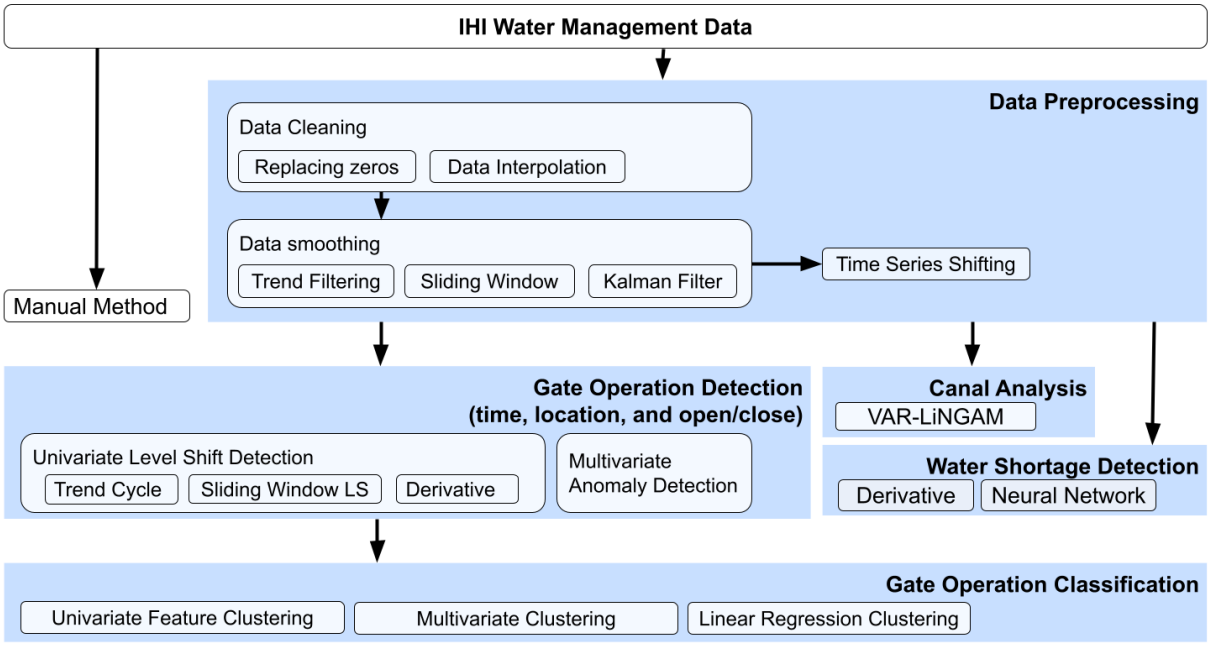


Figure 3: Data Flow. The methods used in this report and the positioning of their objectives. Manual methods were used for both gate operation detection and classification.

3 Methods and Results

3.1 Dataset Preparation

3.1.1 Manual Detection and Classification of Gate Operations

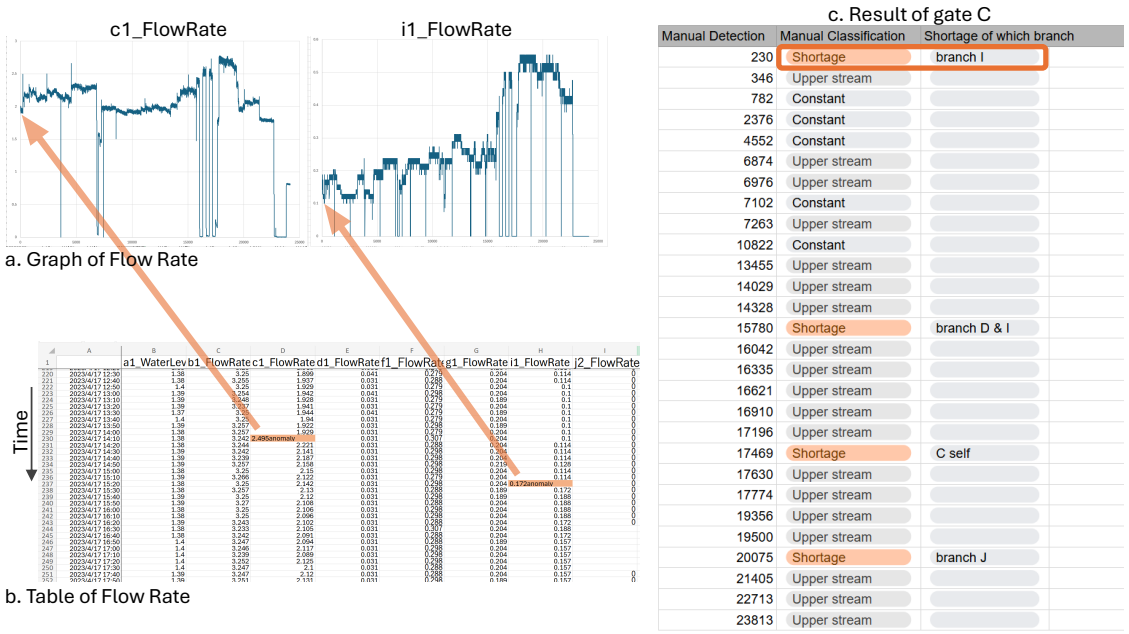
Detection of manual gate operations was carried out using the following procedure. The flow rate data for the dam and all gates (A(a1), B(b1), C(c1), D(d1), F(f1), G(g1), I(i1), J(j2)) were analyzed individually. A graph was created for each gate (Figure 4a), and the times when gate operations might have occurred were recorded (Figure 4b).

For example, on April 17th in Figure 4a, the flow rates of c1 and i1 increased, indicating potential gate operations at gates C and I. However, since gate C is upstream of gate I, the gate I was not operated, and the increase in flow rate at i was naturally due to the operation of gate C. Therefore, the actual gate operation was only at gate C. Observing the graph of i1_flowrate, there is a decrease just before the increase. This indicates that the operation of gate c was performed due to a water shortage at gate I. The same procedure was followed to consider all gate operations and their reasons, as summarized in a table like Figure 4c.

The results were compared with those obtained using other methods for gate operation detection and were used to evaluate the other methods.

3.1.2 Data Cleaning

Prior to time series analysis, it is necessary to clean and smooth the water measurement data. Due to errors in sensor measurement, the data contain erroneous recorded 'zero' measurements, missing observations,



c. Result of gate C

Manual Detection	Manual Classification	Shortage of which branch
230	Shortage	branch I
346	Upper stream	
782	Constant	
2376	Constant	
4552	Constant	
6874	Upper stream	
6976	Upper stream	
7102	Constant	
7263	Upper stream	
10822	Constant	
13455	Upper stream	
14029	Upper stream	
14328	Upper stream	
15780	Shortage	branch D & I
16042	Upper stream	
16335	Upper stream	
16621	Upper stream	
16910	Upper stream	
17196	Upper stream	
17469	Shortage	C self
17630	Upper stream	
17774	Upper stream	
19356	Upper stream	
19500	Upper stream	
20075	Shortage	branch J
21405	Upper stream	
22713	Upper stream	
23813	Upper stream	

Figure 4: Example of manual detection. a: Graph of c1_FlowRate and i1_FlowRate. b: Table that is the basis for a: Graph of Flow Rate. The sections where gate operations might have occurred are highlighted in orange. c: Example of Results. Detected and classified gate operations of gate C by Manual methods.

and fluctuations that distract from the significant and meaningful changes in flow rate and water level (Figure 5). We first removed recorded zeros that we suspect to be errors rather than true measurements and then interpolated the missing data using standard interpolation methods. After the data was cleaned, we compared common methods for data smoothing methods and noise removal and chose one method to use for the rest of our time series analysis.

Replacing Zeros

When flow rate or water level measurement fails, it may be recorded as zero. Such outliers need to be removed for accurate data analysis. We attributed the zero in the case shown in Figure 6(b) to a measurement failure. This figure depicts an instance where in one time step, the measurement value decreases to zero, and in the following time step returns back to the previous value. Since it is unlikely that the water level or flow rate can change by this amount within two-time steps (20 min), we suspect this value is an error. We replace values such as these with 'NaN'.

Interpolating Missing Data

The data contain several missing observations due to sensor-detected anomalies. We apply linear interpolation to predict the missing value from the water values before and after the missing observation. For simplicity, we used linear interpolation[1] as the interpolation method, the algorithm of which is shown in Figure 6(a). The algorithm replaces the missing value with an estimate from a linear equation between the neighboring observations. In this process, we also linearly interpolate the recorded zeros that we suspected to be errors. In the event that a recorded 'zero' was erroneously removed, i.e. the measurement was a true zero, the interpolated value will still be relatively close to zero and hence will not introduce significant error into the measurement.

[1]Other effective methods such as spline interpolation can be mentioned. However, considering factors such as the lack of prior information about the degree, its implementation was not carried out.

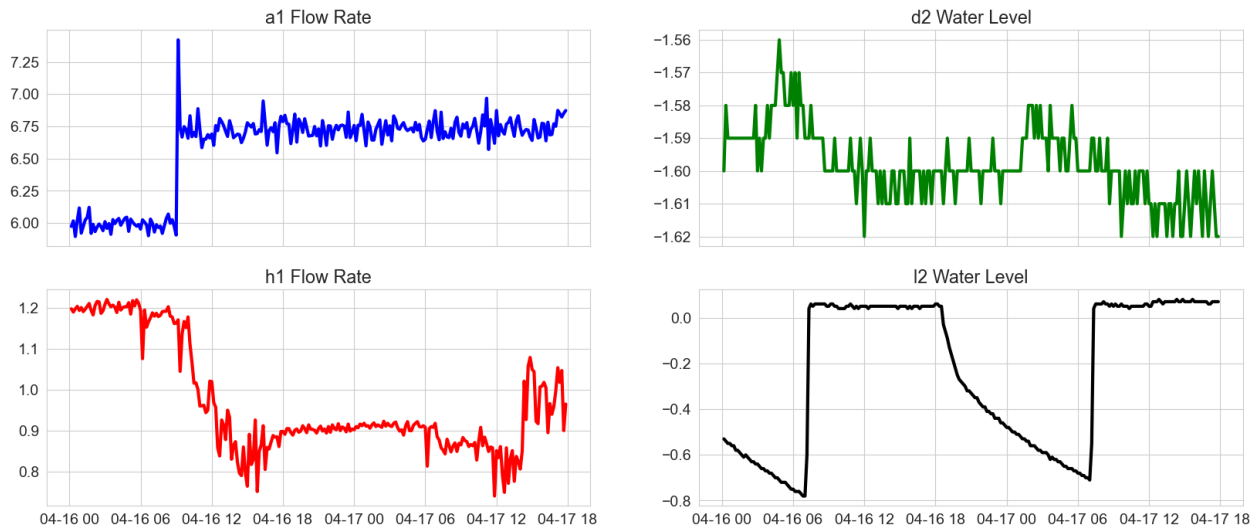
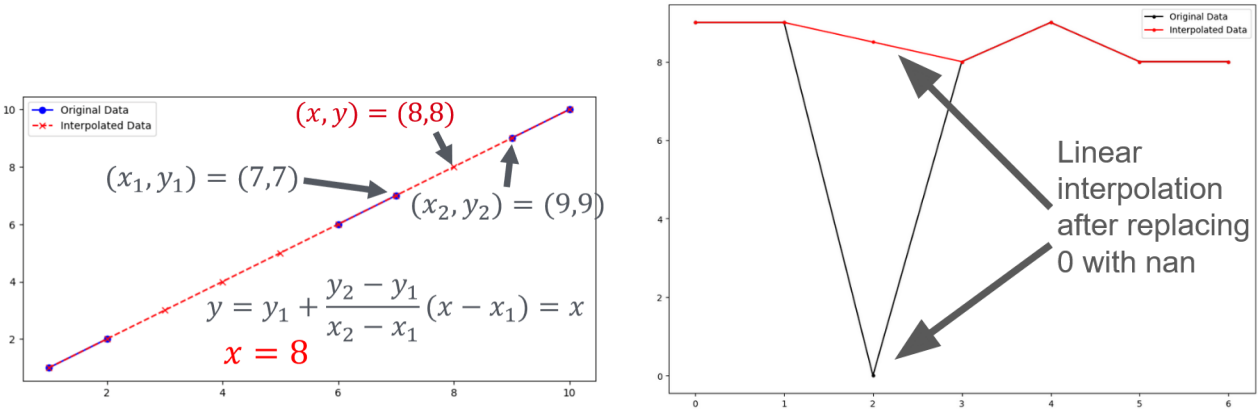


Figure 5: Raw data containing noisy fluctuations at various measurement locations.



(a) (b)

Figure 6: (a) Interpolation algorithm. (b) Replacing Zero algorithm.

Sliding Window Average/Median Filters

3.1.3 Data Smoothing

Trend Filtering

The Hodrick-Prescott filter is used to remove noisy cyclical trends in raw data in order to extract the smooth underlying long-term trend of a time series. We assume the $y_t = \tau_t + c_t$ where τ_t is the long term trend component that is uncorrelated with the noisy cyclical signals, c_t. The algorithm finds an estimate for τ by solving the following minimization problem shown in Equation 3.1.

$$\min_{\tau}\left(\sum_{t=1}^{T}(y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1}\left[(\tau_{t+1} - \tau_t) - (\tau - \tau_{t-1})\right]^2\right) \tag{3.1}$$

The first term, $y_t - \tau_t = c_t$, which represents deviations from the long-term trend and is assumed to have an average of 0 in the long term, is penalized by the first term in Equation 3.1. The λ parameter penalizes the deviations in the growth rate of the trend component by penalizing the sums of squares of its second

difference (where larger values place a higher penalty on the sum of the squares for the long-term trend's second difference) [20]. We implement this algorithm using Python 'statsmodels' [21] for using time series analysis filter,'hp_filter'.

Sliding Window Average/Median Filters

We first considered the sliding window technique for data smoothing. This methods replaces every data value with the average of observations before and after the target data point. For a specified window size, w, an observation $y(t_i)$ at time t_i is replaced by

$$\bar{y(t_i)} = \frac{1}{w} \sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} y(t_j). \tag{3.2}$$

We independently apply this method for each measurement location for series of window sizes ranging from 3 time steps (30 minutes) to 30 time steps (5 hours). In Figure 7, we compare the effect of applying different window sizes to smooth the data. This depicts a trade-off between smoothness and precision when choosing a window size – larger window sizes better minimize the noise yet result in the flow rate beginning to increase sooner and over a longer period of time. If it is sufficient to detect the time of a gate operation with less accuracy, then a larger window size may be preferred to the enhanced noise smoothing.

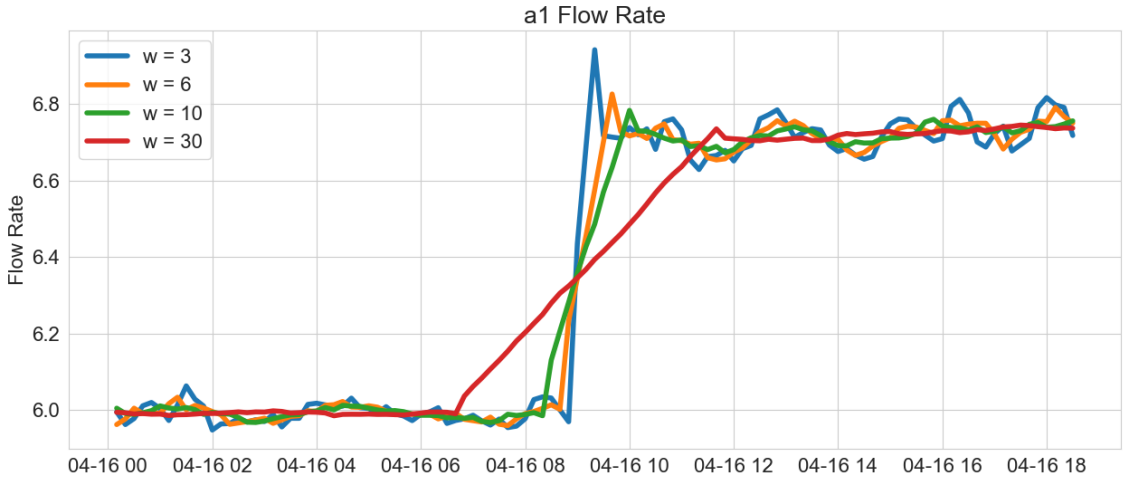


Figure 7: An example of the window size selection noise smoothing using flow rate at measurement site a1.

Rather than using the mean value across a sliding window, we can instead use the median. This can help reduce the issue of causing the slope around changes to decrease but less adequately smooths time series noise.

Kalman Filter

The Kalman filter is an algorithm that is typically used for time series data forecasting given multiple current data measurements. For this problem, we assume that the water measurement state at time t, x_t, can be represented by Equation 3.3 where A represents the transformation of the measurement between time steps and w_k some uncertainty or perturbation in the system. z_k in Equation 3.4 represents the measurement value at time k with added noise v_k and update process H.

$$x_k = Ax_{k-1} + w_k \tag{3.3}$$

$$z_k = Hx_{k-1} + v_k \tag{3.4}$$

It is similar to a predictor-corrector method in that it contains (1) a time update equation that obtains an *a priori* estimate for future time using the current state and error covariance estimates, and (2) a measurement update equation that synthesizes new measurements with the *a priori* estimate to obtain *a posteriori* estimate [22]. Using an initial estimate of x_k and a prior covariance estimate P_t^-, an *a priori* estimate is used to evaluate the Kalman gain (Equation 3.5) to acquire *a posteriori* estimates of the true state and covariance which will serve as our smoothed time series data. We follow the implementation demonstrated by Filippo Bianchi [23].

$$K_k = P_k^- H^T (H P_k^- H^T + R)^{-1} \tag{3.5}$$

Comparison of Smoothing Methods

A visual comparison for each of the smoothing methods is shown in Figure 8. We plot the raw data in grey to help depict the level the effect the smoothing method has on the level of noise and the rate at which flow rate and water level change. Sliding window averaging, shown in blue, was implemented with a window size of 12. This method best captures the steady increase of flow rate after a gate was operated though the increase begins about 1 hour earlier than the raw data shows. The trend filter experiences even more inaccuracies of the timing precision though contains less noise than the other methods. The Kalman filter appears to sufficiently smooth the data as well and seems to retain the timing of the gate operation, though the data take significantly more time to reach the new flow rate/water level value [2].

For the rest of our time series analyses (up to Section 3.4.3), we choose to use the data smoothed with the Hodrick-Prescott trend filter because the objective of classifying the cause of gate operation does not require that we have the precise time a gate was operated.

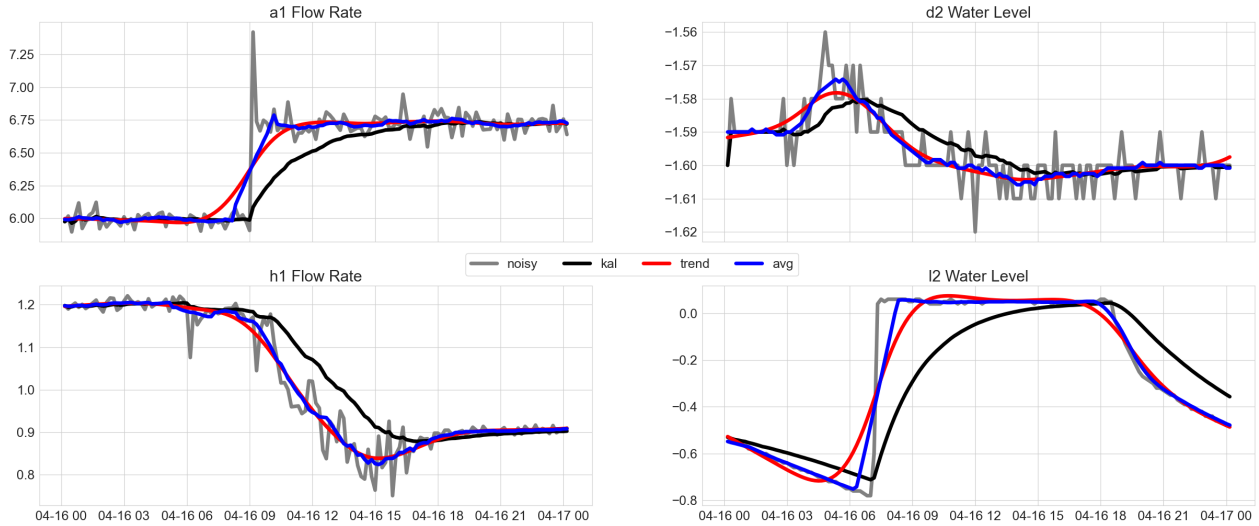
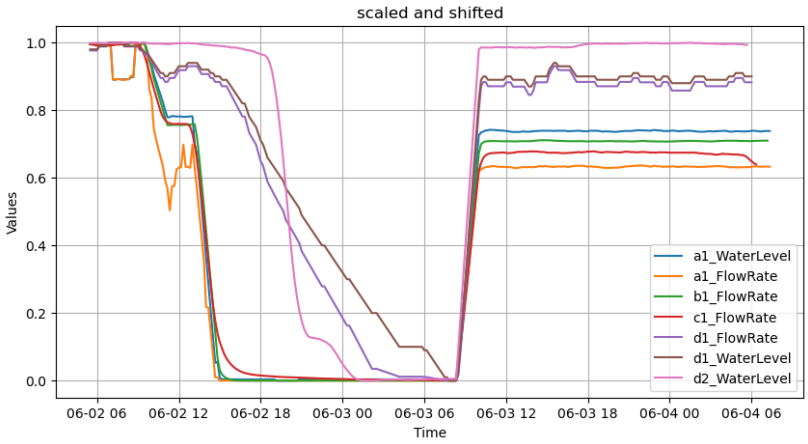


Figure 8: Comparison of smoothing capability of tested filters. For the 4 different measurement locations, we plot the raw data (grey), the Kalman filter (black), the Hodrick-Prescott trend filter (red), and the sliding window averaging (blue).

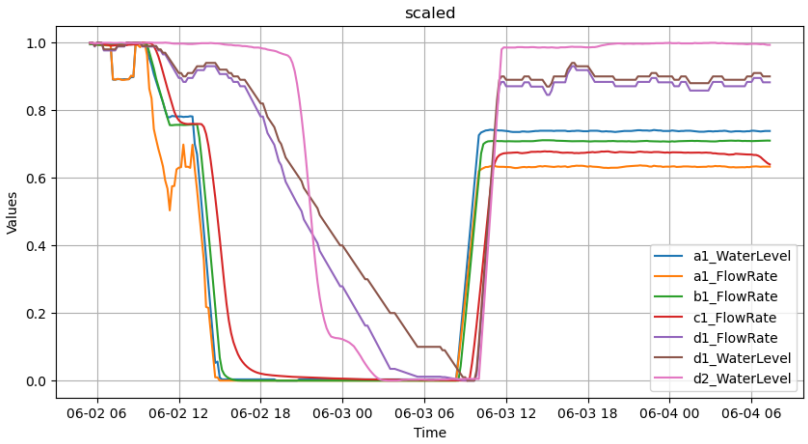
3.1.4 Time Series Shifting

By observing some points, we can estimate delays between each gate. The processed data is shown in Figure 9a. By shifting, we can estimate each delay (see the difference between Figure 9a and b).

[2]This delay was expected to be resolved by using the Kalman Smoother [24]. However, considering the time constraints of the project, its implementation was not carried out.



(a)



(b)

Figure 9: Comparing shifted (a) and non-shifted data (b).

3.2 Detection of Gate Operations

3.2.1 Level Shift Detection

Level shift detection falls under the broader category of change point detection where we attempt to find meaningful or significant changes in the underlying trend in signal data. There are various methods for detecting change points and include several key features such as search method, cost functions, and constraints [25].

For this project, we use a sliding window-based approach for detection level shifts with an l_2 cost function (Equation 3.6). For a given window from t_0 to t_1, the algorithm computes the discrepancy of the cost function between the left and right half of the window. That is, for $t_0 < w < t_1$ with w as the center of the current window, we compute $d(y_{t_0:w}, y_{w:t_1})$ shown in Equation 3.7

$$c(y_{t_0:t_1}) = \sum_{i=a+1}^{t_1} ||y_i - \bar{y}_{t_0:t_1}||_2^2, \tag{3.6}$$

$$d(y_{t_0:w}, y_{w:t_1}) = c(y_{t_0:t_1}) - c(y_{t_0:w}) - c(y_{w:t_1}). \tag{3.7}$$

We obtain a curve where each time value has a discrepancy value given the window size after which a sequential peak search is performed to select the level shifts or change points. We implement this algorithm using 'ruptures' [25] where we estimate the number of level shifts as a hyperparameter. We will use the

number of shifts detected manually as an estimate for the number of peaks to search for (about 32 for each measurement location).

Assuming each gate is operated once a week (four times a month), the amount of this level shift aligns reasonably with the total number of gate operations per facility.

Level Shift Post-Processing

After obtaining the list of level shifts at each location, we need to sort through to determine which shifts actually reflect a gate operated at that measurement site or are the result of an upstream gate being operated.

We first apply time series shifting techniques to line up level shifts with upstream causes. The time index of each downstream measurement is shifted back by the amount of time we estimate the water from facility A to reach that measurement location (described in Section 3.1.4). Then we concatenate our list of possible gate operations across all sites and remove any duplicate events. That is, we will only keep the first time stamp for a group of detected level shifts that occur within 30 minutes. For each of the remaining level shifts, we will determine which sites had a level shift detected around that same time, after which we assume that the most upstream level shift was the gate that was operated and the downstream level shifts were a result of that operation.

An example of the list of gates operated is shown in Table 1. Here we write 'WL' for the water level time series and 'FR' for the flow rate time series. The operated gate was the location of the most upstream measurement that contained a level shift. Note that d2 does not contain a gate and we suspect that times when only d2 experiences a level shift are due to farmland usage.

Time	Gate Operated	a1 FR	a1 WL	b1 FR	c1 FR	d1 FR	d1 WL	d2 WL
2023-04-16 9:20:00	a1	1	1	0	0	0	0	0
2023-04-17 12:40:00	c1	0	0	0	1	0	0	0
2023-04-18 8:40:00	a1	1	1	1	1	1	1	1
2023-04-19 14:00:00	a1	0	1	0	0	0	0	0
2023-04-20 10:50:00	a1	1	1	0	0	0	0	0
2023-04-21 9:10:00	c1	0	0	0	1	0	0	0
2023-04-22 7:50:00	d2	0	0	0	0	0	0	1
2023-04-23 13:50:00	d2	0	0	0	0	0	0	1
2023-04-24 10:10:00	d1	0	0	0	0	1	1	1

Table 1: Example of table containing where level shifts were detected and the assumed gate that was operated. A 1 indicates that the measurement for the respective location experienced a level shift at that time, whereas a 0 indicates there was not a level shift for that measurement data.

3.2.2 Other Methods

Derivative Method

Another method to detect significant shifts in water level and flow rate is to identify points where the change in measurement between time steps exceeds some threshold. Using the smoothed data, we calculate the derivative, or change in value from t_0 to t_1 (Equation 3.8).

$$\Delta y(t_0) = y(t_1) - y(t_0) \tag{3.8}$$

The points where $\Delta y(t)$ exceeds the standard deviation of the derivative for that location will be considered a level shift.

Trend Filter Cycle

Using the trend filtering algorithm described in Section 3.1.3, we can use the cycle information to estimate when the level shifts occur. The method applies the same algorithm as the derivative method, though rather

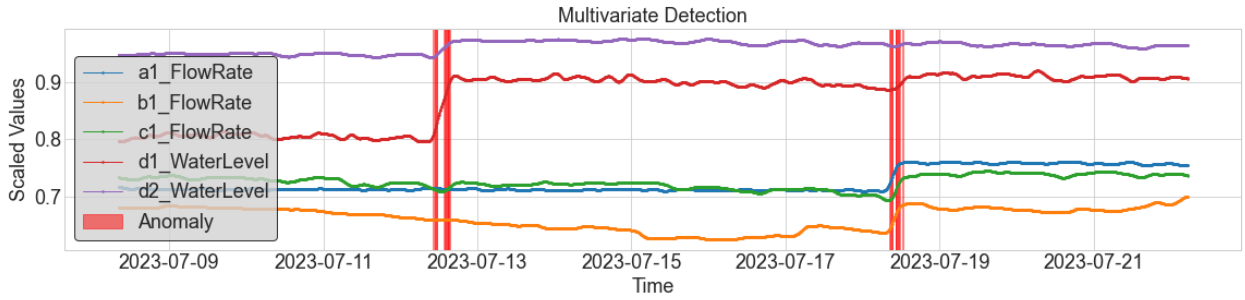
than using the time series data, we apply the derivative method to the cycle data. By definition, the cyclical component $c(t) = y_t - \tau_t$, which is the amount the measurement $y(t)$ deviates form the long term trend, τ_t. Level shifts will occur when the value of $c(t)$ increase or decrease suddenly and rapidly. To evaluate where level shifts occur in the long term trend occur, we can calculate the difference between each time step of c_t and located the time points where the cyclical component exceeds some threshold.

Multivariate Detection

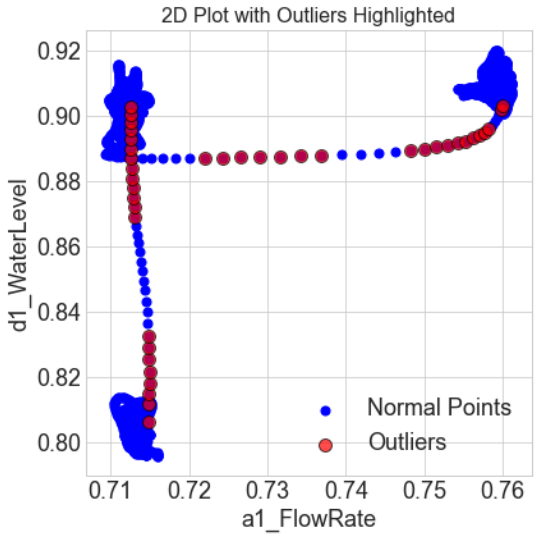
This method detects system-wide anomalies rather than anomalies at individual locations. The goal of multivariate anomaly detection is to find violations of the usual relationship between measurements at individual locations. The input for multivariate anomaly detection are multiple time series, and the output is specific indices (times) that are considered outliers of the system. For example, since our locations are connected via one-directional waterways, we expect that when flow rate increases upstream, it also increases at downstream measurement locations. Multivariate anomaly detection finds violations to this usual relationship.

We use ADTK OutlierDetector algorithm for this method [26]. We start with an input of 5 time series from measurement locations at facilities A-D in our target canal. Each index is converted to a 5-dimensional point, where the coordinates consist of all the measurements at that specific index. Next, we compute the local density of each point, which is estimated by the typical Euclidean distance at which a point can be "reached" from its k-neighbors. For our experiments, we used $k = 20$ neighbors, which is the default for the ADTK algorithm. We then assign a score to each point, which measures the deviation of the local density of each point to its k-neighbors. We identify points that have a significantly lower local density than its neighbors as outliers. Note that "neighbors" is not defined by time between indices or space between measurement locations, but by Euclidean distance between our new 7-dimensional points.

Figure 10 shows the algorithm run on one month of data from the A-D canal. Notice that anomalies correspond with changes in the time series relative to the others. For example, the first anomaly occurs when d1_WaterLevel and d2_WaterLevel increase while the other measurements stay roughly constant or decrease. The anomalies tend to occur in groups which may signal both the start of an event (gate change) and the end of an event (measurements return to an equilibrium where all measurements are constant). Figure 10(b) shows that when measurements stay roughly constant, they tend to form clusters (note that that the figure only shows 2 out of 5 coordinates). However, when one or several measurements begin to change, this can create outliers from the clusters. This method can also be used to track violations of the usual relationship between heads and tails of irrigation canals as in Figure 11, which can reveal water shortages or water usage patterns.



(a)



(b)

Figure 10: Multivariate anomaly detection. (a) Detected anomaly (red vertical line) in A-D canals over one month (b) Plotting two of the coordinates as points where a1_FlowRate is the x-coordinate and d1_WaterLevel is the y-coordinate. The outliers have a significantly lower density that their neighbors. The true points have 5 coordinates and cannot be visualized.

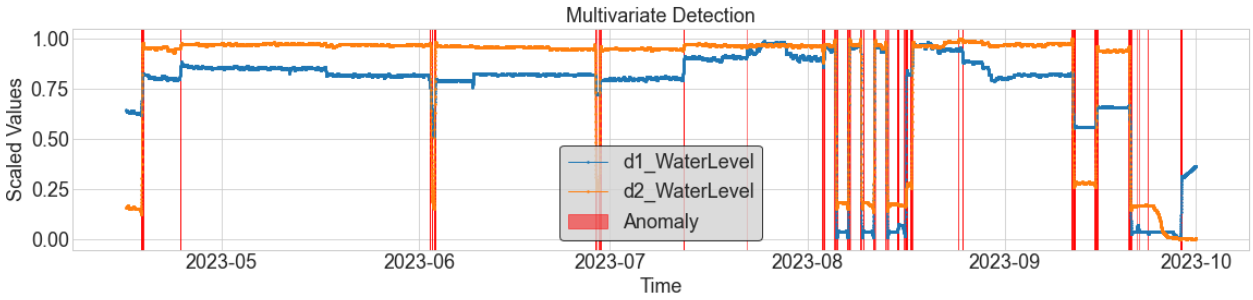


Figure 11: Multivariate anomaly detection on d1_WaterLevel and d2_WaterLevel (head and tail of irrigation canal through farmland) over full time

Gate Detection Method Comparison

After gathering sets of indices for each of the methods for level shift detection, we apply post processing methods to group together the indices that are a part of the same level shift or 'event'. This is done to more

Metric	Location	Method			
		Level shift	Derivatve	Trend cycle	Multivariate
Suceessful Detection	a1	31 (96%)	22 (69%)	26 (81%)	-
	b1	17 (71%)	11 (46%)	13 (54%)	-
	c1	27 (96%)	16 (57%)	20 (71%)	-
	d1	24 (89%)	18 (67%)	17 (63%)	-
	all	-	-	-	64%

Table 2: Number of correct and incorrect level shift detection points by location and method.

easily compare the detected level shifts with the manually detected level shifts. The 'ruptures' level shift detection algorithm outputs a single time stamp per estimated level shift. The other methods (derivative, trend cycle, and multivariate methods), however return several timestamps that all correspond to the same level shift or 'event'. For example, since the value of flow rate and water level will increase over several time steps after a gate operation, the magnitude of the derivative and $c(t)$ value, will be elevated for several time steps. Similar to level shift grouping methods described in Section 3.2.1, we iterate through the list of timestamps and group together points that are within 20 time-steps. Only the first index of each group is kept as the estimated level shift.

To evaluate the accuracy, we iterate through the manually detected level shifts and determine whether or not that level shift was detected by each algorithm. We count how many level shifts were correctly detected and how many were incorrectly detected, i.e., how many shifts were detected that did not appear in the manually detected list.

Since there is variation in the exact time that the level shift was detected, we will search for any level shifts in a range around the manually detected level shift (20-time step interval). We combine the list of detected level shifts at each location. For example, since we apply each method to each time series separately, we combine the shifts for a1 flow rate and a1 water level since they are related to facility A. We do the same for d1 flow rate and water level. The number of level shifts that were successfully detected using each method is shown in Table 2. By applying the 'ruptures' level shift detection algorithm and specifying an estimate for the number of gates we suspect to have occurred throughout the time series data, we obtain a list of gate operations.

3.3 Classifying Cause of Gate Operation

After detecting the locations and times that gate operations occur, we classify the cause of a gate operation as water shortage, surplus, or scheduled (i.e. it was not caused by perceived changes in the measurements). That is, we want to understand why each gate was operated so that future water gate management can be automated to replicate the current strategy.

3.3.1 Feature-Based Clustering

The goal of clustering the time series is to understand how to group the patterns before gate operations, so that we can classify the causes of each gate operation. With future data, this helps us identify patterns in the data that indicate when a gate operation should occur. We used our method on facilities A-D and considered all gate operations in facilities A-D and measurements a1 flow rate, b1 flow rate, c1 flow rate, d1 water level and d2 water level. To cluster the gate operations, we first normalized the data and then segmented one day (144 time steps) before each gate operation. Thus, for each gate operation, we had 5 segments. For each segment, we computed four basic features: mean, range, slope of a linear regression, and standard deviation.

We then used k-means clustering on the feature vectors, which is an iterative process that partitions vectors into k clusters with the nearest mean. To choose the number of clusters, we considered both our

goals (how many classifications we wanted) and the silhouette score of each clustering process. The silhouette score measures cluster quality (how similar an object is to its own cluster versus objects in other clusters).

We applied k-means clustering to the feature vectors in two different ways. The first was to concatenate the 4 features on all 5 segments together, so that each gate operation was associated with a 20-coordinate feature vector. This method seeks to directly cluster and classify the gate operations itself. Through experimentation, we found that the optimal number of clusters was 9 (the silhouette score increased until 9 clusters and then began to diminish at 10 clusters). Thus, each gate operation was assigned a number from 0-8.

The second method clusters the segments at each measurement location separately so that each location receives its own cluster number. All segments were classified together so that cluster numbers could be directly compared across locations. Using this method, we can indirectly classify gate operations by first clustering and classify the time series of individual water measurement points. This will provide more clarity as to which canal locations are the cause of gate operations. Using the silhouette score, we determined that only two mean and slope with four clusters would optimize the clusters formed. The quantification of each cluster is shown in Figure 12.

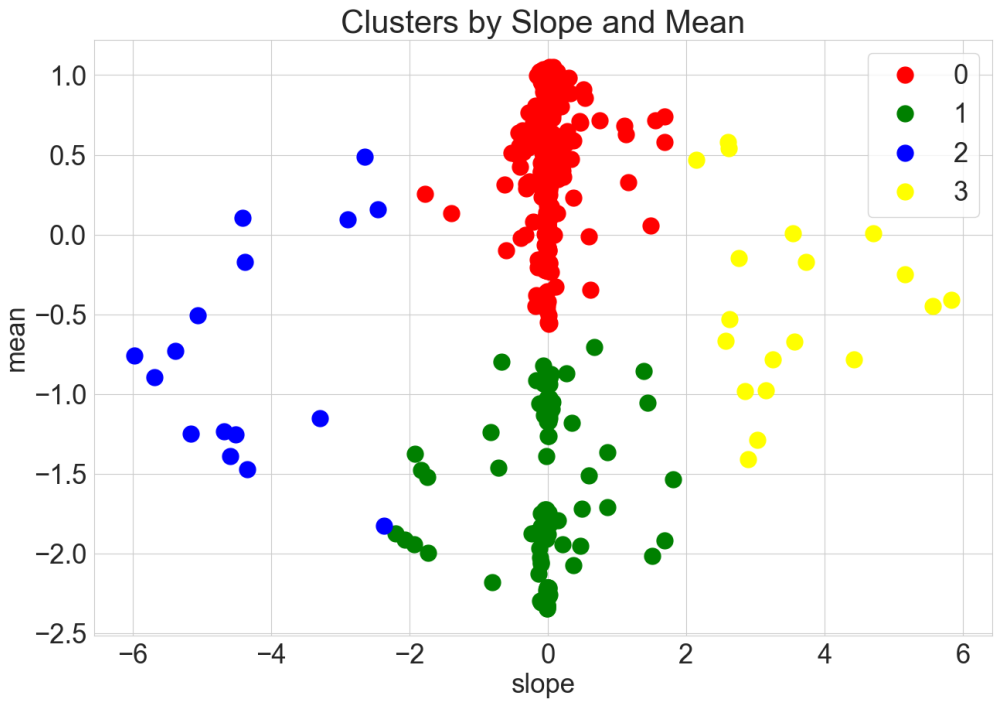


Figure 12: Quantification of univariate clusters using slope and mean.

3.3.2 Comparing Feature Based Clustering

In Table 3, we show an sample of the results after using the two described methods for feature based clustering. For each gate operation, we show the gate operated, the date and time, the cluster number (0-4) for each independent measurement location, and the gate operations classification from concatenating features (column 'Clusters').

We also compare these clusters with detected shortages. We can combine this information to determine if there are trends in the collection of cluster numbers and the gate that was operated. In Figure 13, we determine the frequency that each gate is operated for different patterns. This can help us determine in which circumstances each gate should be operated. For example, when we have many locations being clustered as '2' (large negative slope), this may indicate declining water availability. Of the 3 instances where the cluster collection is dominated by '2', gate D was operate 66% of the time.

Future gate management could then use this information to know to operate gate D when water values

are decreasing significantly. This is one example of how to use this information, but expert knowledge can be used to understand the meaning of the cluster patterns. The cluster patterns can also be used to bring a qualitative understanding to the cluster number that uses several features across multiple locations.

Pattern	Total	A (%)	B (%)	C (%)	D (%)	Usage(%)
[0, 0, 0, 0 ,1, 0, 1]	5	40.00%	0.00%	20.00%	40.00%	0.00%
[0, 0, 0, 0 ,1, 0, 0]	29	48.28%	6.90%	17.24%	17.24%	10.34%
2 Dominated	3	33.33%	0.00%	0.00%	66.67%	0.00%
3 Dominated	4	25.00%	25.00%	25.00%	25.00%	0.00%
[0, 0, 0, 0 ,0, 0, 0]	32	37.50%	31.25%	6.25%	18.75%	6.25%
[1, 1,1, 1, 1, 1,1]	7	28.57%	28.57%	0.00%	14.29%	28.57%
[0, 0,1, 1, 1, 1,1]	6	50.00%	0.00%	16.67%	33.33%	0.00%
Other	5	80.00%	0.00%	0.00%	20.00%	0.00%

Figure 13: Frequency of cluster patterns.

3.3.3 Linear Regression Classification

During the clustering process, we noticed that the most important feature for deciding the cause of a gate operation was the slope, and we are most interested in whether the measurements are increasing, decreasing, or near constant before and after a gate operation. We tried classifying each gate operation directly based on slope. We input all the time series for our canal of interest (in this case, A-D) and the indices for the suspected gate operations. We compute the slope of the linear regression for each time series during a segment of 144 indices (one day) before and after a detected gate operation. We then assigned the segment a value of -1 (decreasing), 1 (increasing), and 0 (near-constant).

We decided on a threshold for near-constant of 0.001, but this can be adjusted based on how the operator defines "increasing" and "decreasing." Since gate operations can be influenced by measurement locations nearby, we wanted a complete picture of all the measurement locations in the system at the time of the gate operation. For each gate operation, we obtain two vectors whose entries are -1, 1, and 0: one vector for "before" the operation and one vector for "after" the operation. This allows us to see at a glance which directions are changing before and after a gate operation.

To further investigate, we visualize the classifications by plotting each time series for 2 days before and after the gate change. We color the day before and after blue for surplus, red for shortage, and green for near-constant. Figure 14 shows classifications before and after a suspected gate change at index 16900. We suspect the gate operation occurred at facility B or C and was due to a shortage. The resulting vectors for "before" and "after" respectively are:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

If we can find gate changes with similar "before" and "after" vectors, we can identify patterns that indicate when a gate change should happen and its result. We then cluster the 14-dimensional vector associated with each gate operation to obtain one cluster for each gate.

Index	Time	Gate	a1 FR	b1 FR	c1 FR	d1 WL	d2 WL	Cluster	Shortage
55	4/16/2023 9:20	A	0	0	0	0	1	4	
219	4/17/2023 12:40	C	0	0	0	0	1	4	
339	4/18/2023 8:40	A	0	0	0	0	1	4	X
515	4/19/2023 14:00	A	0	0	0	0	0	3	
640	4/20/2023 10:50	A	0	0	0	0	0	3	
774	4/21/2023 9:10	C	0	0	0	0	0	3	
910	4/22/2023 7:50	–	0	0	0	0	0	3	X
1090	4/23/2023 13:50	–	0	0	0	0	0	3	X
1212	4/24/2023 10:10	D	0	0	0	0	0	3	
1360	4/25/2023 10:50	A	0	0	0	0	0	3	
1404	4/25/2023 18:10	A	0	0	0	0	0	3	
2369	5/2/2023 11:00	C	0	0	0	0	0	3	
4030	5/13/2023 23:50	A	0	0	0	0	0	3	
4275	5/15/2023 16:40	A	0	0	0	0	0	3	
4350	5/16/2023 5:10	A	0	0	0	0	0	3	
4505	5/17/2023 7:00	D	0	0	0	0	0	3	
4534	5/17/2023 11:50	B	0	0	0	0	0	3	
5849	5/26/2023 15:00	C	0	0	0	0	0	3	
6845	6/2/2023 13:00	A	0	0	0	0	0	3	
6887	6/2/2023 20:00	D	2	2	0	0	0	2	
6965	6/3/2023 9:00	A	3	2	2	0	2	2	X
7089	6/4/2023 5:40	C	3	3	3	0	3	2	
7145	6/4/2023 15:00	A	0	0	0	0	0	3	
7220	6/5/2023 3:30	A	2	0	0	0	0	3	
7254	6/5/2023 9:10	A	1	0	0	0	0	3	
7570	6/7/2023 13:50	A	0	0	0	0	0	3	
7837	6/9/2023 10:20	D	0	0	0	0	0	3	
8539	6/14/2023 7:20	C	0	0	0	0	0	3	
8550	6/14/2023 9:10	A	0	0	0	0	0	3	
9555	6/21/2023 8:40	B	0	0	0	0	0	3	

Table 3: Sample table of gate operations, the univariate clustering, and the feature clustering with concatenated vector.

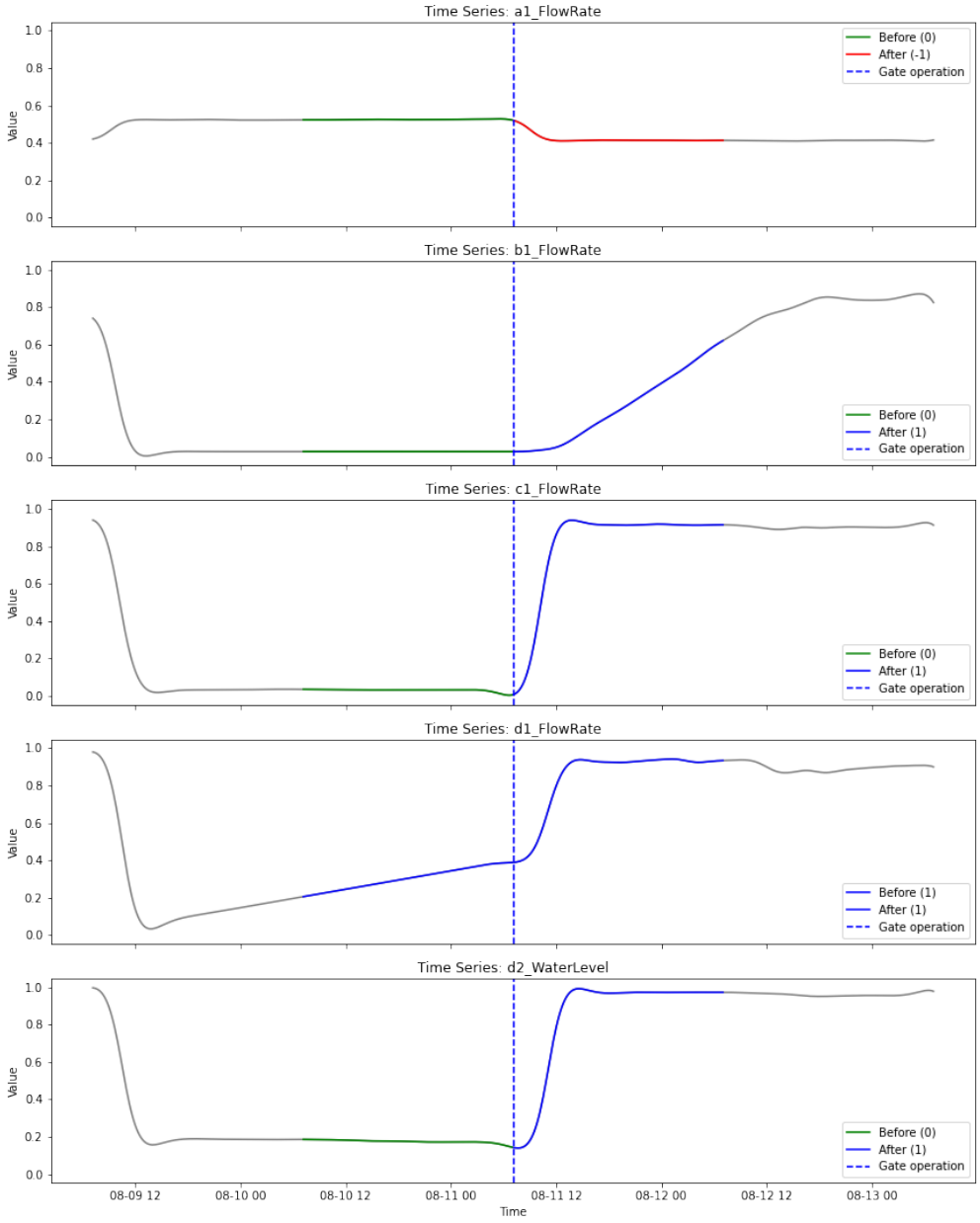


Figure 14: Classifying the slope of the linear regression for segments before and after a suspected gate change at index 16900. 0 is near-constant, 1 is increasing, -1 is decreasing.

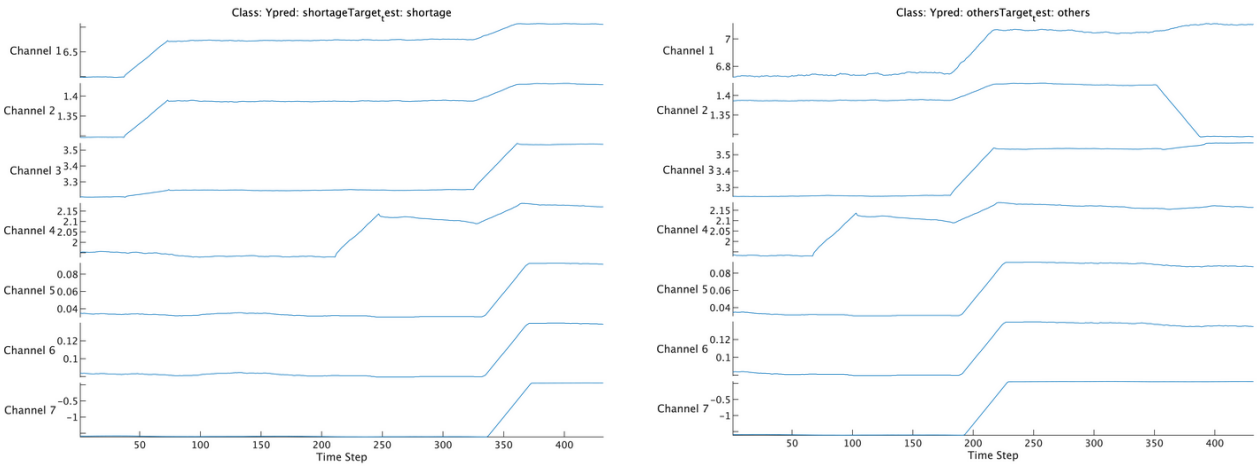
3.3.4 Neural Networks

A neural network (also known as an artificial neural network or ANN) is an adaptive system that uses nodes, or neurons, interconnected in a layered structure similar to the human brain to perform learning. Neural networks can learn from data, so they can be trained to recognize patterns, classify data, and predict future events. Using data from b1 to f1 and Manual Classification, we created a learning model that classifies into five classes: “surplus”, “shortage”, “normal”, “No gate operation” and “others”. Similar to ”Manual Classification”, this classification focuses on ”gate operation.” ”normal” is used when a level shift was detected during manual classification, but the reason could not be determined.

To train the data, we created a sequence by taking three days as one sequence and shifting the start date by one day. Specific example: April 16th, 17th, and 18th data will be Sequence 1. Next, April 17th, 18th,

and 19th data will be Sequence 2. Total number of sequences is 165.

The reason for setting the range to 3 days is that the definition of "shortage" is a decrease in the measured quantities that persists for several days as defined in Section 2.2.3. The labeling method was based on whether or not F gate operation occurred on the third day of sequence. For example, according to the Manual Classification, a F gate operation due to "shortage" is performed at the change point on the third day of Figure 15(a). The channels in Figure 15 are a1's Flow Rate, a1's Water Level, b1's Flow Rate, c1's Flow Rate, d1's Flow Rate, d1's Water Level and d2's Water Level from the top. Now label this sequence as "shortage." In a case like Figure 15(b), according to the Manual Classification, there is no gate operation on the third day, so we label it as "others."



(a) Manual Classification ; shortage, Model Classification ; shortage

(b) Manual Classification ; others, Model Classification ; shortage

Figure 15: (a) Before. (b) After.

Figure 16 shows a classification result of data from b1 to f1. The results of manual classification and that of trained model are compared. Each number inside the square indicates the number of sequences it contains. For example, if Manual Classification is "others" and model classification is "shortage", it corresponds to the square written as 11. It doesn't match completely and the correct percentage is 79 percent. For the reason, the first possibility is that the same signal may be counted twice. This is due to the way the sequence is taken. The second possibility is that we only labeled f1. There was a possibility that signals such as "surplus" were occurring elsewhere, which is why there was such a discrepancy.

As a solution, we came up with an idea of using other gate operation and classification data as training data. It may become possible to determine which gate was operated by the f1 signal. We also used LSTM (Long Short-Term Memory) as a neural network. We can determine the amount of data to be stored by setting parameters. If this characteristic can be used to store data 2 to 3 days prior to the gate operation signal, it may be possible to perform more accurate classification. We built a learning model that included all upstream observation data. The result was as shown in Figure 17 (a). The correct percentage was 83 percent, which was better than the previous result.

Furthermore, in order to see the relationship between each gate, we performed classification using a learning model that excluded data from each gate, and the results of the learning model that excluded data from gate B became interesting. The result is shown in Figure 17 (b), and the correct percentage is 82 percent. From this result, it can be concluded that the influence of gate B on gate D is extremely small.

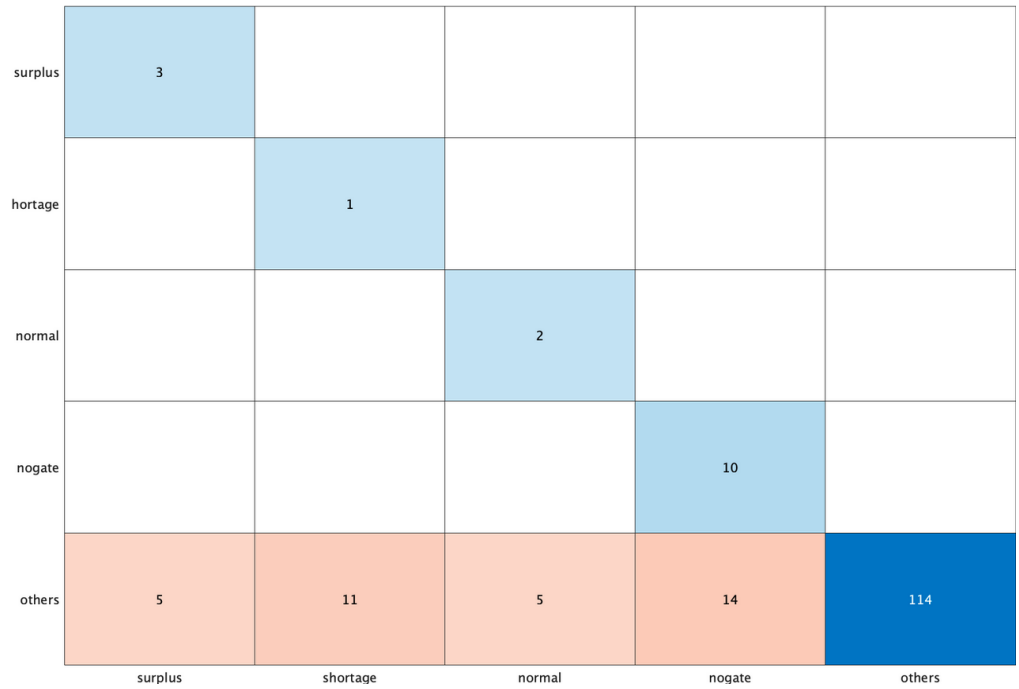
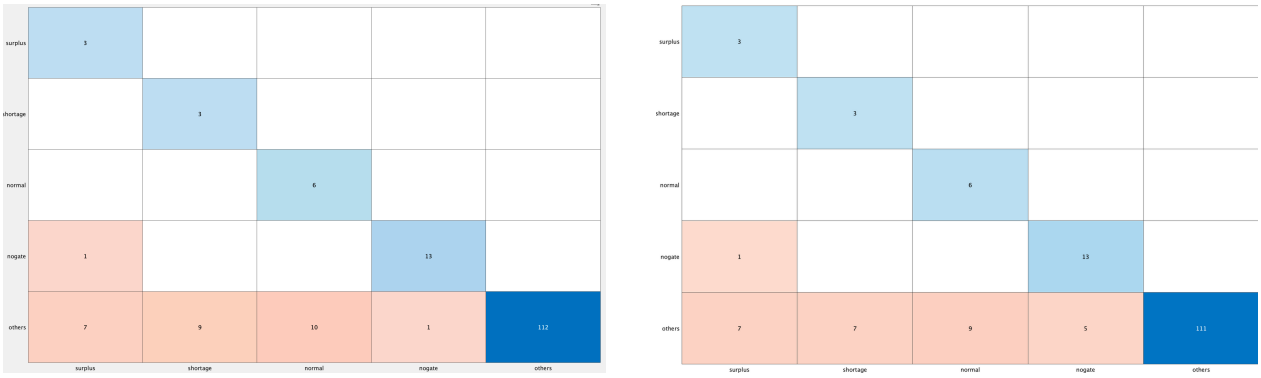


Figure 16: The vertical line shows Manual Classification and the horizontal line shows leaning model results.



(a) (b)

Figure 17: (a) Whole data. (b) Except b1 FlowRate.

3.3.5 Water Shortage Identification

Here, we introduce a method to detect options for water shortages and water surpluses by using derivatives and values.

The first program is to detect the end point of water shortages at d2 to help classify gate operations. Figure 18 is made by the following algorithm. First, we made a list of points that keep decreasing or the value is smaller than some constant. Then, we extract endpoints. We applied the same kind of program for water surplus (Figure 19).

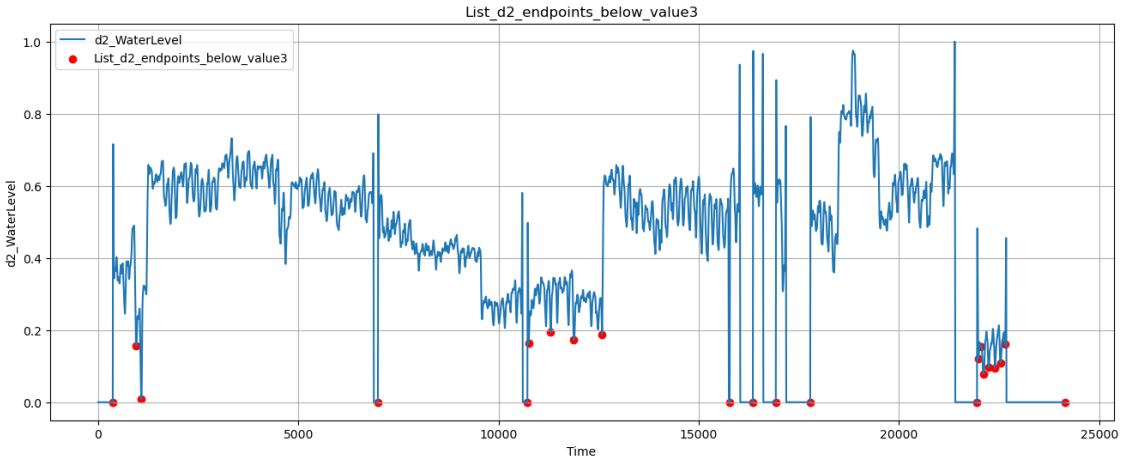


Figure 18: Options of the endpoint of water shortages at d2

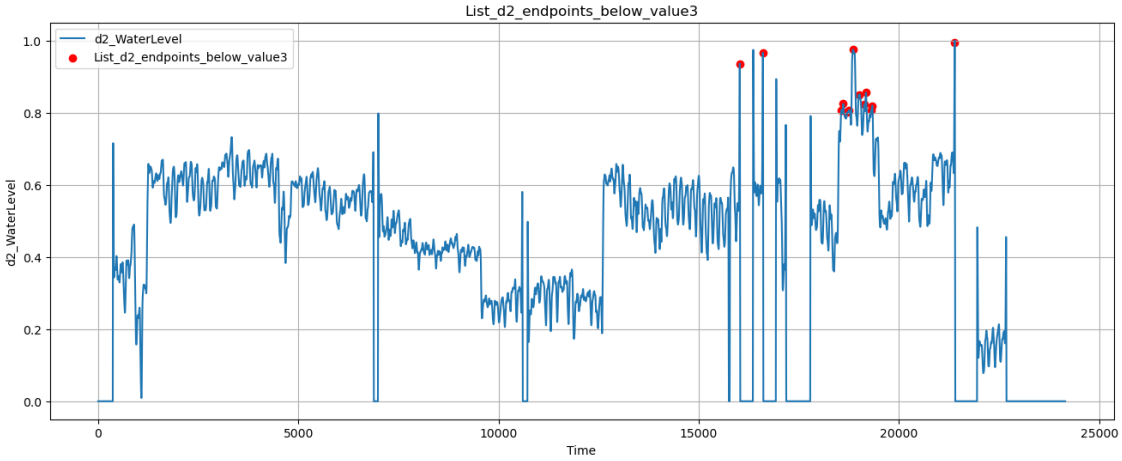


Figure 19: Options of the endpoint of water surpluses at d2

The second program is to detect water shortages which were caused only by increasing water demand between d1 and d2. At these points, the operator probably should have operated the gate. Figure 20 shows points that the water level at d2 keeps decreasing and is lower than some constant, or lower than another constant. We applied the same program to the water level at d1 (Figure 21). Figure 22 shows the result by subtracting the second one from the first one.

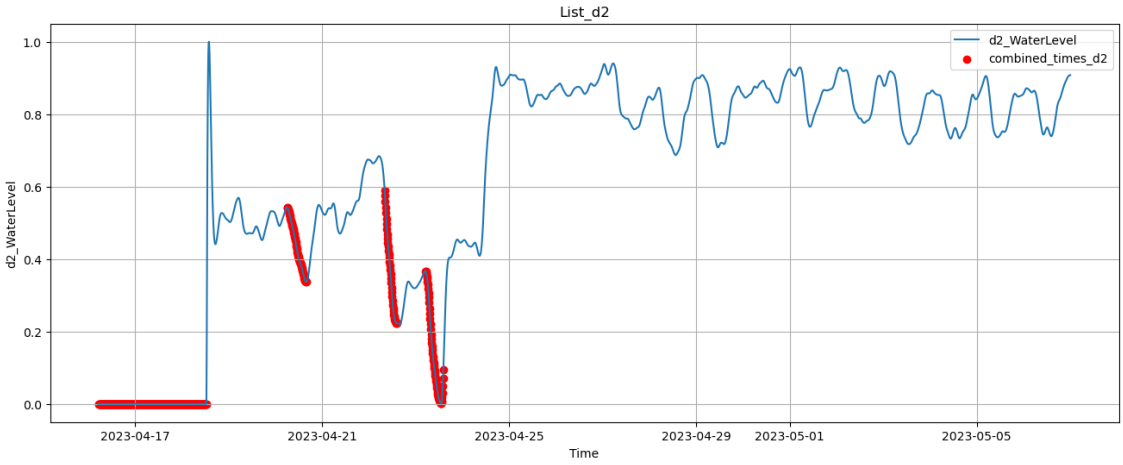


Figure 20: Points at d2

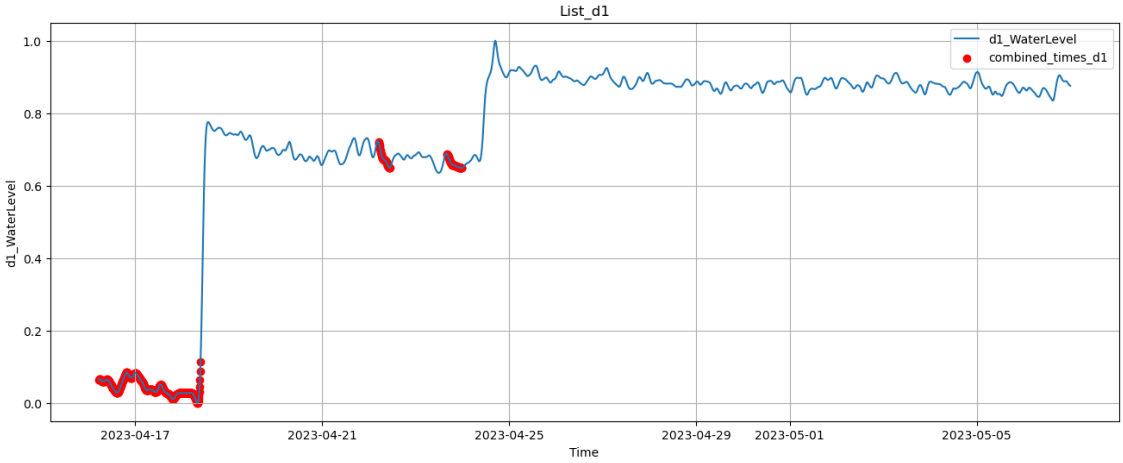


Figure 21: Points at d1

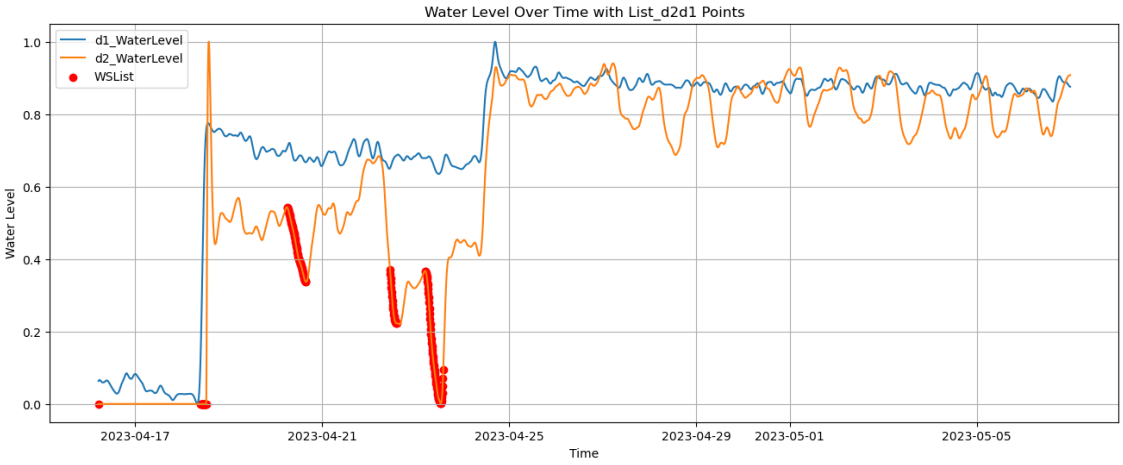


Figure 22: After subtracting

3.4 Identifying trends

In addition to identifying gate operations and classifying their cause, we also want to understand trends at each location and in the whole water system.

3.4.1 Trends at One Location

We return to feature clustering described in Section 3.3.1. We consider a time series at one location and use a moving window of 1 days (with offset of 1/2 day) to obtain 334 segments for each time series. We do not normalize the data, since we only compare measurements at one location. We then computed features (slope, mean, standard variation, and range) for each segment and used k-means clustering to find clusters. We used the silhouette score to determine the number of clusters and found for each of the 5 locations, 5-6 clusters were optimal.

In Figure 23, we show the algorithm run on a1 flow rate, c1 flow rate, and d1 flow rate over the whole time series. Clustering at one location allows us to identify both recurring and unusual trends. For example, in a1 WaterFlow, we see that there are two different periods of oscillating levels. In the first period (blue), we've identified that it is around the time of the start of the rainy season, and so the dam gate may have opened and closed rapidly to adjust to the changing waterfall. The second period (yellow) corresponds to the rice harvest season, and so the dam opens and closes at more regular intervals to allow water to reach different irrigation canals and maintain the flow to farmland. However, the cluster (red) is by the far the most common cluster and recurs several times during the time period studied.

For each time series, we can also obtain a list of clusters in order, as shown in Figure 23(d). We can then perform statistical analysis on this list, such as the maximum run length of a cluster. A larger maximum run length suggests that a time series has a longer period of consistent values. A very small maximum run length suggests that a time series moves frequently between different clusters, which might suggest a more erratic time series.

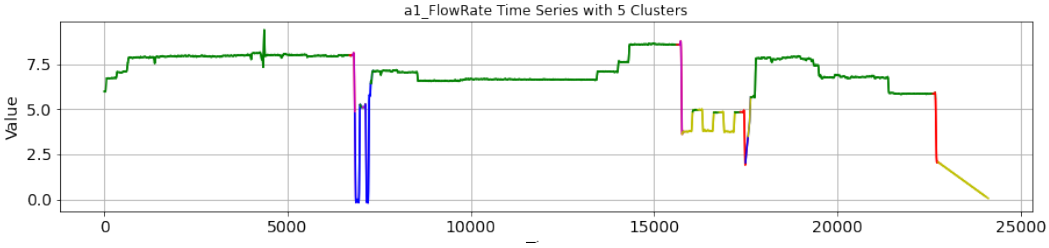
3.4.2 Trends at Multiple Locations

To find trends at multiple locations, we considered our measurement locations as part of a multivariate system. We considered each time step as a multi-dimensional point whose coordinates come from each location. For example, when considering 5 measurement locations from Facilities A-D, we obtained 24132 5-dimensional points. We also normalized the measurements for each location, so that one location did not skew the results. We have two methods for analyzing the data: clustering and anomaly detection.

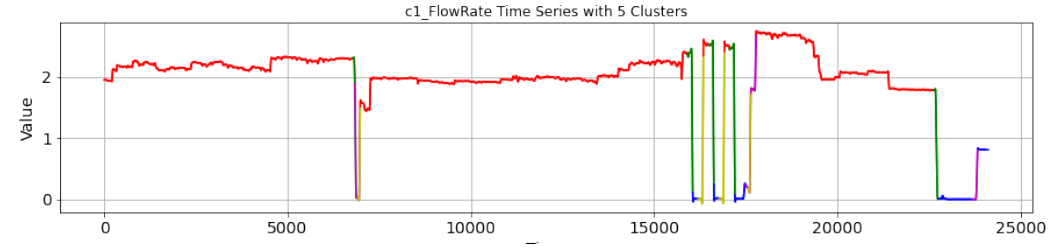
When clustering, we use k-means clustering as before. We also use the silhouette score to determine the optimal number of clusters. In most cases, 5-6 clusters were optimal. Figure 24 shows the clustering algorithm run on 5 locations from the A-D canal over a period of one month. Also included is a 2-dimensional graph showing only a1 FlowRate and d1 WaterLevel and the cluster colors assigned. Points (time stamps) with similar measurements tend to cluster together, so we can see if there are recurring patterns in the measurements when taken together.

We see that the clusters change when there is a significant shift at one or several locations. We can use this information in two ways: First, we can detect that significant events have occurred when the cluster changes (such as a gate operation). We call these cluster change points. Figure 24(c) shows the indices where the clusters change in canals A-D over the whole time period. Note that these indices are highly dependent on the number of clusters, and we will get more indices if we choose a larger number of clusters.

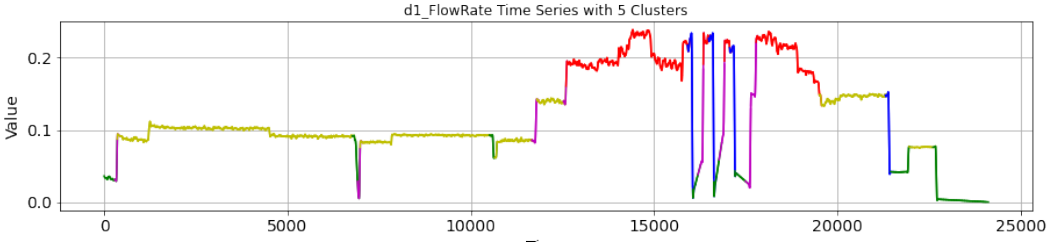
Second, we can use the clusters to learn about the behavior of the canal system around a gate operation. For example, at index 18902 and index 14907, we manually detected a gate change at facility D caused by a surplus. After multivariate clustering, both points were put in the same cluster. This indicates that there is a common pattern before facility D is operated when there is a surplus in the system. However, more work is needed to understand what these clusters represent.



(a)



(b)

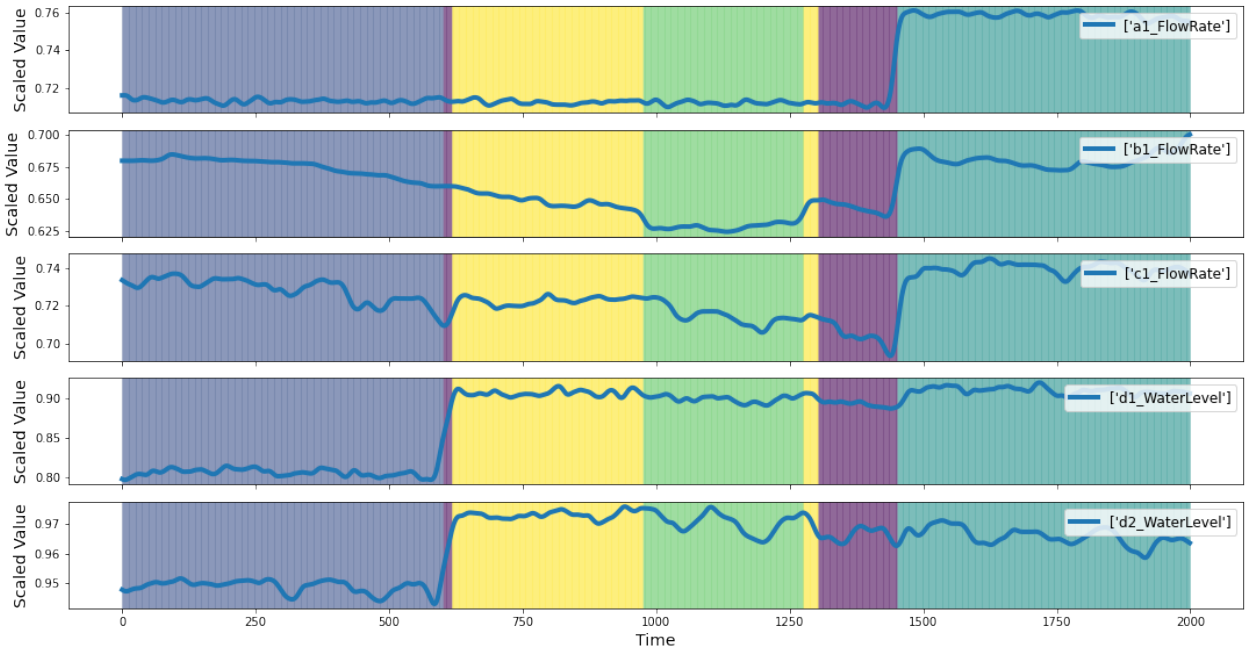


(c)

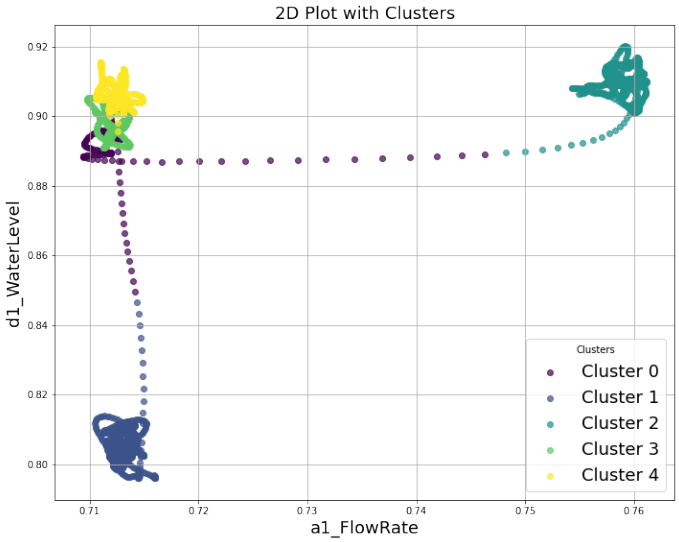
```
[1 1 1 1 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 4 3 3 3 3 3 3 3 3 3 3 3 3
 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1
 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 3 3 3 3 3 3 3 3 3 3 4 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2
 2 1 1 4 4 0 0 2 2 1 1 4 4 0 0 2 2 1 1 1 1 4 4 4 4 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 2 2 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1]
```

(d)

Figure 23: (a)-(c) Clustering the time series into 5 clusters based on 4 features: slope, mean, range, and standard deviation. Note that the data is not normalized, since we are not comparing time series. (d) List of clusters for d1 Flow Rate.



(a)



(b)

351, 628, 4353, 4381, 6824, 6831, 6841, 6853, 6956, 6968, 6972, 6974, 7138,
7151, 7213, 7225, 7268, 9560, 10614, 10706, 14015, 14030, 14319, 15742,
15744, 15748, 15750, 15756, 16043, 16333, 16619, 16620, 16918, 17193, 17474,
17555, 17648, 17768, 17772, 19488, 21391, 21399, 21414, 21936, 22689, 22701,
23802

(c)

Figure 24: (a) One month of data from A-D facilities clustered into 5 clusters by multivariate clustering. Colors indicate cluster. (b) Plotting two of the coordinates as points where a1 flow rate is the x-coordinate and d1_WaterLevel is the y-coordinate with 5 clusters. Note that the points are 5-dimensional but only 2 coordinates are shown, so some clusters appear more similar than they are. (c) Indices where clusters change in canals A-D over the whole time period.

3.4.3 VAR-LiNGAM Causal Discovery

VAR-LiNGAM is an extension of basic LiNGAM to analyze time series data. The word VAR means Vector Auto Regressive. Mathematically, the model is written as

$$x(t) = \sum_{\tau=0}^{k} B_\tau x(t-\tau) + e(t), \tag{3.9}$$

where $x(t) \in \mathbb{R}^n$ is variables, $e(t)$ is error variables, and B_τ is coefficient matrix with time lag $\tau \in [0, k]$. There are several approaches for estimating the parameters as described in [27]. In Appendix B, we will explain the outline of the Two-Stage Method, which is one of the approaches. Lag time τ should be the maximum value in delays between measurement points. In essence, the calculation can be reduced to a VAR model.

When we use VAR-LiNGAM, it requires the following 4 assumptions:

1. Linearity,
2. Non-Gaussian continuous error variables (except at most one),
3. Acyclic of contemporaneous causal relations,
4. No hidden common causes between contemporaneous error variables.

As assumed in Section 2.2.2, the network we are considering does not have directed cycles, therefore the third assumption is satisfied. As for the other assumptions, they are anticipated to be greatly influenced by the preprocessing (smoothing) methods described in Section 3.1.3. Their impact will be investigated in this section by using the following hypothesis test: Hilbert-Schmidt independence criterion (HSIC) test and Shapiro–Wilk test.

Next, we will show three results. we applied the VAR-LiNGAM model to preprocessed data collected over a period of 4 weeks from July 1 to July 28. In that season, farmers use a massive quantity of water.

For smoothing, we removed spikes by applying one of "moving mean with window size 6", "moving median with window size 6", and "trend filter." After that, we standardized the data so that they have a mean of 0 and a standard deviation of 1.

How VAR-LiNGAM Works

Moving Mean with Window Size 6

First, we will show results of moving average with window size 6. We used data from July 1 to July 28. The estimated lag time is 8. It is automatically estimated by the VAR-LiNGAM package. As for the adjacency matrices, please refer to Appendix B. Figure 25 shows the result of predicting time series.

After calculating lag time and adjacency matrices, we apply two tools to verify them. The first tool is the Hilbert-Schmit independence criteria (HSIC test) for the independence of each noise $e_i(t)$. VAR-LiNGAM requires that "No hidden common causes between contemporaneous error variables." When the result of HSIC (probability value) does not reject the null hypothesis, we don't have hidden causes and satisfy one assumption.

Figure 26 shows a heat map of the matrix, where each value is represented with a significance level of 0.05. By this result, at a significance level of approximately 0.05, we can confirm that *moving average* meets one of the assumptions that no hidden common causes between contemporaneous error variables.

The second tool is the Shapiro-Wilk test for Non-Gaussian continuous error variables (Figure 27). We set some significance levels. When the back is gray, it means they meet the assumption.

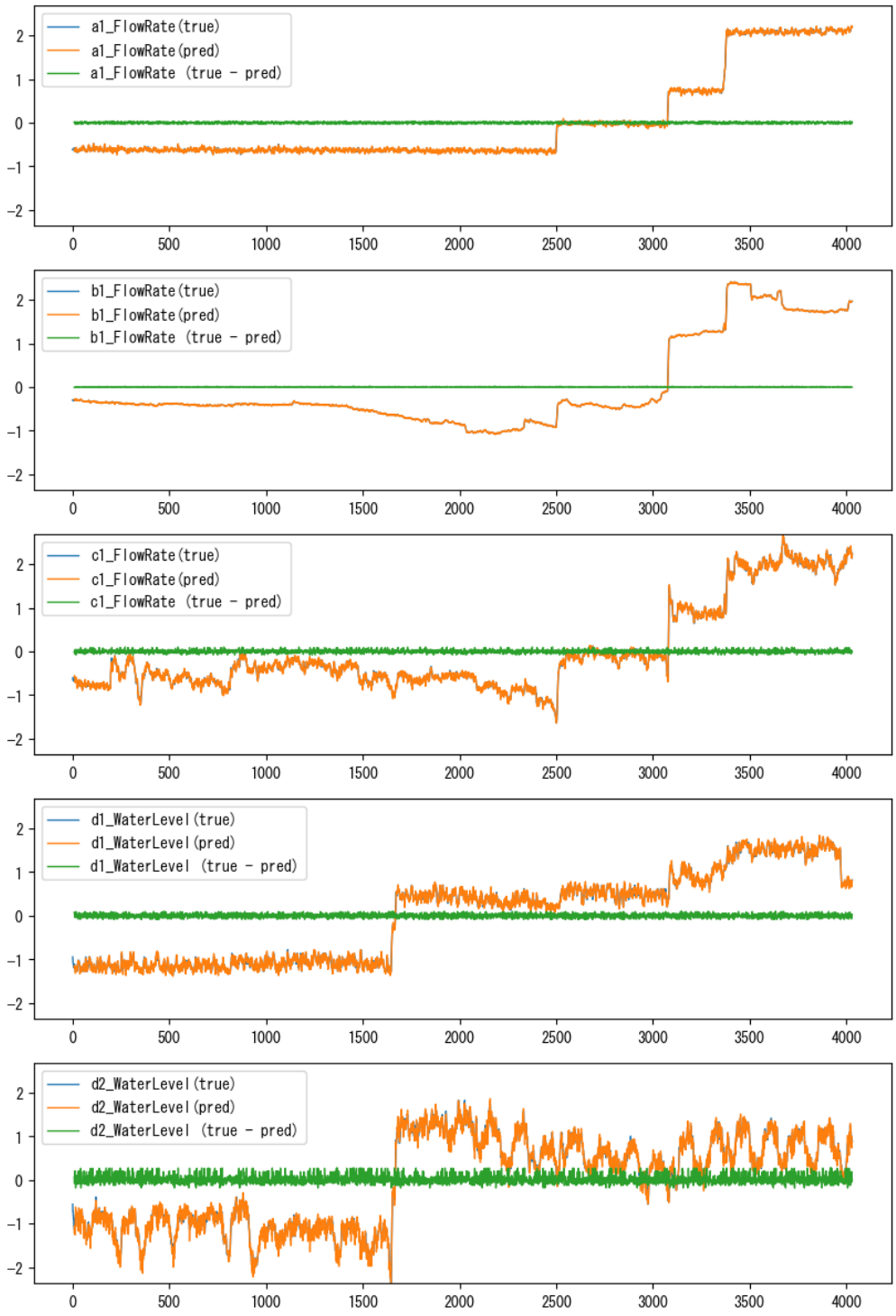


Figure 25: Results of prediction (preprocessed by moving mean)

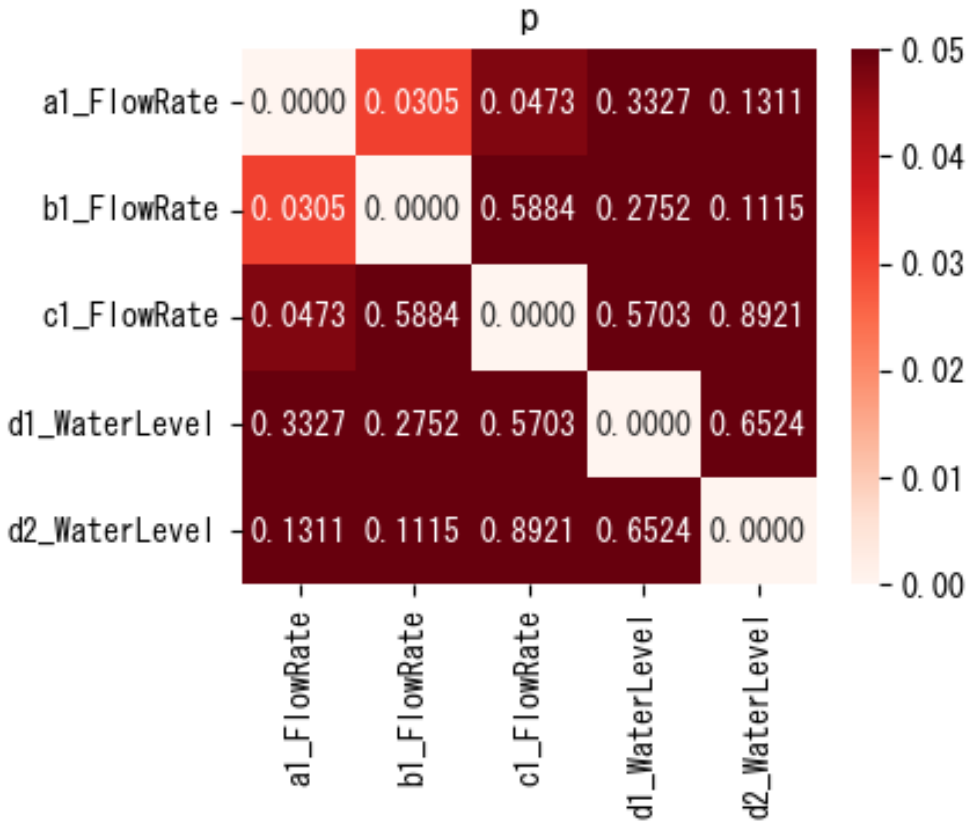


Figure 26: Hilbert-Schmidt Independence Criteria test (preprocessed by moving mean filter)

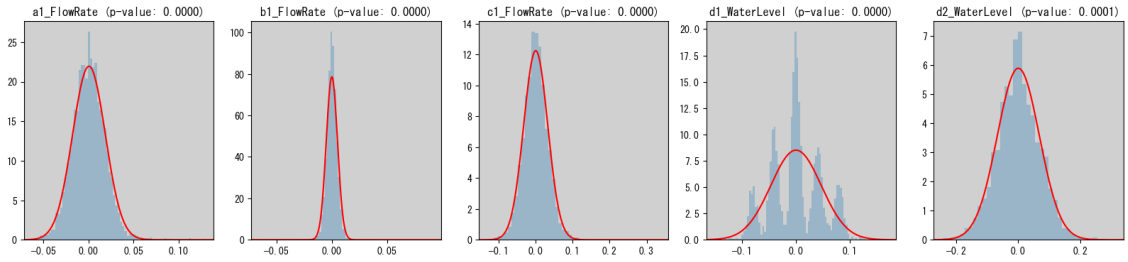


Figure 27: Shapiro-Wilk test (preprocessed by moving mean filter)

Next, by using bootstrap, we estimate the total effect on each variable. Table 4 shows the list of effects, sorted in order of probability. We can estimate each effect, probability, and lag. The probability refers to the proportion of total effects from *from* to *to* that were non-zero, and the effect refers to the median of these non-zero total effects. For example, at No.29, the flow rate at b1 has an effect 0.269016 on a1 with lag7, and the probability is 0.98. It means more than 98 percent of every point have the median of these effects is 0.269016.

No	effect	probability	from	to	lag
0	-0.095987	1.00	d2_WaterLevel	d2_WaterLevel	lag8
1	-0.117510	1.00	a1_FlowRate	a1_FlowRate	lag8
2	-0.471616	1.00	a1_FlowRate	a1_FlowRate	lag6
3	0.082157	1.00	a1_FlowRate	b1_FlowRate	lag1
4	1.793416	1.00	b1_FlowRate	b1_FlowRate	lag1
5	-0.065689	1.00	a1_FlowRate	b1_FlowRate	lag2
6	-0.719023	1.00	b1_FlowRate	b1_FlowRate	lag2
7	-0.152706	1.00	b1_FlowRate	b1_FlowRate	lag4
8	-0.314752	1.00	b1_FlowRate	b1_FlowRate	lag6
9	0.768102	1.00	b1_FlowRate	b1_FlowRate	lag7
10	-0.035959	1.00	a1_FlowRate	b1_FlowRate	lag8
11	-0.374023	1.00	b1_FlowRate	b1_FlowRate	lag8
12	1.197097	1.00	c1_FlowRate	c1_FlowRate	lag1
13	-0.509836	1.00	c1_FlowRate	c1_FlowRate	lag6
14	0.605447	1.00	c1_FlowRate	c1_FlowRate	lag7
15	-0.153987	1.00	c1_FlowRate	c1_FlowRate	lag2
16	1.032371	1.00	d1_WaterLevel	d1_WaterLevel	lag1
17	0.532379	1.00	d2_WaterLevel	d2_WaterLevel	lag7
18	-0.511730	1.00	d2_WaterLevel	d2_WaterLevel	lag6
19	1.062264	1.00	d2_WaterLevel	d2_WaterLevel	lag1
20	1.129086	1.00	a1_FlowRate	a1_FlowRate	lag1
21	0.422287	1.00	b1_FlowRate	a1_FlowRate	lag1
22	-0.146788	1.00	c1_FlowRate	c1_FlowRate	lag8
23	0.559796	1.00	a1_FlowRate	a1_FlowRate	lag7
24	-0.096752	1.00	a1_FlowRate	a1_FlowRate	lag2
25	-0.498998	1.00	d1_WaterLevel	d1_WaterLevel	lag6
26	0.030955	1.00	d2_WaterLevel	d1_WaterLevel	lag1
27	0.479416	1.00	d1_WaterLevel	d1_WaterLevel	lag7
28	-0.444514	0.99	b1_FlowRate	a1_FlowRate	lag2
29	0.269016	0.98	b1_FlowRate	a1_FlowRate	lag7

Table 4: Effects and probabilities of various variables with different lag times (preprocessed by trend filter).

Moving Median with Window Size 6

Second, we will show results of using a moving median with a window size 6. The estimated lag time is 7.

Figure 28 shows the result of applying VAR-LiNGAM. The result of HSIC test for the independence of each noise $e_i(t)$ is shown in Figure 29. Figure 30 shows the results of the Shapiro-Wilk test (focus on the water level at d1). Table 5 shows the list of effects and probability.

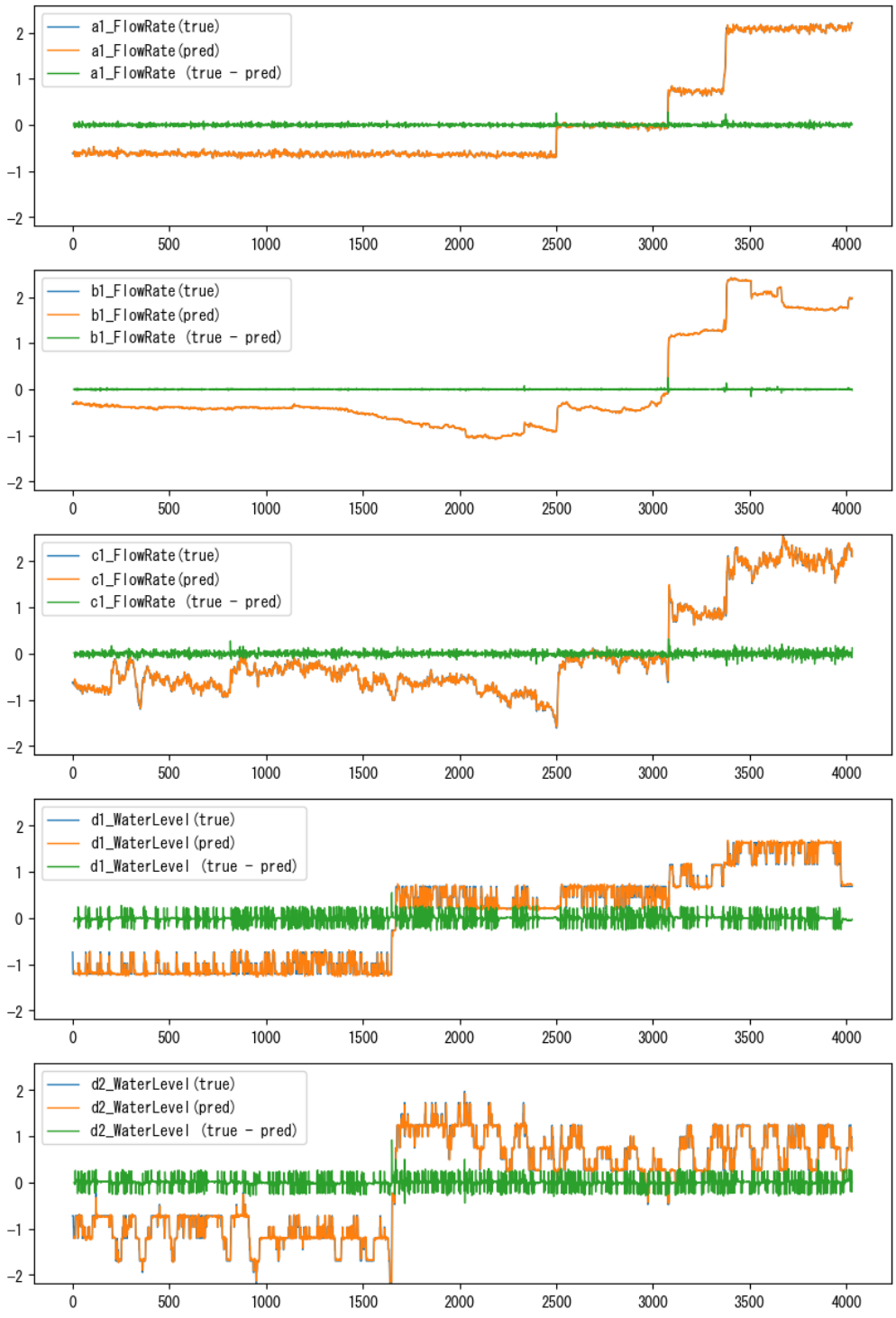


Figure 28: Results of prediction (preprocessed by moving median filter)

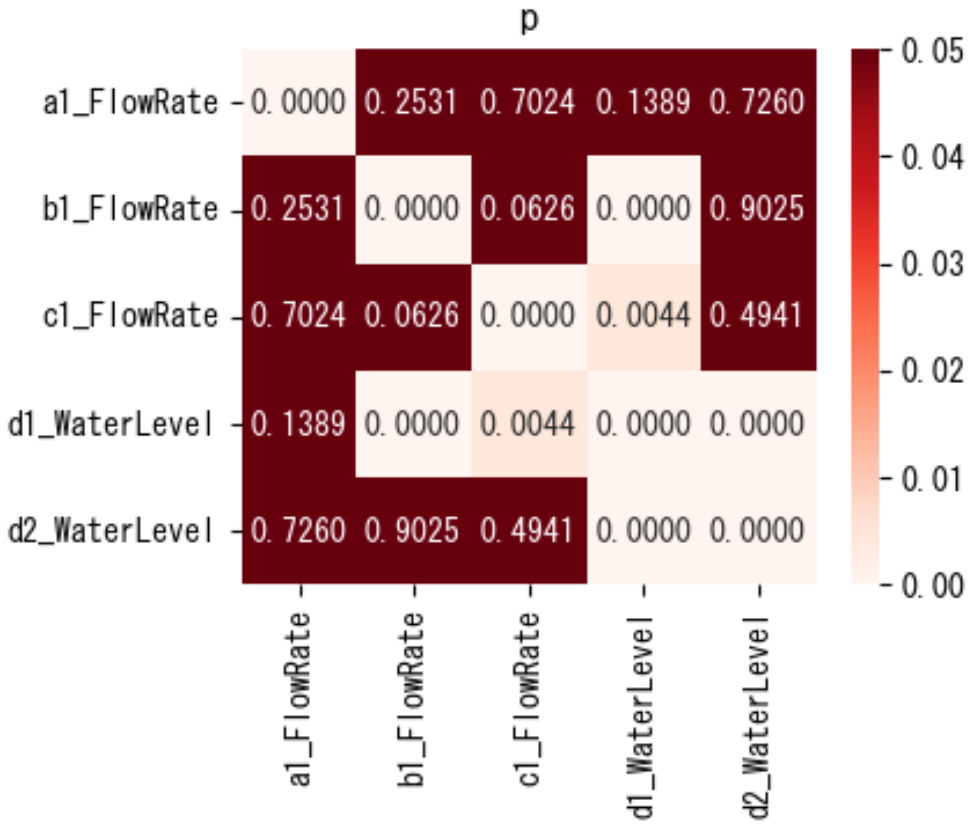


Figure 29: Hilbert-Schmidt Independence Criteria test (preprocessed by moving median filter)

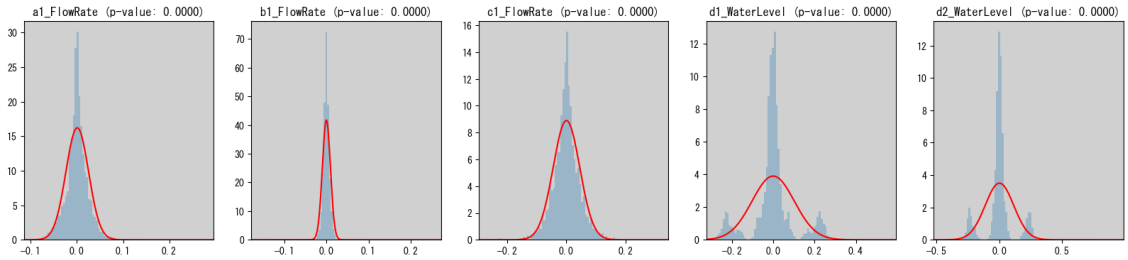


Figure 30: Shapiro-Wilk test (preprocessed by moving median filter)

No	from	to	effect	probability	from_var	to_var	from_lag
0	39	4	0.161662	1.00	d2_WaterLevel	d2_WaterLevel	lag7
1	35	0	0.269674	1.00	a1_FlowRate	a1_FlowRate	lag7
2	6	1	1.782726	1.00	b1_FlowRate	b1_FlowRate	lag1
3	32	2	-0.288218	1.00	c1_FlowRate	c1_FlowRate	lag6
4	30	0	-0.265764	1.00	a1_FlowRate	a1_FlowRate	lag6
5	37	2	0.262726	1.00	c1_FlowRate	c1_FlowRate	lag7
6	8	3	0.867672	1.00	d1_WaterLevel	d1_WaterLevel	lag1
7	33	3	-0.214354	1.00	d1_WaterLevel	d1_WaterLevel	lag6
8	38	3	0.246195	1.00	d1_WaterLevel	d1_WaterLevel	lag7
9	7	2	1.043355	1.00	c1_FlowRate	c1_FlowRate	lag1
10	34	4	-0.154796	1.00	d2_WaterLevel	d2_WaterLevel	lag6
11	11	1	-1.039581	1.00	b1_FlowRate	b1_FlowRate	lag2
12	5	0	1.088091	1.00	a1_FlowRate	a1_FlowRate	lag1
13	9	4	0.935843	1.00	d2_WaterLevel	d2_WaterLevel	lag1
14	5	1	0.074092	0.98	a1_FlowRate	b1_FlowRate	lag1
15	8	4	0.042818	0.97	d1_WaterLevel	d2_WaterLevel	lag1
16	8	2	0.002066	0.97	d1_WaterLevel	c1_FlowRate	lag1
17	9	2	0.009723	0.97	d2_WaterLevel	c1_FlowRate	lag1
18	4	2	0.065230	0.96	d2_WaterLevel	c1_FlowRate	lag0
19	39	2	0.007527	0.96	d2_WaterLevel	c1_FlowRate	lag7
20	34	2	-0.009103	0.96	d2_WaterLevel	c1_FlowRate	lag6
21	9	3	0.028833	0.94	d2_WaterLevel	d1_WaterLevel	lag1
22	12	2	-0.061300	0.93	c1_FlowRate	c1_FlowRate	lag2
23	6	2	0.261983	0.91	b1_FlowRate	c1_FlowRate	lag1
24	16	1	0.249780	0.90	b1_FlowRate	b1_FlowRate	lag3
25	16	2	-0.479593	0.89	b1_FlowRate	c1_FlowRate	lag3
26	10	0	-0.096222	0.87	a1_FlowRate	a1_FlowRate	lag2
27	11	2	0.456425	0.84	b1_FlowRate	c1_FlowRate	lag2
28	5	2	0.017198	0.82	a1_FlowRate	c1_FlowRate	lag1
29	5	3	0.051435	0.78	a1_FlowRate	d1_WaterLevel	lag1

Table 5: Effects and probabilities of various variables with different lag times (preprocessed by moving median filter).

Trend Filter

Finally, we will show results of trend filter. Figure 31 shows the result of applying VAR-LiNGAM. Figure 32 shows a heat map of the result of the HSIC test. Figure 33 shows the result of Shapiro-Wilk Test. Table 6 shows the list of effects and probability.

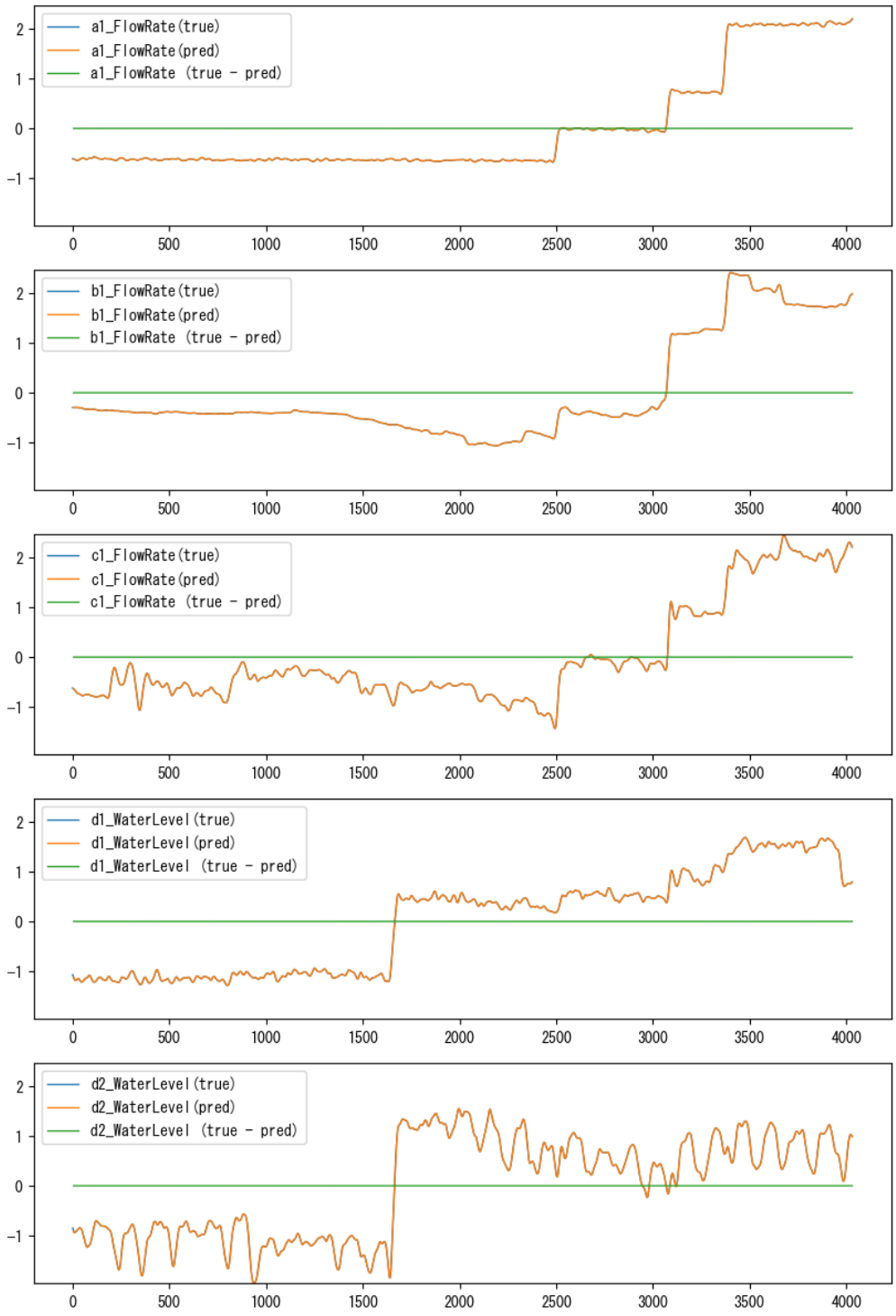


Figure 31: Results of prediction (preprocessed by trend filter)

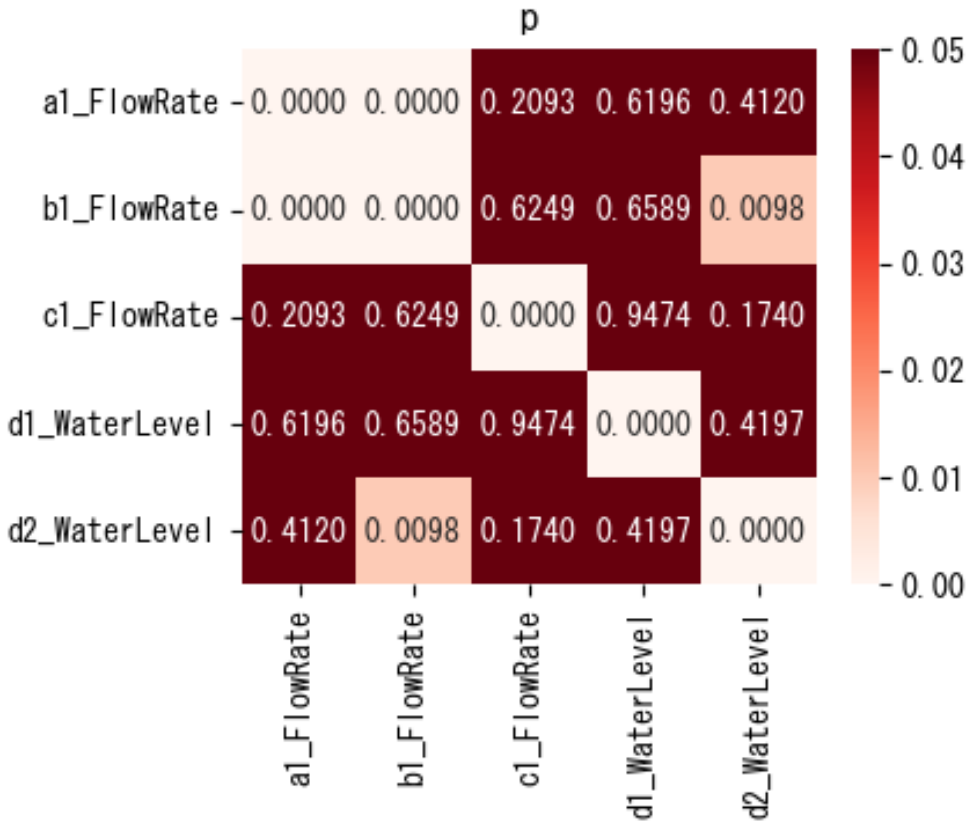


Figure 32: Hilbert-Schmidt Independence Criteria test (preprocessed by trend filter)

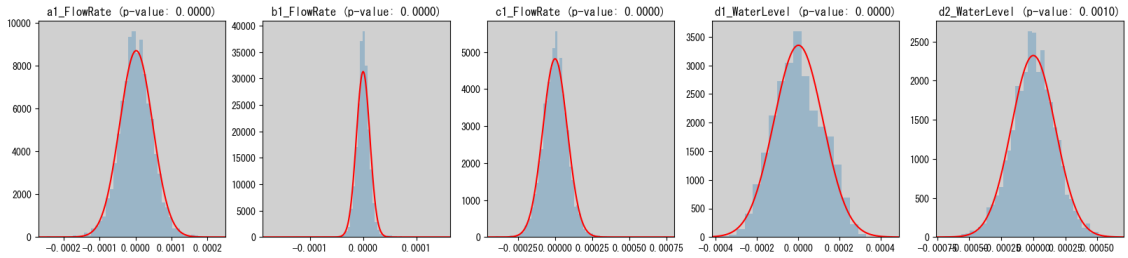


Figure 33: Shapiro-Wilk test (preprocessed by trend filter)

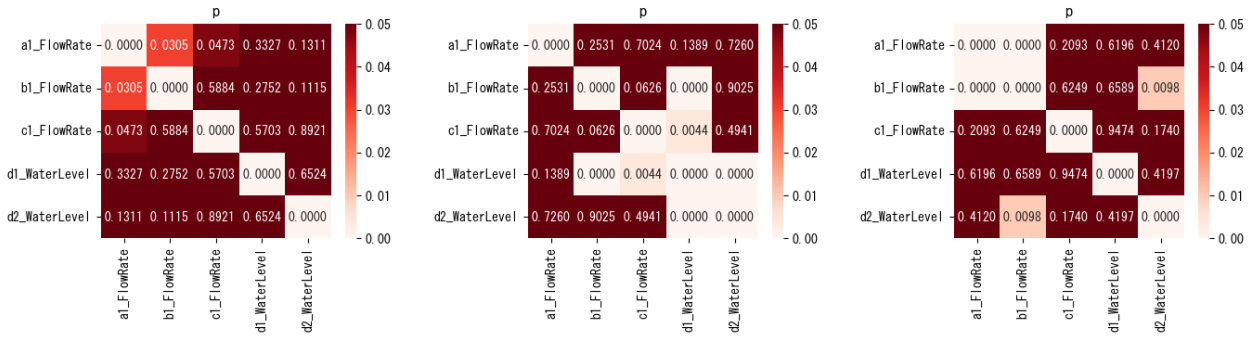
No	from	to	effect	probability	from_var	to_var	from_lag
0	5	0	2.058925	1.00	a1_FlowRate	a1_FlowRate	lag1
1	22	2	0.325477	1.00	c1_FlowRate	c1_FlowRate	lag4
2	12	2	-1.960476	1.00	c1_FlowRate	c1_FlowRate	lag2
3	7	2	2.634759	1.00	c1_FlowRate	c1_FlowRate	lag1
4	8	3	2.631244	1.00	d1_WaterLevel	d1_WaterLevel	lag1
5	13	3	-1.956770	1.00	d1_WaterLevel	d1_WaterLevel	lag2
6	6	1	2.649720	1.00	b1_FlowRate	b1_FlowRate	lag1
7	9	4	2.971761	1.00	d2_WaterLevel	d2_WaterLevel	lag1
8	14	4	-2.952649	1.00	d2_WaterLevel	d2_WaterLevel	lag2
9	11	1	-1.979562	0.99	b1_FlowRate	b1_FlowRate	lag2
10	19	4	0.980942	0.99	d2_WaterLevel	d2_WaterLevel	lag3
11	10	0	-1.062003	0.87	a1_FlowRate	a1_FlowRate	lag2
12	21	1	0.329877	0.82	b1_FlowRate	b1_FlowRate	lag4
13	23	3	0.325297	0.78	d1_WaterLevel	d1_WaterLevel	lag4
14	21	0	0.024187	0.77	b1_FlowRate	a1_FlowRate	lag4
15	6	2	0.002311	0.76	b1_FlowRate	c1_FlowRate	lag1
16	16	3	-0.000873	0.73	b1_FlowRate	d1_WaterLevel	lag3
17	6	3	0.000942	0.72	b1_FlowRate	d1_WaterLevel	lag1
18	17	3	-0.000066	0.68	c1_FlowRate	d1_WaterLevel	lag3
19	17	4	-0.000256	0.61	c1_FlowRate	d2_WaterLevel	lag3
20	21	4	0.001497	0.61	b1_FlowRate	d2_WaterLevel	lag4
21	18	0	0.000024	0.58	d1_WaterLevel	a1_FlowRate	lag3
22	15	4	-0.000088	0.56	a1_FlowRate	d2_WaterLevel	lag3
23	11	0	-0.030166	0.56	b1_FlowRate	a1_FlowRate	lag2
24	13	4	0.000075	0.54	d1_WaterLevel	d2_WaterLevel	lag2
25	7	4	0.000665	0.52	c1_FlowRate	d2_WaterLevel	lag1
26	11	2	-0.005469	0.52	b1_FlowRate	c1_FlowRate	lag2
27	11	4	-0.001182	0.51	b1_FlowRate	d2_WaterLevel	lag2
28	19	2	-0.000019	0.50	d2_WaterLevel	c1_FlowRate	lag3
29	19	3	-0.000065	0.47	d2_WaterLevel	d1_WaterLevel	lag3

Table 6: Effects and probabilities of various variables with different lag times (preprocessed by trend filter).

Which Preprocessing is Better for VAR-LiNGAM

In this paragraph, we compare the three result. By comparing the results of the HSIC (Figure 34), we conclude that the result of the *moving average* is the best and that of the *moving median* is the worst. As original data is the same, these differences are caused by how we preprocessed the data.

Table 7 shows the variances of prediction errors. The variance of *trend filter* is extremely small. This means a better result in terms of fitting. However, over smoothing can be detrimental to hypothesis tests for common causes between the error variables.



(a) Moving average. (b) Moving median. (c) Trend filter.

Figure 34: Comparison of each heat map.

Point	*mean*	*median*	*trend*
a1_FlowRate	0.000301	0.001329	1.407542e-09
b1_FlowRate	0.000014	0.000037	1.151261e-10
c1_FlowRate	0.001287	0.002369	6.823062e-09
d1_WaterLevel	0.002314	0.004385	1.434665e-08
d2_WaterLevel	0.005657	0.013354	2.237900e-08

Table 7: Variance of errors

Evaluating Effects of Each Value

In the previous paragraph, we compared the results from three preprocessed methods. Next, we will elaborate each effect focusing on the water level at d1. Using bootstrap, we list effects on each point. Figure 35 shows the result of applying VAR-LiNGAM. This means the value of d1 is influenced by the value of d1 from time step earlier with the effect 1.066, the value of d1 from 6 steps earlier with the effect -0.510, and the value of d1 from 6 steps earlier with the effect 0.537.

From the result, VAR-LiNGAM could capture very reasonable facts that the value of d1 is most influenced by the value of d1 from one time step earlier, and other values mean are not dominant. As shown in Figure 36, other matrices shows similar results. This phenomenon occurs at 4 points out of 5 points from a1 to d2. When we try different passes, from a1 to j2, a1 to g2, and a1 to h1, we obtain the same results in almost all location except b1. The influence observed at 6 and 7 steps earlier is assumed to be due to the time delay introduced by preprocessing.

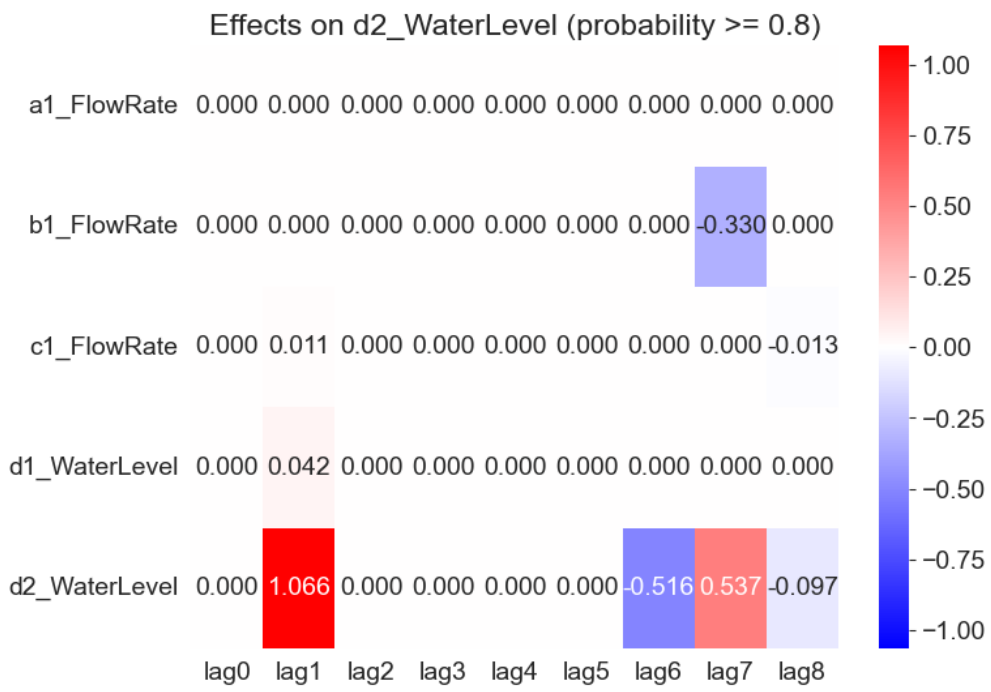
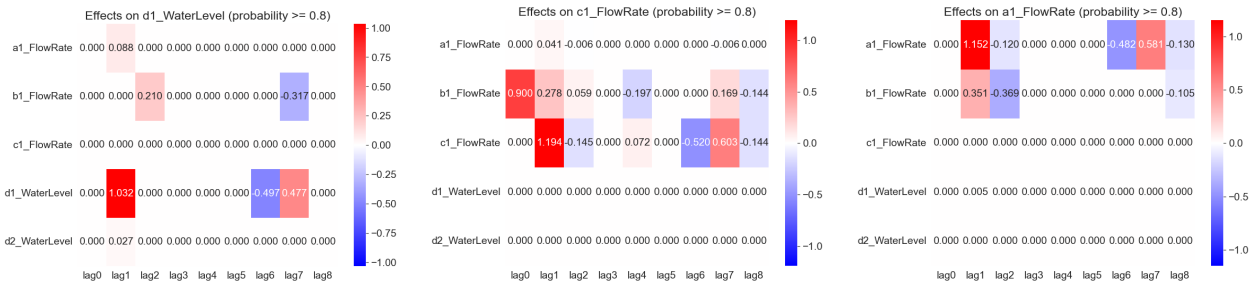


Figure 35: Effects on d1



(a) Effects on d1 (b) Effects on c1 (c) Effects on a1

Figure 36: Effects

4 Discussion

4.1 Insights from Detection Methods

The detection of gate operations using level shift detection methods proved effective in identifying significant changes in water levels that correspond to gate operations. The univariate feature-based clustering method provided a straightforward approach to detecting these shifts, while the multivariate clustering method offered more nuanced insights by considering multiple variables simultaneously.

Manual detection of gate operations was important for validating automated methods. Our comparison of some detection methods demonstrated that automated techniques could replicate the accuracy of manual detection, reducing the need for constant human oversight.

4.2 Classifying Cause of Gate Operation

Univariate Feature-Based Clustering segments time series data into clusters based on specific features such as slope, mean, range, and standard deviation. By analyzing patterns before gate operations, we can

distinguish periods of surplus or shortage. The method involves two main strategies: clustering individual time series and segmenting time series before anomalies. Both strategies show promise in identifying critical patterns that lead to gate operations.

Linear Regression Clustering uses linear regression to cluster data points, providing a clear representation of the relationship between different measurements. It helps in understanding how changes in one variable may influence others, thus aiding in identifying the causes of gate operations more accurately.

By considering multiple variables simultaneously, this method offers a comprehensive view of the factors influencing gate operations. It employs anomaly detection in a multivariate space, identifying points with significantly different densities compared to their neighbors. This approach is effective in spotting complex interactions between various measurements that precede gate changes.

Leveraging the power of neural networks, this method provides a sophisticated tool for classifying gate operations. Neural networks can capture intricate patterns in the data, offering high accuracy in predicting the causes of gate operations. This method shows potential for automating the classification process with minimal human intervention.

Overall, our study demonstrates that automated techniques, particularly those involving clustering and neural networks, can effectively replicate the accuracy of manual detection. These methods reduce the need for constant human oversight, thus enhancing the efficiency and reliability of water management systems. Future work will focus on refining these techniques and integrating them into operational workflows to optimize gate operations and minimize water shortages and surpluses.

4.3 Identifying trends

Clustering using used k-means considering information on one location or multiple locations allowed us to identify both recurring and unusual trends. The clustering at multiple locations can indicate that there is a common pattern before any facility is operated when there is a surplus in the system. The method makes it possible to visualize canal status, allowing operators to understand whether something is normal or abnormal at a glance even without any knowledge of mathematics. VAR-LiNGAM, which is a tool of the causal discovery, found more detailed causal relationships such as effects of the last few minutes.

5 Further research

Future research should challenge other models with more diverse data sets to create better models. In addition, the inclusion of other components of the water management system, such as pumps and spillways, would further increase the efficiency of the overall system. Furthermore, there is still room to investigate the impact of meteorological (weather) data, which is strongly related to climate change.

By continuing to develop and validate these automated methods, we can contribute to more sustainable and resilient water management practices, addressing the challenges posed by climate change and increasing water demand.

6 Conclusion

In order to use water efficiently, this study analyzed irrigation canal flow rates and water levels over a period of approximately six months. Comparison between several methods indicated that level shift detection can adequately replace manual detection, particularly in identifying gate operations. Furthermore, it is suggested that methods similar to those addressed here show promise for the classification of gate operations.

References

[1] Yu Fan, Haorui Chen, Zhanyi Gao, Yumiao Fan, Xiaomin Chang, Mingming Yang, and Benyan Fang. Water distribution and scheduling model of an irrigation canal system. *Computers and Electronics in Agriculture*, 209:107866, 2023. https://www.sciencedirect.com/science/article/pii/S0168169923002545, Accessed: 2024-10-15.

[2] Paulo C.F. Erbisti. Design of hydraulic gates. https://www.taylorfrancis.com/books/mono/10.1201/b16954/design-hydraulic-gates-paulo-erbisti, 2004. Accessed: 2024-10-15.

[3] Won Chang, Michael L. Stein, Jiali Wang, V. Rao Kotamarthi, and Elisabeth J. Moyer. Changes in spatiotemporal precipitation patterns in changing climate conditions. *Journal of Climate*, 29(23):8355–8376, December 2016. http://dx.doi.org/10.1175/JCLI-D-15-0844.1, Accessed: 2024-10-15.

[4] Jinichi Koue, Hikari Shimadera, Tomohito Matsuo, and Akira Kondo. Analysis of the effects of climate change on the gyre in Lake Biwa, Japan. *Journal of Hydroinformatics*, 25(2):243–257, 01 2023. https://doi.org/10.2166/hydro.2023.075, Accessed: 2024-10-15.

[5] UNESCO. World water resources at the beginning of the 21st century. *International Hydrology Series*, 2003. https://assets.cambridge.org/97805218/20851/sample/9780521820851ws.pdf, Accessed: 2024-10-15.

[6] Sandra L. Postel, Gretchen C. Daily, and Paul R. Ehrlich. Human appropriation of renewable fresh water. *Science*, 271, 1996. https://www.science.org/doi/abs/10.1126/science.271.5250.785, Accessed: 2024-10-15.

[7] Stephen R. Carpenter, Stuart G. Fisher, Nancy B. Grimm, and James F. Kitchell. Global change and freshwater ecosystems. *Annual Review of Ecology and Systematics*, 23:119–139, 1992. http://www.jstor.org/stable/2097284, Accessed: 2024-10-15.

[8] Igor A. Shiklomanov. World water resources -a new appraisal and assessment for the 21st century-. https://www.caee.utexas.edu/prof/mckinney/ce385d/papers/Shiklomanov.pdf, 1998. Accessed: 2024-10-15.

[9] Kevin E. Trenberth. The impact of climate change and variability on heavy precipitation, floods, and droughts. *Encyclopedia of Hydrological Sciences*, 2005. https://www2.cgd.ucar.edu/staff/trenbert/books/EHShsa211.pdf, Accessed: 2024-10-15.

[10] Ministry of Land, Infrastructure, Transport, and Tourism. Water resource issues. https://www.mlit.go.jp/mizukokudo/mizsei/mizukokudo_mizsei_tk2_000021.html, 2015. Accessed: 2024-10-15.

[11] Loiy Al-Ghussain. Global warming: review on driving forces and mitigation. *Environmental Progress & Sustainable Energy*, 38(1):13–21, 2019. https://aiche.onlinelibrary.wiley.com/doi/full/10.1002/ep.13041, Accessed: 2024-10-15.

[12] Keyvan Malek, Jennifer C Adam, Claudio O Stöckle, and R. Troy Peters. Climate change reduces water availability for agriculture by decreasing non-evaporative irrigation losses. *Journal of Hydrology*, 561:444–460, 2018. https://www.sciencedirect.com/science/article/pii/S0022169417308119, Accessed: 2024-10-15.

[13] Sean A. Woznicki, A. Pouyan Nejadhashemi, and Masoud Parsinejad. Climate change and irrigation demand: Uncertainty and adaptation. *Journal of Hydrology: Regional Studies*, 3:247–264, 2015. https://www.sciencedirect.com/science/article/pii/S2214581814000524, Accessed: 2024-10-15.

[14] T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley (eds.). Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change. https://www.ipcc.ch/report/ar5/wg1/, 2013. Accessed: 2024-10-15.

[15] Suido-Sangyo newspaper company. Overview of water utilization in the lake biwa-yodo river basin. http://www.byq.or.jp/kankyo/r02/pdf/2022chapter_201.pdf, 2012. Accessed: 2024-10-15.

[16] Keizrul bin Abdullah. Use of water and land for food security and environmental sustainability. *Irrigation and Drainage*, 55(3):219–222, 2006. https://onlinelibrary.wiley.com/doi/abs/10.1002/ird.254, Accessed: 2024-10-15.

[17] Yoshihide Wada, Dominik Wisser, Stephanie Eisner, Martina Flörke, Dieter Gerten, Ingjerd Haddeland, Naota Hanasaki, Yoshimitsu Masaki, Felix T. Portmann, Tobias Stacke, Zachary Tessler, and Jacob Schewe. Multimodel projections and uncertainties of irrigation water demand under climate change. *Geophysical Research Letters*, 40(17):4626–4632, 2013. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50686, Accessed: 2024-10-15.

[18] Masao Ono and Fumio Hasegawa. Project for proposal (project members only), 2024.

[19] Hitachi Power Solutions. What is a stage-discharge rating curve? https://diovista-en.hitachi-power-solutions.com/faq/faq_4_20-what-is-stage-discharge-rating-curve.html, 2023. Accessed: 2024-10-15.

[20] *Appendix: The Hodrick-Prescott Filter*, pages 361–367. John Wiley & Sons, Ltd, 2011.

[21] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python, 2010.

[22] Greg Welch and Gary Bishop. An introduction to the kalman filter, July 2006.

[23] Filippo Maria Bianchi. Time series analysis with python. https://github.com/FilippoMB/python-time-series-handbook, 2024. Accessed: 2024-10-15.

[24] Roger Labbe. Kalman and bayesian filters in python. *Chap*, 7(246):4, 2014.

[25] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. https://www.sciencedirect.com/science/article/pii/S0165168419303494, Accessed: 2024-10-15.

[26] He Zhao. Adtk: A python toolkit for rule-based/feature-based anomaly detection in time series. https://github.com/arundo/adtk, 2019. Accessed: 2024-10-15.

[27] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010. http://jmlr.org/papers/v11/hyvarinen10a.html, Accessed: 2024-10-15.

Appendix A: Executive Summary

Table 8 shows the objectives of this study, each main result, and future challenges (this part was edited by industrial mentors after the submission).

Objectives	Achievement/Methods	Main contribution	Future work
Prepare the dataset for analysis	++ / Manual data preparation	Sensor data was visualized for the upstream of the target canal, and the timing of gate operation was recorded.	Comparing with actual operation timing. Expanding the network to record.
	+++ / Data cleaning and smoothing	Smoothing performance was compared using multiple smoothing methods and parameter settings.	N/A
Identify gate operation events	+++ / Level shift detection	Methods for detecting and labeling significant water flow changes considering the direction of waterways were developed.	N/A
	+++ / Comparison with several filters	Verification using manual detection results showed that the level shift method had high accuracy in detecting significant shifts.	Increase sample size of datasets and further validation.
Classify cause of gate operation	+++ / Machine learning such as feature-based clustering	Automated techniques such as clustering demonstrated that effectively replicate the accuracy of manual detection.	N/A
	++ / Time series causal discovery	VAR-LiNGAM, one of the causal discovery tool, found causal effects in water flows over the past tens of minutes.	More detailed parameter studies and discussions, especially regarding smoothing.
Suggest methods to automation	**N/A** / Not achieved due to time running out	N/A	Integrating and embedding into water management systems and IoT devices.

Table 8: Executive summary.

The definition of each symbol for Achievement is as follows: +++: Very good, ++: Good, +: Slightly good, -: Bad, and N/A: Not applicable.

Appendix B: VAR-LiNGAM

B.1 Estimation procedure

Here, refer to reference [27], we supplement an estimation algorithm for the VAR-LiNGAM parameters in Equation 3.9. This process can be summarized as following two steps:

1. For $x(t)$ in Equation 3.9, estimate coefficient matrices and error variables as a VAR model. By excluding the components at the same time, i.e., $\tau = 0$, the classical AR method can be applied.
2. For the calculated $e(t)$, estimate an adjacency matrix B_0 using DirectLiNGAM, which is an estimation method in the original LiNGAM. Using the adjacency matrices obtained above, calculate adjacency matrices B_τ with each lag τ.

To estimate the optimal lag time k in Equation 3.9, we can use an algorithm based on BIC (Bayesian Information Criterion). That is, we estimate coefficients of VAR-LiNGAM for each lag time k, and the model with the smallest BIC among them is selected as the final estimation result.

B.2 Adjacency Matrices calculated in Section 3.4.3

This section is withheld from publication based on the judgment of industrial mentors, considering the non-disclosure agreement.