An Effective and Efficient Method for Identification of Contamination Sources in Water Distribution Systems Based on Manual Grab-Sampling

Ji, Yiran; Zheng, Feifei; Du, Jiawen; Huang, Yuan; Bi, Weiwei; Duan, Huan Feng; Savic, Dragan; Kapelan, Zoran

DOI
10.1029/2022WR032784

Publication date
2022

Document Version
Final published version

Published in
Water Resources Research

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE

10.1029/2022WR032784

Key Points:

- A new manual grab-sampling method is proposed to localize continuous water distribution system (WDS) contamination sources
- The new method employs a dynamic and cyclical sampling strategy based on WDS hydrants
- Proposed method is effective and efficient to localize contamination sources for many scenarios

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

F. Zheng,
feifeizheng@zju.edu.cn

Citation:

Ji, Y., Zheng, F., Du, J., Huang, Y., Bi, W., Duan, H.-F., et al. (2022). An effective and efficient method for identification of contamination sources in water distribution systems based on manual grab-sampling. *Water Resources Research*, *58*, e2022WR032784. https://doi.org/10.1029/2022WR032784

Received 11 MAY 2022
Accepted 16 OCT 2022

An Effective and Efficient Method for Identification of Contamination Sources in Water Distribution Systems Based on Manual Grab-Sampling

Yiran Ji[1], Feifei Zheng[1], Jiawen Du[1], Yuan Huang[2], Weiwei Bi[3], Huan-Feng Duan[4], Dragan Savic[5,6,7], and Zoran Kapelan[8]

[1]College of Civil Engineering and Architecture, Zhejiang University, Zhejiang, China, [2]College of Water Conservancy & Hydropower Engineering, Hohai University, Nanjing, China, [3]College of Civil Engineering, Zhejiang University of Technology, Zhejiang, China, [4]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, [5]KWR Water Research Institute, Nieuwegein, The Netherlands, [6]Centre for Water Systems, University of Exeter, Exeter, UK, [7]Universiti Kebangsaan Malaysia, Bangi, Malaysia, [8]Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

Abstract Most of the contamination source localization methods for water distribution systems (WDSs) assume the availability of accurate water quality models and multi-parameter online sensors, which are often out of reach of many water utilities. To address this, a novel manual grab-sampling method (MGSM) is developed to effectively and efficiently locate continuous contamination sources in a WDS using a dynamic and cyclical sampling strategy. The grab samples are collected at a pre-specified number of hydrants by the corresponding teams followed by laboratory tests. The MGSM optimizes the sampling plan at each cycle by making the probability of contamination source(s) in each sub-network as equal as possible, where sub-networks are determined by the selected hydrants and current flow pipe directions. The CS's size is reduced at each cycle by exploiting sample testing results obtained in the previous cycle until there are no further hydrants to sample from. Two real-world WDSs are used to demonstrate the effectiveness of the proposed MGSM. The results obtained show that the MGSM can significantly reduce the spatial range of the CS (to about 5% of the entire WDS) for a range of scenarios including multiple contamination sources and pipe flow direction changes. We found that an optimal number of sampling teams exists for a given WDS, representing a balanced trade-off between detection efficiency and sampling/testing budgets. Due to its relative simplicity, the proposed MGSM can be used in engineering practice straightaway and it represents a viable alternative to the methods associated with water quality models and sensors.

1. Introduction

A water distribution system (WDS) represents a basic lifeline infrastructure that closely relates to the daily life and health safety of its served population (Qi et al., 2018). Typically, a WDS is spatially distributed and thus inherently vulnerable to accidental and/or intentional contamination intrusion (Ostfeld et al., 2014; Yang & Boccelli, 2016; Zhang et al., 2020). For instance, over a 5-day period in October 2007, a boil-water notice was served on the majority of Oslo, Norway, as a result of a combination of bacteriological, Cryptosporidium oocysts and Giardia cysts found in the samples taken from the WDS (Robertson et al., 2008). More recently, on July 26, 2020, a contamination event was reported in Hangzhou, China, where a sewer pipe was misconnected to a drinking water pipe in a small suburb (ChinaNews, 2020). Unfortunately, these events were not detected by the water quality warning systems of the local water utilities. The events were reported by the residents and/or diagnosed by the hospitals. This implies that monitoring and protecting water quality safety are still nontrivial challenges for many WDSs (Asheri Arnon et al., 2019).

To secure water quality safety in a WDS, extensive studies have been carried out to develop contamination response systems (CRSs) (Giudicianni et al., 2020). In principle, an effective CRS should at least consist of a contamination warning and source identification (Rodriguez et al., 2021). Regarding the contamination warning, a straightforward manner is to deploy online water quality sensors within the WDS (Hart & Murray, 2010). A warning is triggered once the concentration of some particular water quality parameters (e.g., pH and turbidity) is above or below the sensor's safety threshold. Ideally, placing a sensor at each possible location in the WDS can maximize the capability to generate a warning when a contamination intrusion event occurs (Zheng et al., 2018).

However, it is difficult, if not impossible, to implement this approach due to the high capital and maintenance costs associated with so many water quality sensors (Winter et al., 2019).

Consequently, many studies have focused on optimally deploying a limited number of water quality sensors to maximize their detection/warning performance (Rathi & Gupta, 2014). These studies range from the use of different objective functions to identify appropriate water quality sensor placement strategies (He et al., 2018; Naserizade et al., 2018), to the development of various algorithms to enable effective optimization on this design problem (Hu et al., 2017). More recently, efforts have been increasingly made to identify design solutions that provide a resilient water quality sensor strategy. The approach does not only perform well when all sensors function perfectly, but also can detect contamination events even under possible sensor failures (Ostfeld et al., 2008; Zhang et al., 2020). Typically, the objective functions designed for the water quality sensor placement problems are very complex as different aspects of contamination detection need to be taken into account (e.g., detection likelihood, detection time delay, sensor reliability, different consequences of non-detection, and various uncertainties; Khorshidi et al., 2018). Studies have been undertaken to develop various algorithms to effectively identify optimal water quality sensor placement strategies based on these objective functions (Ung et al., 2017). Specifically, those studies focus on developing either sophisticated search algorithms that enhance the design solution's quality (Di Nardo et al., 2018; Hu et al., 2020) or advanced water quality modeling approaches that improve the optimization efficiency (Naserizade et al., 2018; Ohar et al., 2015).

In parallel to the research progress on the early warning systems for contamination detection, efforts have also been made to develop various algorithms for sourcing/localizing the contamination injection locations according to the analysis of sensor data (Preis & Ostfeld, 2007). These developments started by using traditional optimization techniques, such as linear programming (LP) scheme (Preis & Ostfeld, 2006). This was followed by the use of various evolutionary algorithms (EAs) as they possess superior search capabilities compared to the traditional LP and nonlinear programming (NLP) techniques (Hu et al., 2015; Li et al., 2021; Preis & Ostfeld, 2008). While these algorithms have reliable performance in locating contamination sources in hypothetical case studies, their practical application can be challenging. This is mainly due to the "equifinality" issue associated with the identification of the source of the incident (Jia, Zheng, Zhang, et al., 2021), where many different injection scenarios (contaminant concentration and starting time) indicate a similar contamination impact. To address this issue, Bayesian-based approaches have been proposed to identify contaminant sources, where the location with the highest posterior probability is interpreted as the most plausible (Jerez et al., 2021; Sankary & Ostfeld, 2019; Yang & Boccelli, 2014). More recently, machine learning algorithms have been increasingly employed to facilitate contamination localization, such as the Random Forest algorithm (Grbčić et al., 2020) and Convolutional Neural Network (Sun et al., 2019).

Detailed analysis of previous studies in terms of the CRS research shows that the majority of contamination warning and source identification methods rely heavily on an accurate water quality model (Vrachimis et al., 2020). This is one of the main reasons that may hinder their implementation as a well-calibrated water quality model is usually not available for many water utilities (Sankary & Ostfeld, 2018). In addition, existing water quality modeling techniques are still incapable of accurately reproducing contaminant reaction dynamics in WDSs, especially for biochemical contaminants (Hart et al., 2019). While online sensors may provide reliable warning information by measuring the contaminant concentration in real-time, they generally can only measure a limited number of water quality parameters such as pH, turbidity, chlorine, and conductivity (Sun et al., 2019). Consequently, many other contaminants, such as organics and pathogenic microorganisms, cannot be detected with certainty using online in-situ sensors. In addition, water quality sensors are often expensive in both the purchase and maintenance, especially for advanced sensors that are used to measure complex substances (He et al., 2018). Therefore, the water quality sensors are often sparsely distributed in many WDSs (Ostfeld et al., 2014).

The contamination events within the WDS can be classified into three different types, which are intentional events (Type 1), accidental events (Type 2), and events caused by WDS itself (Type 3). For Type 1, the contamination can be toxic substances that are intentionally injected into the WDS, typically during a short time period. Such events can result in serious consequences and hence need a quick response at all costs (Ostfeld et al., 2014). Type 2 is often represented by the misconnections between water supply pipes and greywater/sewer pipes that have been reported in China (He et al., 2018). Type 3 can be caused by structural damages to pipes (e.g., contamination due to pipe corrosion or leaks; Zhang et al., 2020) or biochemical substances (e.g., microorganisms) activated by the water at a particular level of turbulence (He et al., 2019).

10.1029/2022WR032784

Typically, within Types 2 and 3, the contamination exists *continuously* in the WDS until the source(s) is localized and eliminated. These contamination substances (e.g., metal, microorganism, and organic) often have the following properties: (a) they can be colorless and tasteless, and hence cannot be directly detected by tap-water users; (b) they do not induce quick, serious public health consequences (i.e., this study focuses on the contamination events with chronic but no acute health effects) and hence their source(s) localization needs to be conducted without interrupting water supply; and (c) they may not be directly detected by online water quality sensors as the majority sensors typically monitor simple quality parameters such as chlorine, pH, turbidity, and conductivity. These properties motivate the development of the proposed manual grab-sampling method (MGSM) to efficiently and effectively identify continuous contamination sources of Types 2 and 3 in WDSs. This is particularly the case for Type 3 events as there are a number of practical issues that cause such contamination events and utilities are interested in locating these sources in the most efficient way possible.

The proposed MGSM is an iterative MGSM to enable effective contaminant detection and localization. This is followed by gathering comprehensive water quality parameter information with the aid of laboratory tests. The MGSM is particularly useful for cases where the online quality sensors are sparsely distributed (or completely unavailable) or sensors cannot measure the contaminants (Wong et al., 2010). The MGSM does not need water quality modeling and can identify the contamination location without encountering the "equifinality" issue. In addition, for the cases where the labor is plentiful with low cost, the MGSM is preferred as it provides the spatial distribution of water quality measurements at a reduced cost when compared to fixed sensors (Mann et al., 2012). Therefore, manual grab-sampling can be an important strategy for water utilities interested in water quality safety in the WDS, which can supplement the information obtained from existing online sensors.

Despite the merits and practical significance of the MGSM for the cases with sparsely distributed sensors and relatively low labor costs, relevant research on this topic is surprisingly rare. Among few relevant studies, one significant example is from the work of Wong et al. (2010), where a Mixed-Integer Linear Programming formulation is proposed to determine optimal locations for manual grab sampling after a contamination event is detected in a WDS. In their study, the optimal manual grab sample locations are identified by maximizing the total pair-wise distinguishability of candidate contamination events (eliminate unlikely events as much as possible). While Wong et al. (2010) showed that a contamination event can be identified by their proposed method with significantly improved efficiency, its success was conditioned on a few critical assumptions. These assumptions include: (a) each node in the WDS has an equal probability of being the source of contamination intrusion, (b) only one contamination event can occur in the WDS, and (c) the pipe flow direction cannot change during the entire sampling process. However, these assumptions can significantly violate the real conditions as the contamination intrusion can occur at any pipe location and a long pipe is typically associated with a higher contamination probability (He et al., 2018). Furthermore, although the probability of simultaneous multiple contamination intrusions is low, their occurrence is still possible in large WDSs (Butera et al., 2021). In addition, flow direction changes are likely to occur in some pipes in a large WDS with multiple supply sources (Qi et al., 2018).

The main contribution of this paper is the proposal of an improved water quality MGSM for detecting and localizing continuous contamination sources in WDSs. The newly developed method employs a dynamic and cyclical sampling strategy based on the hydrant locations in a WDS. The novel aspect of the proposed method is the simple and effective way developed to split the network after each round of sampling, thereby significantly enhancing the efficiency of the entire detection process. In addition, the proposed method is novel in that the optimal sampling locations are determined by making the probability of contamination source in each sub-network based on the current flow pipe directions as equal as possible at each cycle. The results of these samples are subsequently analyzed and employed to drive the sampling strategy for the next cycle. It is highlighted that the proposed MGSM is an alternative to these literature methods (sensor-based methods) in the cases where: (a) sensors are sparsely distributed or not available (e.g., lack of existence of suitable sensors), (b) the low-cost labor force is available, and (c) the contamination events have slow or low impacts to the water quality in the WDSs.

2. Methods

The basic premise of the proposed MGSM is: (a) select a given number of sampling points (hydrants of the WDS) in the studied area based on the testing capacity of the laboratory (i.e., the number of samples that can be tested simultaneously) and the number of sampling teams, with all pipes within the candidate area considered as possible contamination sources, (b) narrow down the range of the candidate areas containing contamination source(s)

10.1029/2022WR032784

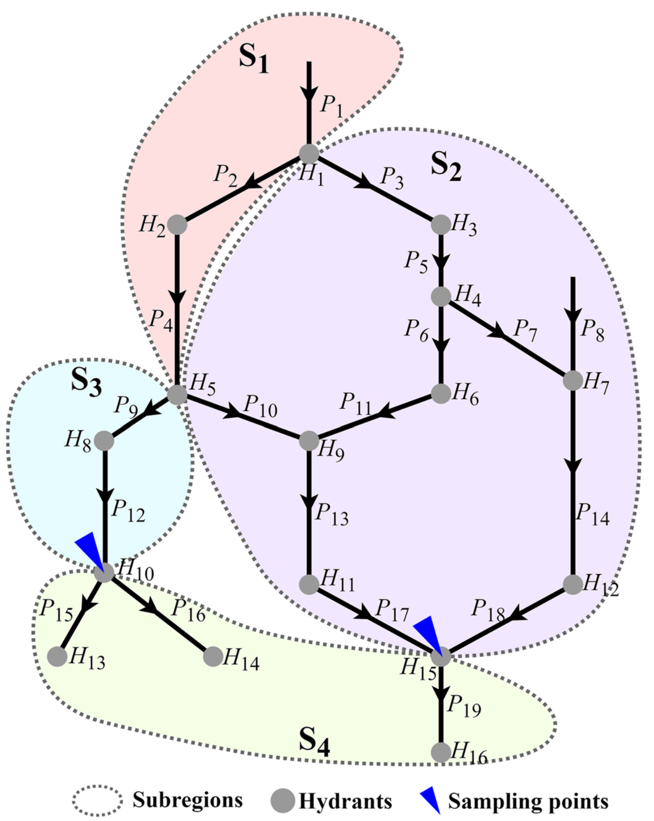


Figure 1. Illustration of the WDS sub-networks identified by the proposed MGSM based on two sampling locations, with arrows representing pipe flow directions.

based on sample testing results, and (c) repeat steps (a) and (b) until the range of candidate areas with contamination source(s) cannot be further narrowed down. The key to effectively implementing this new MGSM is how to automatically select the appropriate hydrants in each cycle of the above methodology to reduce the total number of cycles, thereby quickly localizing the pollution source(s) in the WDS. It is noted that every length of pipe between two hydrants within the WDS is considered as a contamination source. Therefore, the proposed MGSM can account for both scenarios where the contamination sources are in pipes or junctions. While the proposed MGSM is demonstrated using hydrants in this study, any other sampling facilities (e.g., taps) can be easily handled by simply treating them as hydrants within the algorithm implementation.

Section 2 presents the details of the proposed MGSM, including the associated theoretical foundations (e.g., the development of the objective function), the MGSM algorithm structure, the illustration of the proposed MGSM, and the optimization method to implement the MGSM.

2.1. Theoretical Foundations for the Proposed MGSM

Section 2.1 introduces the theoretical foundations of the proposed MGSM, including the proposal of a method to enable the WDS partitioning and the development of the objective function of the proposed MGSM. The details are given below.

2.1.1. WDS Partitioning Based on Sampling Locations and Flow Directions

As previously stated, the proposed MGSM attempts to identify the optimal sampling locations (hydrants) at each cycle, aimed to minimize the total number of cycles (equivalent to the efficiency and cost of the entire process). Within the MGSM, the entire WDS is partitioned into different sub-networks based on sampling locations and flow directions at a given point in time. Specifically, if a hydrant H in the system is selected as the sampling point, all pipes in the WDS can be divided into two sub-networks: all upstream pipes relative to the selected hydrant H, denoted as $\mathbf{U}_H$, and remaining pipes whose flows do not go through H, denoted as $\mathbf{N}_H$. If two hydrants (H_1 and H_2) are selected as the sampling points, four sub-networks can be identified, respectively representing the common group of pipes upstream of both selected hydrants ($\mathbf{U}_1 \cap \mathbf{U}_2$), the unique group upstream of one hydrant only ($\mathbf{U}_1 \cap \mathbf{N}_2$ and $\mathbf{U}_2 \cap \mathbf{N}_1$), and not the upstream of both hydrants ($\mathbf{N}_1 \cap \mathbf{N}_2$). Using this process, for a number of n sampling points in a WDS, for example, $\{H_1, H_2, \ldots, H_n\}$, a total of $T = 2^n$ sub-networks, $\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_T\}$, can be obtained theoretically.

Figure 1 illustrates how the proposed MGSM identifies the WDS sub-networks based on two sampling locations. A total of 16 hydrants are available that can be considered as the potential sampling points, where the arrows represent pipe flow directions. For illustration, hydrants 10 (H_{10}) and 15 (H_{15}) are selected as sampling points to enable network partitioning. Four different sub-networks are identified using the proposed MGSM, which are $\mathbf{S}_1 = \{P_1, P_2, P_4\}$, $\mathbf{S}_2 = \{P_3, P_5, P_6, P_7, P_8, P_{10}, P_{11}, P_{13}, P_{14}, P_{17}, P_{18}\}$, $\mathbf{S}_3 = \{P_9, P_{12}\}$, $\mathbf{S}_4 = \{P_{15}, P_{16}, P_{19}\}$. It can be observed that pipes in $\mathbf{S}_1$ are in the common upstream group for H_{10} and H_{15} and flows for pipes in $\mathbf{S}_4$ do not go through any of the two hydrants. Pipes in $\mathbf{S}_2$ are those that are upstream of H_{15} but not H_{10}, and Pipes in $\mathbf{S}_3$ are upstream of H_{10} but not H_{15}.

For the n sampling points $\mathbf{A} = \{H_1, H_2, \ldots, H_n\}$, the outcome of the test at each sampling point is either that the sample is contaminated or non-contaminated. Therefore, there are 2^n possible results for n sampling points, in which each contaminated outcome corresponds to the contamination source being located in a certain sub-network or many sub-networks when contaminations are found in many sampling locations. For example, if the contamination is detected at both H_{10} and H_{15}, as in Figure 1, it can be derived that the contamination source(s) may be located in the common upstream group of pipes ($\mathbf{S}_1$ in Figure 1). The source can also be in the two sub-networks ($\mathbf{S}_2$ and $\mathbf{S}_3$) upstream of one of the two sampling locations. When only one sampling point indicates contamination, it can be determined that the source is located in the area upstream of the sampling point where contamination is detected, that is, $\mathbf{S}_2$ or $\mathbf{S}_3$. When results show no contamination at both sampling points, then the contamination source(s) is located in an area outside all the upstream parts of the two sampling points, that is, $\mathbf{S}_4$ in Figure 1. This is the basic localization principle used in the proposed MGSM in this study.

Once a sub-network or a few sub-networks are selected as potential contamination sources based on the sample testing results, all pipes in this/these sub-network(s) are considered as candidates. This is followed by the further use of the partitioning method to narrow down the spatial range to localize the source. In other words, the network partitioning needs to be carried out at each cycle of the entire sampling process based on the updated candidate pipes with potential contamination sources.

2.1.2. The Development of the Objective Function of the Proposed MGSM

Conditioned on the identified T sub-networks, the mathematical expectation ($E(\mathbf{A})$) of a given set of sampling points ($\mathbf{A}$) in localizing the location of the contamination source can be expressed as

$$E(\mathrm{A}) = \sum_{i=1}^{T} p_i \cdot L_i \tag{1}$$

where p_i is the probability of the ith sub-network that has the contamination source, and L_i is the corresponding total pipe length of this sub-network. Since the proposed MGSM mainly aims to detect contamination Types 2 and 3 (see Section 2 for details), the probability of a contamination source being located on each unit length of pipe can be considered identical. This results in the probability of the contamination source being in any sub-network i equal to the ratio of the pipe length of the sub-network L_i to the total pipe length L_{all} in the entire WDS. Mathematically, it gives,

$$E(\mathrm{A}) = \sum_{i=1}^{T} \frac{L_i}{L_{all}} \cdot L_i = \frac{1}{L_{all}} \sum_{i=1}^{T} L_i^2 \tag{2}$$

Thus, the objective function for calculating the optimal sampling group can be expressed as follows:

$$\text{Minimize}: F(\mathrm{A}) = \frac{E(\mathrm{A})}{L_{all}} = \frac{1}{L_{all}^2} \sum_{i=1}^{T} L_i^2 \tag{3}$$

where $F(\mathrm{A})$ is a dimensionless number by dividing $E(\mathbf{A})$ using L_{all}, representing the ratio of candidate area with contamination source identified by the sampling group relative to the total pipe length of the entire WDS being considered. $\mathbf{A}$ is the decision variables, representing the hydrant sampling strategy. The minimization of $F(\mathrm{A})$ physically indicates a minimum pipe length of the sub-network with contamination source(s) to be identified by the selected sampling points.

Cauchy-Schwarz Inequality (Bhatia & Davis, 1995) can be used to further explain the minimization of Equation 3, which is

$$T \times \left(L_1^2 + L_2^2 + \cdots + L_T^2\right) \geq (L_1 + L_2 + \cdots + L_T)^2 \tag{4}$$

$$\text{Namely } F(\mathrm{A}) = \frac{1}{L_{all}^2} \sum_{i=1}^{T} L_i^2 \geq \frac{1}{T} \tag{5}$$

For $L_1 = L_2, \ldots, = L_T$, the equation holds. Under this condition, when only one hydrant is selected as the sampling point in each cycle, the optimal hydrant divides the WDS into two sub-networks such that the pipe length of its upstream section is half of the total length. When n hydrants are selected as the sampling points in each cycle, theoretically, the optimal hydrant group bisects the WDS to 2^n sub-networks with identical pipe lengths across different sub-networks. In other words, the minimization of Equation 3 (i.e., $L_1 = L_2, \ldots, = L_T$) can be interpreted as using a specified number of sampling points to assign the pipes into T sub-networks with the minimum difference in pipe length at each cycle. This is equivalent to the bi-section approach in computer science, and hence it is expected that such a method can achieve a statistically efficient sampling strategy to localize the contamination

```
Specify the number of sampling points n
Set the cycle c=1, flag=0, the candidate sub-network (CS) as the entire WDS
While True
{
      If n = 1
      {
            Select n sampling hydrant for the CS by minimizing Equation (3)
            Update the CS according to sample testing results
               Case A1: the sample is contaminated
                           Select the sub-network (US) upstream of the selected hydrant
               Case A2: the sample is contamination free
                           Select the sub-network that is not the upstream of the selected hydrant
            c = c + 1
      }
      Else
      {
            If flag=0
            {
               Select n sampling hydrants for the CS by minimizing Equation (3)
               Update the CS according to sample testing results using Selection strategy 1 (SA1):
                 Case B1: more than one hydrant sample are contaminated
                             Select the common sub-network (CUS) upstream of the contaminated hydrants
                             Set flag=1
                 Case B2: only one hydrant sample is contaminated
                             Select the unique sub-network upstream of the contaminated hydrant
                 Case B3: no hydrant samples are contaminated
                           Select the sub-network that is not the upstream of the selected hydrant
               c = c + 1
            }
            If flag=1
            {
               If the CUS exists and its most downstream hydrant is not sampled
                    Assign one sample point at the most downstream hydrant of the CUS
                    Select n-1 sampling hydrants for the CS by minimizing Equation (3)
                    Update the CS(s) according to testing results at the end hydrant
                        If the end hydrant is contaminated
                           Select the CS using the SA1 mentioned above
                        Else
                           Selection strategy 2 (SA2):
                           Select the CS(s) as the union of USs of the hydrants showing evidence of
                           contamination minus the union of USs of contamination-free hydrants and the CUS
                    c = c + 1
              Else If the CUS does not exist
                    Select the CS(s) using the SA2 mentioned above
             Set flag=0
            }
      }
      If no hydrant can be sampled in the selected sub-network
            break
}
```

Figure 2. The algorithm of the proposed MGSM.

source. It is noted that the proposed optimization method may not be able to guarantee global optimality, but it can offer a near-optimal solution that can be efficiency found at each cycle.

The pipe length is used to split the WDS in this study due to its simplicity and efficiency. However, a more refined method may need to account for water velocities or flow volumes, both of which can be correlated with pipe diameters and can account for the amounts of contaminants moving through the pipes. Therefore, partitioning the WDS with the aid of both pipe length and water velocity can be an important future research focus.

2.2. The Algorithm of the Proposed MGSM

The implementation of the proposed MSGM can be triggered by (a) the routine water quality checks done by the water utilities, (b) abnormal signals from the sparsely distributed online water quality sensors if they are installed (e.g., chlorine sensors), where these sensors with a rather limited number are often installed at the outlets of the districted metering areas (DMAs) or WDSs, or (c) contamination warning based on test results of samples at the outlets of the DMAs or the important locations within the WDS area. Figure 2 shows the algorithm details of the proposed MGSM in localizing contamination source(s). As shown in this figure, when the number of sampling locations at each cycle is $n = 1$, the sampling hydrant is selected by minimizing Equation 3, where the minimization method is elaborated in Section 2.4. The candidate sub-network (**CS**) that may contain contamination source(s) is updated at each cycle based on the sample testing results (Cases A1 and A2 in Figure 2). If n is greater than 1, the algorithm of the proposed MGSM becomes more complex, with details given in Figure 2. At the beginning (i.e., flag = 0, and the MGSM is triggered), the n optimal sampling locations are identified by minimizing Equation 3 for the entire WDS being considered (i.e., **CS** is the entire WDS). This is followed by the application of selection strategy 1 (SA1) to update the **CS** for the next cycle, where three different cases (Cases B1, B2, and B3) can be available. For Case B2 (only one sample hydrant has contamination) and Case B3 (all sample hydrants are contamination free), it is straightforward to select the **CS** for the next cycle as shown in Figure 2.

When more than one sample hydrant is contaminated (Case B1), the common upstream sub-network (**CUS**, which is theoretically available) is selected as the **CS** for the next cycle ($c = c + 1$). If this **CUS** exists and its most downstream hydrant is not sampled, one sampling location is assigned to this hydrant. The remaining $n-1$ sampling locations are determined by minimizing Equation 3. The **CS**, which is temporally considered as the **CUS**, is now updated using the following method based on test results of the most downstream hydrant. If that hydrant is contaminated, the **SA1** is employed to update the **CS**; otherwise, the **SA2** (see Figure 2) is used to update the **CS**. Specifically, the **SA2** selects the **CS**(s) as the union of all upstream sub-networks (**US**s) of hydrants where contamination was detected, minus the union of **US**s of contamination-free hydrants and the **CUS**. Note that if the selected **CUS** does not exist in the WDS, the **SA**2 is used to update the **CS**(s).

The proposed MGSM in Figure 2 can handle both the single and multiple contamination sources in a DMA of a WDS. However, each MGSM run identifies only a sub-network that contains a contamination source of the smallest spatial extent. This identified region may need to be blocked for engineering operations (e.g., disconnect the misconnections, repair the leaks, or replace the pipes), to remove the contamination source(s). Sampling tests with a few contaminated hydrants may indicate the presence of multiple contamination sources in different WDS regions. For such cases, once the identified contamination source(s) is fixed, the proposed MGSM can be applied to the potential **CS**s (instead of the entire WDS) derived by the sampling test results combined with knowledge of pipe flow directions. Such a CS selection can be easily performed by engineering experience, but it is difficult to be shown by formal procedures. However, it is also straightforward to apply the MGSM to the entire WDS to identify the other contamination source(s), after the localized source(s) are fixed or isolated.

The methodology assumes that all hydrants selected in one cycle can be sampled at the same flow direction status. This assumption is practically reasonable as the time required to grab samples is often short and the frequency of flow direction change is typically low (e.g., once a day; Wong et al., 2010). While flow direction changes may exist within the supply boundary of some real large WDSs, its associated region is often rather small. Therefore, the change of the flow directions will not significantly affect the application of the proposed MGSM. If the WDS region with changing flow direction is large and known, it can be easily accounted for by the proposed MGSM subject to an important assumption. This assumption is that the time between the start of the flow direction change and the next sampling cycle is significantly greater than the longest travel time from the source to the sample locations. In other words, the contaminant distribution is assumed to be consistent with the current flow regime and without residual effects from the previous flow regime. Based on this assumption, the flow direction changes can be considered by the WDS partitioning process as described in Section 2.1.1, which would accordingly affect the selection of sub-networks and hence the identification of the optimal sampling locations (Equation 3).

2.3. Illustration of the Proposed MGSM

The proposed MGSM is illustrated with two scenarios, including the single contamination source and the two contamination sources simultaneously existing in the WDS, with details given below.

2.3.1. Single Contamination Source

We first illustrate the application of the proposed MGSM (Figure 2) using a single contaminating source as shown in Figure 3. The single contamination source is in P_2, and two sampling locations ($n = 2$) are identified at each cycle. At the first cycle, the entire WDS is set as a candidate sub-network (**CS**), and a total of 120 sampling combinations (2 out of 16 total hydrants) are possible. The mathematical expectations (Equation 3) corresponding to these 120 combinations are calculated by enumeration and the combination with the minimum $F(\mathbf{A})$ value is selected. Consequently, two hydrants $\{H_{11}, H_{12}\}$ are identified as the sampling points yielding the lowest objective function value (Equation 3), as shown in Figure 3a. Based on the assumed location for the contamination

10.1029/2022WR032784

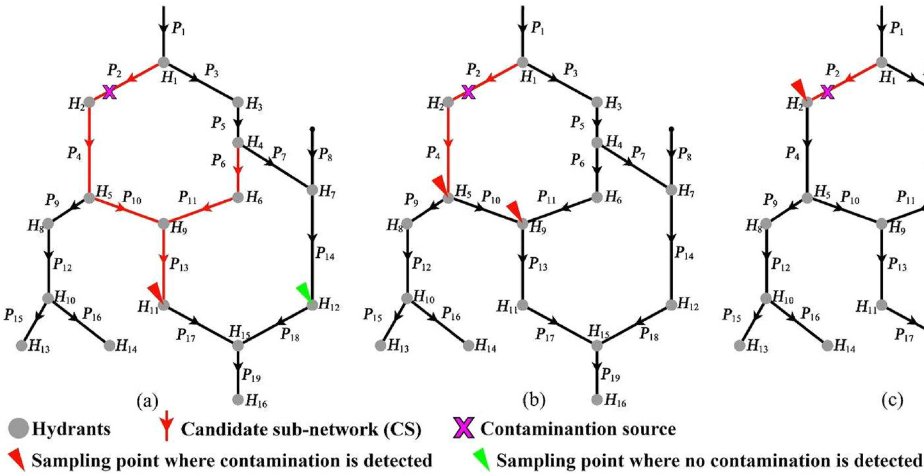


Figure 3. Source localization process for the contamination at P_2: (a) the first cycle ($c = 1$) of sampling and testing; (b) sampling and testing at $c = 2$; (c) sampling and testing at $c = 3$.

source, the sample from hydrant H_{11} is contaminated while the sample from H_{12} is not based on the laboratory tests. Therefore, the **CS** is updated to be a unique sub-network upstream of H_{11} (and not pipes upstream of H_{12}) based on Case B2 in Figure 2, that is, the red pipes shown in Figure 3a.

In the second cycle of sampling, the mathematical expectations corresponding to different hydrant groups are calculated according to the updated **CS** determined in the previous cycle. The resultant optimal strategy is the combination of H_5 and H_9 as it produces the lowest objective function value. Testing results on these two hydrant samples show that both are contaminated, indicating that the contamination source exists in the common upstream sub-network (**CUS**) of H_5 and H_9. Therefore, the **CS** is updated as the **CUS** based on Case B1 (Figure 2), which is $\{P_2, P_4\}$ as represented by red lines in Figure 3b. In the third cycle of sampling, there is only one hydrant location, H_2, so the contamination source is successfully detected on P_2, which is the exact location of the contamination source.

2.3.2. Two Contamination Sources

Figure 4 illustrates the application of the proposed MGSM (Figure 2) in dealing with two contamination sources. In this figure, the contamination sources are in P_7 and P_{10}, and two sampling locations ($n = 2$) are identified at each cycle. As the same with the single contamination source in Figure 3a, the hydrants H_{11} and H_{12} are selected as the sampling points at the first cycle by minimizing Equation 3 (the enumeration method is used for this small WDS). The testing results show both hydrants are contaminated, and accordingly, the **CS** is updated to be the common upstream sub-network (**CUS**, red pipes in Figure 4a) using Case B1 in Figure 2. Since the **CUS** exists and its most downstream hydrant (H_4) is not sampled, H_4 is selected as one sampling location and the other location (H_1) is identified with the aid of Equation 3 in the second cycle ($c = 2$).

Based on the locations of the two contamination sources, the end hydrant H_4 should show no contamination in the laboratory test and selection strategy 2 (**SA2**) is used to update the **CS**. More specifically, for such cases, the **CS** can be described as **UA**-**UB**-**CUS** (**CUS** = $\{P_1, P_3, P_5\}$), where **UA** is the union of sub-networks (**US**s) upstream of contaminated hydrants (i.e., H_{11} and H_{12} at $c = 1$) and **UB** is the union of **US**s sampling hydrants without contaminations (it is null at $c = 1$). This is followed by the application of the proposed method at $c = 3$, where two hydrants (H_5 and H_9) are selected as the sampling points. The resultant **CS** is P_{10} using Case B2 in Figure 2 based on test results (H_5 is not contaminated, but H_9 is), which is the unique upstream sub-network of H_9. Since no hydrants can be sampled in the current **CS** (i.e., P_{10}), P_{10} is successfully identified with the contamination source. The run of the proposed MGSM (Figure 2) is finalized.

To identify the second contamination source in P_7, the s localized source in P_{10} needs to be fixed or isolated before the next MGSM run. This is because the proposed MGSM identifies only one contamination source per run. Prior to the application of the next MGSM run, the identified contamination source(s) need to be eliminated. In addition, all the test results of hydrant samples during the previous MGSM run and pipe flow direction

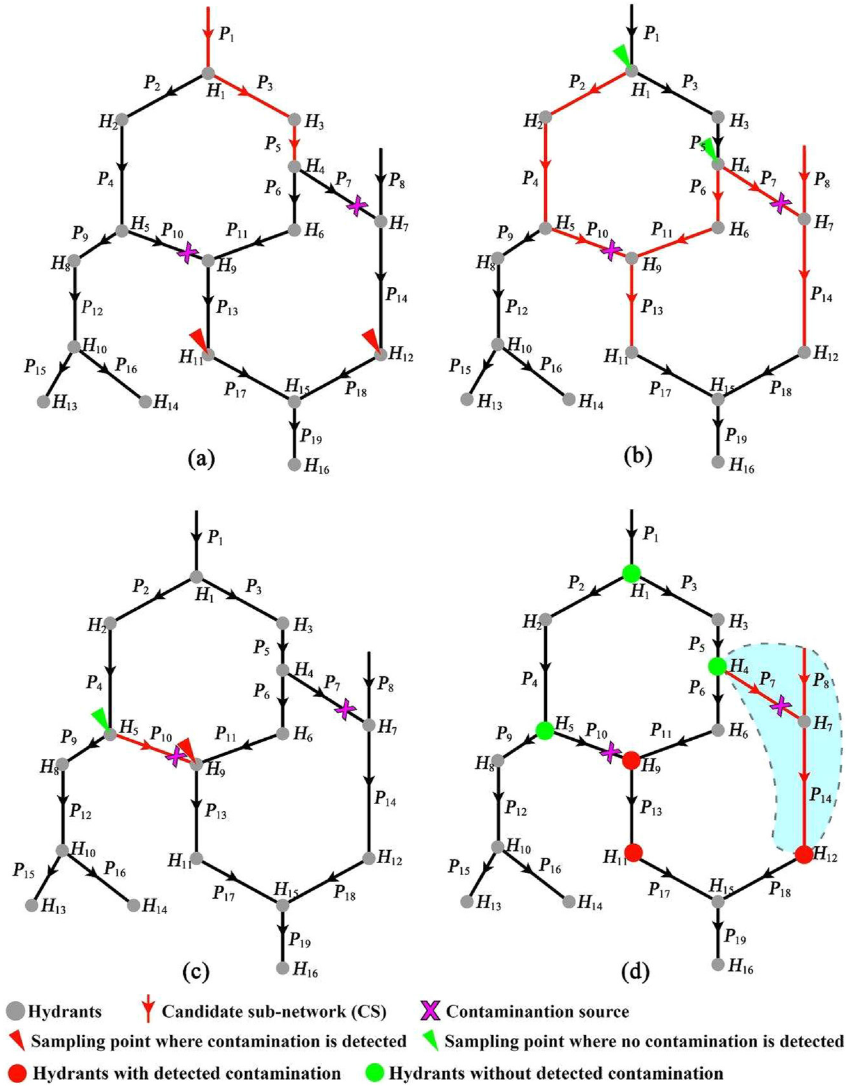


Figure 4. Source localization process for two contamination cases at P_7 and P_{10}: (a) the first cycle ($c = 1$) of sampling and testing; (b) sampling and testing at $c = 2$; (c) sampling and testing at $c = 3$; (d) the CS identified (shaded pipes) for the next MGSM run, where the red and green dots represent test results of the previous MGSM run.

information can be jointly used to derive the potential **CS** for the next MGSM run. For the given example, the **CS** can be identified as the red pipes in Figure 4d based on the test results of the previous MGSM run (red and green dots) since (a) the test on H_4 shows no contamination but H_{12} does, and (b) the identified source at P_{10} is not upstream of H_{12}. This **CS** is only a small proportion of the entire WDS, thereby greatly improving the efficiency of the next MGSM run. However, for cases when the **CS** cannot be determined by the existing information provided by sample test results and pipe flow directions, the entire WDS (after the identified contamination source(s) is eliminated) is considered as the **CS** again to enable the application of the proposed MGSM.

In this subsection, one and two contamination sources are used to illustrate the proposed MGSM due to the high likelihood of those events occurring in real WDS. In addition, two sampling locations are used at each cycle for illustration purposes, where the pipe flow directions are not changed. However, the application procedures with details given in Figure 2 are generic, and hence can be applied to other scenarios such as different number of sampling locations, different contamination sources, and the WDS with possible pipe flow changes (further explanation of which is given in Section 4).

10.1029/2022WR032784

2.4. Optimization Method to Minimize the Objective Function

As shown in Figure 2, the proposed MGSM algorithm requires an optimization method to minimize the objective function (Equation 3). While the enumeration method can be effective when dealing with small WDSs and with a low number of sampling locations at each cycle, it is computationally intractable for real and large WDSs. More specifically, for a case with n sampling points applied to a WDS with a total of N hydrants, the number of all possible combinations is C_N^M. This value increases exponentially with n and N becoming larger, leading to a rapid increase in computing time and deterioration of detection effectiveness.

To solve the computational issue, the Monte Carlo (MC) method is used in this study as an alternative to the enumeration approach in the process of determining the optimal sampling group to improve detection efficiency for large-scale WDSs. The selection of the MC method is mainly due to its simplicity and reasonable performance in offering near-optimal solutions (Maier et al., 2014). This is practically meaningful as in many engineering cases providing near-optimal solutions within a given time framework are more important than identifying global optimums with large computational overheads (Maier et al., 2014). Nevertheless, an advanced optimization algorithm can be developed for the proposed MGSM in future, which is not the focus of the present paper.

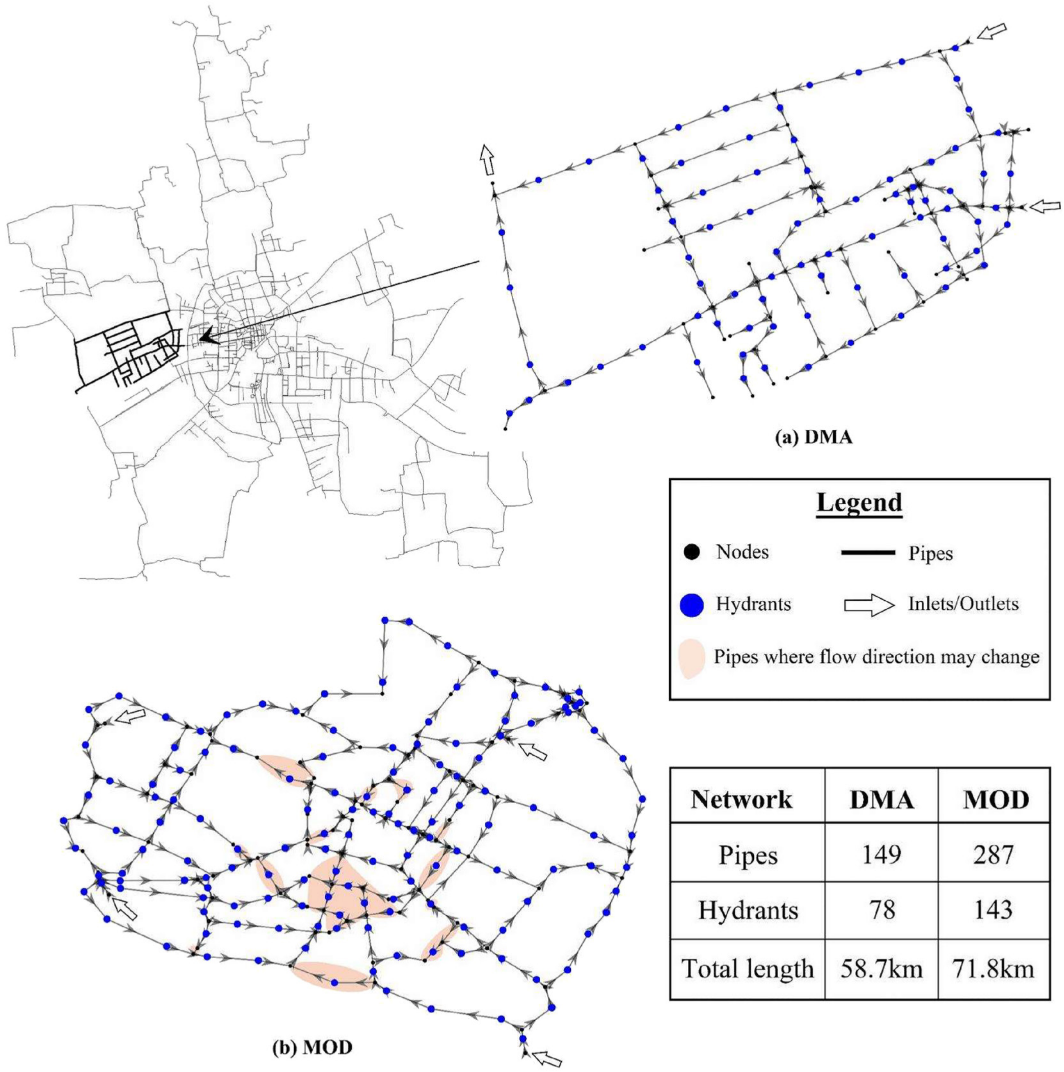
3. Case Studies

Two distribution networks (Figure 5) are used to demonstrate the utility of the proposed MGSM. Specifically, the DMA (district meter area) case study is a part of a real-world WDS in China (Figure 5a) that consists of 149 pipes (58.7 km in length) and 78 fire hydrants. It has two inlets and one outlet, and the flow direction in this network (shown in Figure 5a) does not change. The MOD pipe network is a benchmark WDS of the city of Modena in Italy (Bragalli et al., 2012). This network consists of 4 reservoirs (sources), 287 pipes (71.8 km in length), and 143 fire hydrants. Due to the water level changes in the four reservoirs and variations in residential water consumption, the flow directions of some pipes (shaded pipes in Figure 5b) in the MOD network change over time.

While the demonstration of the proposed MGSM using a very large WDS is academically necessary, in practice, the MGSM is intended for use on a DMA or a region of the entire WDS. This is because (a) many WDSs have been managed into regions, zones, or DMAs, which can greatly enhance the operation efficiency, and (b) for the WDSs with no DMAs, water quality testing or contamination source identification is likely to be conducted region by region. It is highly unlikely to simultaneously consider all the pipes of the entire large network as the contamination sources. Therefore, we demonstrate the proposed method using two case studies at a DMA scale level.

For both case studies, we have analyzed a series of different combinations of sampling locations (i.e., the number of hydrants that can be simultaneously sampled) at each cycle, with n ranging from 2 to 10. The number of potential contamination sources varies from one to three for these two WDSs. The size of the MC method is determined to be 10,000 based on a preliminary analysis for both case studies, but a larger value may be required for larger WDSs. The proposed MGSM is coded in C++ computing language with the aid of EPANET2.0 as the hydraulic solver to identify pipe flow directions (He et al., 2018). For the DMA case study with 78 hydrants and 2 contamination sources, the proposed method was tested using 2 and 10 potential sampling locations at each cycle required an average of 102 and 54 s, respectively, on a PC with Intel i5-9400F CPU@2.90 GHz. For the MOD network with 143 hydrants and 2 contamination sources, the proposed MGSM with 2 and 10 sampling locations at each cycle needs an average of 212 and 92 s, respectively. This implies that the proposed method is very efficient to identify the optimal sampling locations based on the test results. To enable the statistically rigorous analysis, for the single contamination source, we considered all possible scenarios with one source assigned to each pipe of the network. For two and three contamination sources, a total of 100 different randomly generated scenarios are considered.

10.1029/2022WR032784



Network	DMA	MOD
Pipes	149	287
Hydrants	78	143
Total length	58.7km	71.8km

Figure 5. (a) The DMA case study and (b) the MOD case study, where arrows indicate flow directions.

4. Results and Discussion

The proposed MGSM is demonstrated using the effectiveness (Section 4.1), the efficiency (Section 4.2), and the cost (Section 4.3) as shown in Section 4. The effectiveness is measured by the length of finally identified pipes relative to the total pipe length of the entire WDS, and the efficiency is measured by the total number of sampling cycles to identify these pipes with contamination sources. The cost associated with the sampling process is measured by the total number of samples that need to be tested in the laboratory.

4.1. Effectiveness of the Proposed MGSM

Figure 6 illustrates the application procedures of the proposed MGSM in dealing with the DMA case study with two contamination sources (1 and 2 in Figure 6a) and two sampling locations at each cycle. Two different MGSM runs (MR1 and MR2) are performed for this scenario, where the second run assumed that the contamination source identified in the first run was eliminated. As shown in this figure, in the beginning, the entire DMA is considered as the candidate sub-network (**CS**, Figure 6a) assuming that the water sample test at the outlet of this DMA shows contamination. This is followed by the application of the MGSM, where six and four cycles were carried out to localize contamination sources 1 and 2, respectively. The final identified pipe lengths associated with contamination sources 1 and 2 are 741 and 762 m, which represent only 1.26% and 1.30% of the entire

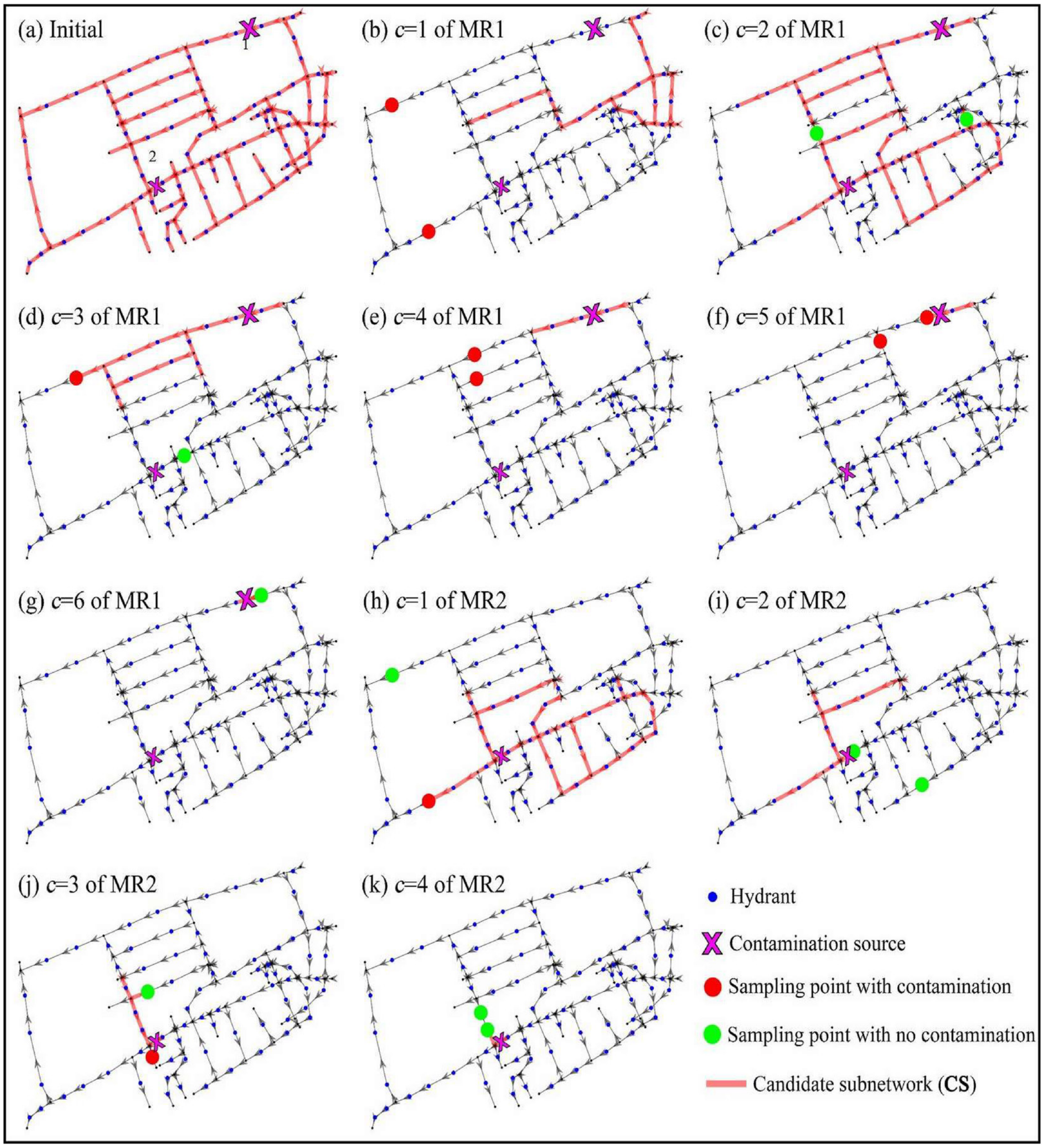


Figure 6. Source localization for the DMA case study with two contamination sources and two sampling locations at each cycle, where arrows indicate flow directions.

DMA, respectively. This implies that the proposed MGSM is able to effectively narrow down the spatial range of pipes that contain contamination sources, which can greatly facilitate the subsequent field investigations to eliminate the cause of the problem.

Figure 7 illustrates the proposed MGSM applied to the WDS with possible pipe flow changes. As shown in this figure, if the pipe flow directions do not change, the two sampling locations identified by the proposed MGSM are 1 and 2 (Figure 7b) based on the candidate sub-network (**CS**) determined at $c = 3$ (Figure 7a). However, if the flow directions change after the sample tests at $c = 3$, the CS for the next cycle needs to account for such variation. For the given example, one pipe is added to the **CS** due to its flow changing. This addition affects the optimal sampling locations selected by the MGSM (the location of 2 is changed as shown in Figure 7c). Based on this example, the flow direction changes can be easily handled by the proposed MGSM. For the MOD case study, we assume the change in the flow direction status occurs (Figure 7c) after $c = 3$, followed by a change to the original direction of flow after another two cycles.

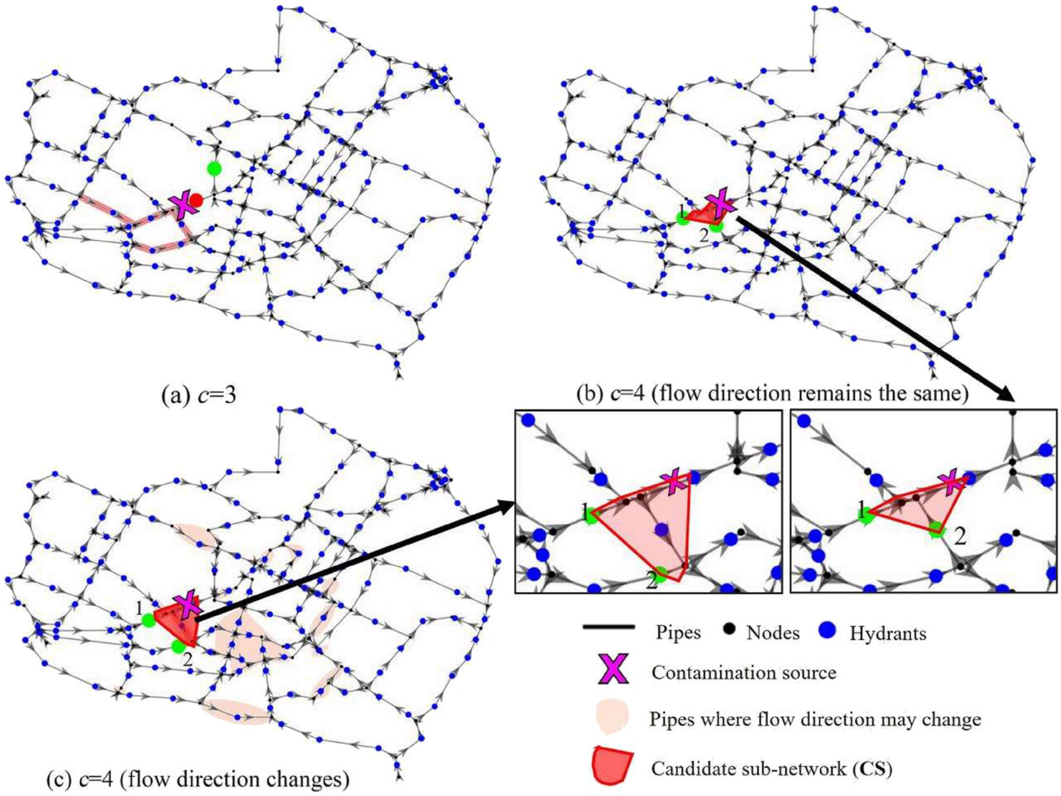


Figure 7. Source localization for the MOD case study with one contamination source and two sampling locations at each cycle, where arrows indicate flow directions.

It is found that the proposed MGSM is able to identify the contamination sources for all scenarios considered in both case studies, implying its great effectiveness to localize contamination sources. In this study, we define a detection effectiveness (%) metric as follows,

$$\text{Detection effectiveness} = \left(1 - \frac{L_f}{L_{all}}\right) \times 100\% \tag{6}$$

where L_f is the pipe length of the finally identified sub-network with contamination source(s) and L_{all} is the total pipe length of the entire WDS being considered. A high detection effectiveness represents that the proposed method can greatly reduce the efforts or budgets of the subsequent field investigations that are needed to micro-locate and eliminate contamination sources.

Figure 8 presents the probability density of the detection effectiveness (%) for all contamination scenarios considered. The probability density is estimated as the ratio between the length of the finally identified pipes with contamination and the total length of pipes in the WDS, across all contamination events. It is seen from this figure that the majority of the detection effectiveness (%) is higher than 95% and 98% for the DMA and MOD case studies, respectively. This indicates that the finally identified pipes with contamination source(s) represent a very small portion of the entire network, which can greatly improve the efficiency of the subsequent engineering effort to fix the contamination problem. The detection effectiveness (%) ranges between 80% and 90% for some contamination scenarios for the DMA case study, as shown in Figure 8a. This is due to the sparse distribution of hydrants for these events, and hence the length of the candidate sub-network identified by the proposed MGSM is relatively large. The detection effectiveness (%) decreases when dealing with a larger number of contamination sources that simultaneously exist in the WDS. It is noted that the detection effectiveness (%) values are the same as those obtained using the average pipe length distance between hydrants divided by the total pipe length of the network. This implies that the proposed method is able to identify the pipe with contamination source between the two hydrants for each scenario considered.

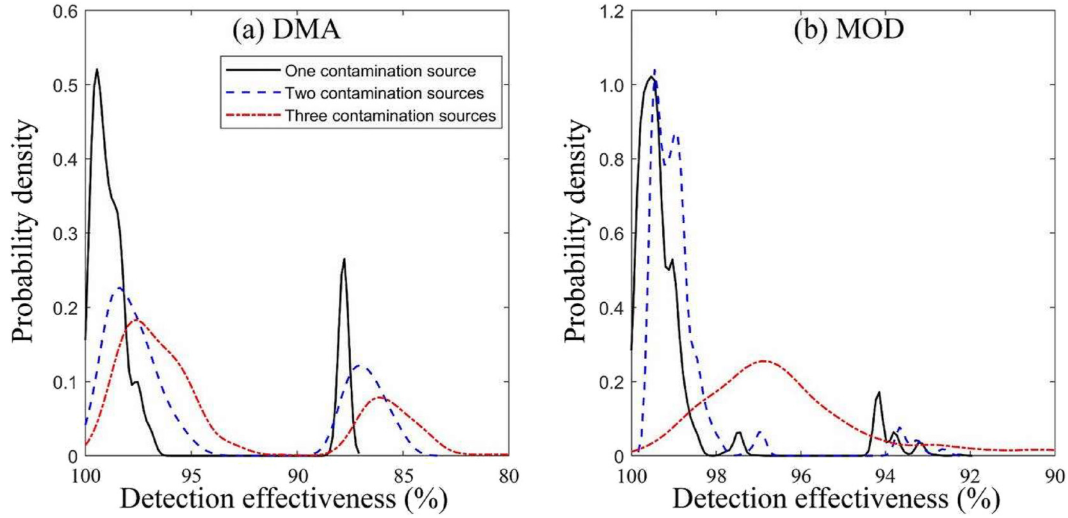


Figure 8. Detection effectiveness (%) of the proposed MGSM applied to the two case studies.

4.2. Detection Efficiency of the Proposed MGSM

The detection efficiency of the proposed MGSM can be evaluated using the number of total cycles required for the entire procedure. The total time used in each cycle includes the time required to collect and test samples, as well as the computation time needed to identify the sampling locations. As previously stated, both the computation and sample collection times are negligible compared to the laboratory tests. Figure 9 shows the total number of cycles used to localize contamination sources of the two case studies as a function of the varying number of samples per 100 km of pipe length at each cycle (n_k), where $n_k = n/L_{all} \times 100$. Such normalization is used to enable the generalization of the results to other WDSs.

As shown in Figure 9, an obvious trend that can be observed is that the detection efficiency is improved when n increases for all different contamination scenarios (n_k ranges from about 1.5 to 5) for both case studies. A

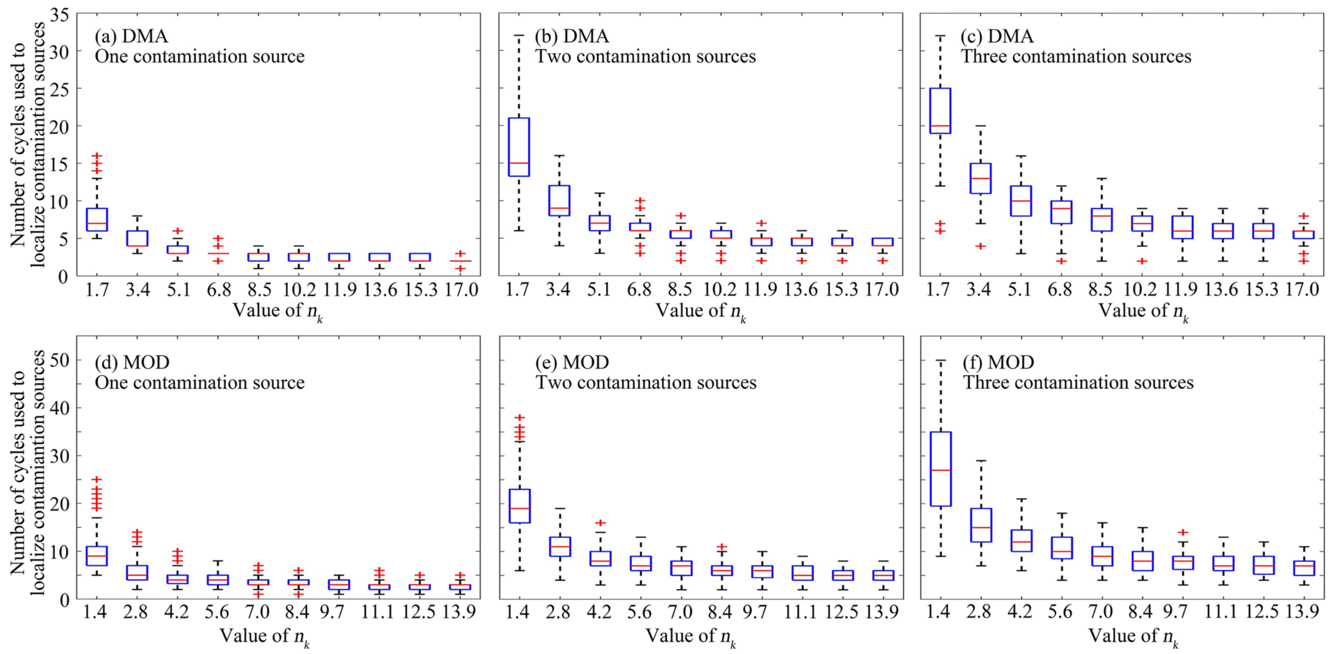


Figure 9. The number of cycles used to localize contamination sources versus the number of sampling points for every 100 km pipe length at each cycle (n_k) for the proposed MGSM applied to the two case studies.

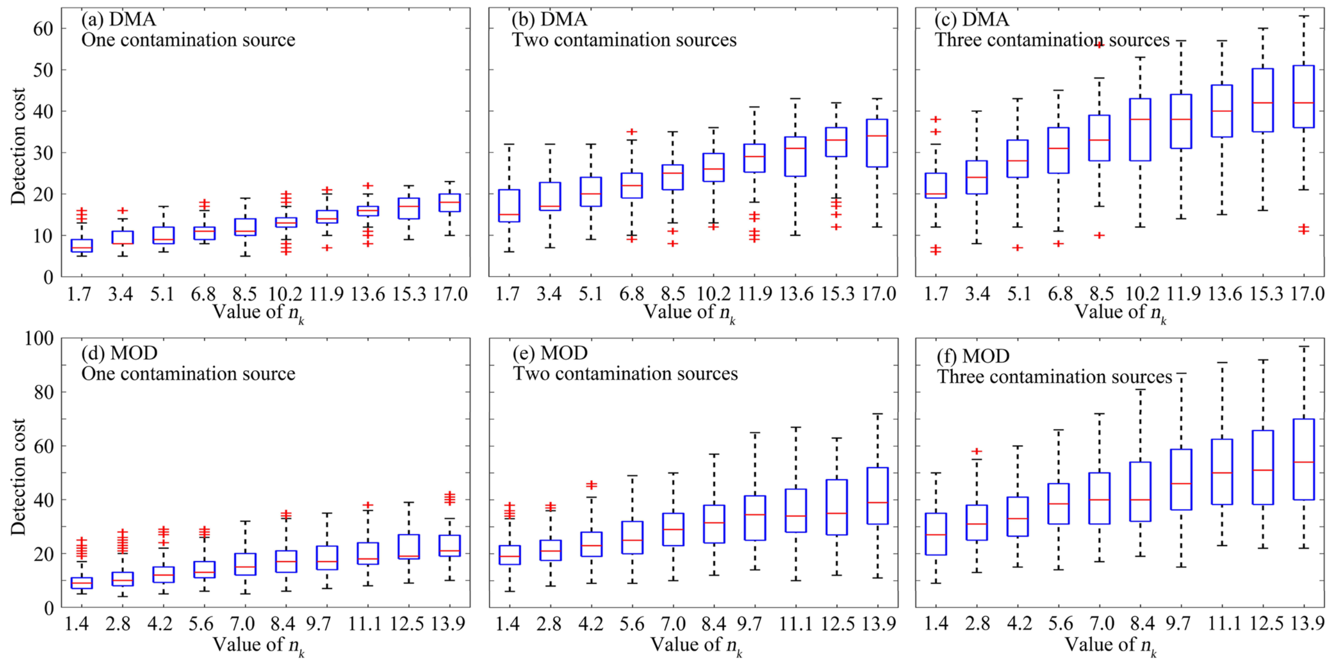


Figure 10. Detection cost (i.e., the number of total samples) versus the number of sampling points for every 100 km pipe length at each cycle (n_k) for the proposed MGSM applied to the two case studies.

significant increase in efficiency occurs for $n_k > 1.5$, with improvements diminishing when $n_k > 6$. This is expected as a high n_k value indicates a larger number of available teams for collecting samples and a significant laboratory capacity for simultaneously testing multiple samples. The diminishing efficiency improvement for large n_k implies that an optimal sampling size exists for the WDS when the efficiency is considered. For the DMA and MOD case studies, the optimal n_k value can be between 7.0 and 8.5 as a further increase in n_k value does not significantly improve the MGSM's detection efficiency, as shown in Figure 9. However, the optimal n_k value for detection efficiency can be case study dependent as it can be related to the size of the WDS being considered. In addition, a large n_k value corresponds to a significant financial commitment, and hence the decision process can be also affected by the budgets available.

Interestingly, for the same number of sampling locations at each 100 km pipe length n_k , when n_k is relatively low, the total number of cycles can vary significantly. For example, for the DMA case study if $n_k = 1.7$, the detection efficiency can vary from 5 to 15 cycles for one contamination source, and range from 7 to 25 cycles when three contamination sources are simultaneously considered. Similar observations can be made for the MOD case study. This implies that the location of the contamination sources can appreciably affect the detection efficiency when there is a low number of sampling teams available and/or a limited laboratory capacity for testing multiple samples. When a sufficiently large n_k is considered, the detection efficiency variations become small, as observed in Figure 9. This implies that the choice of n_k will also affect the uncertainty associated with method efficiency, which should be also accounted for in engineering practice.

4.3. Detection Cost of the Proposed MGSM

In this study, the detection cost of the proposed MGSM is measured by the total number of samples that have been tested to localize the contamination sources. Figure 10 shows the detection cost as a function of varying n_k for both case studies. Despite some variations, a large n_k value is generally associated with a greater detection cost for both case studies. In addition, the simultaneous presence of a larger number of contamination sources also causes an overall increase in detection costs. This information combined with the efficiency results in Figure 9 can be used as guidance for developing effective water quality sampling plans or budgets for a given WDS.

10.1029/2022WR032784

5. Summary and Conclusions

Existing research on water quality management and contamination source localization in WDSs has focused mainly on developing methods that assume the availability of accurate water quality models and multi-parameter online sensors. However, that is not true for many water utilities. A promising way to address such problems is through the iterative manual grab-sample strategies, thereby enabling effective contaminant localizing. To this end, this study proposes a new method for water quality manual grab-sampling (termed as MGSM in this paper) to enable the identification of contamination sources in WDSs.

The proposed MGSM is suitable for situations where online multi-parameter water quality sensors are sparsely available or completely missing, which is the case with many utilities. This is mainly due to the high purchase and maintenance cost associated with these sensors, as well as their inability (or inaccurate) to detect the complex water quality parameters (e.g., metals, microorganisms, and personal care products; Jia, Zheng, Maier, et al., 2021). In addition, a grab-sampling method is tailored for the cases when contamination is *continuously* presented in the WDS and with slow or low impacts to the WDSs. That is the case with misconnections between water supply pipes and sewer (or grey) pipes and contaminations caused by pipe leaks, corrosion, or hydraulic turbulence. For events with serious consequences, the candidate sub-networks (CSs) with contamination sources may need to be shut down or sampled manually as much as possible.

Based on the results obtained for two real-world cases, the following findings and conclusions can be drawn:

1. The newly proposed MGSM can successfully detect and locate continuous contamination source(s) for a wide range of scenarios, including multiple contamination source(s) in complex WDSs with varying pipe flow directions. This is a significant advantage over the traditional approach that works only with one contamination source and fixed flow directions, as described in Wong et al. (2010).
2. For the two case studies, the new MGSM identified contamination source(s) within 5% of the total pipe length of the WDS. This indicates the high effectiveness of the proposed MGSM in narrowing narrow down the spatial range of the sub-network with potential contamination sources. From the practical point of view, it also improves the efficiency of maintenance efforts to eliminate the sources of contamination.
3. The detection efficiency (measured by the number of sampling and testing cycles) of the MGSM can be significantly improved when the number of sampling points per 100 km pipe length at each cycle (n_k) increases from about 1.5 to a moderate value (e.g., $n_k \approx 7$). The increase in efficiency diminishes with further increases in n_k. This implies that there exists an optimal n_k value for a given WDS, representing the balanced trade-off between detection efficiency and costs associated with methodology. The detection cost grows with the increase in the number of sampling points per 100 pipe length, n_k. All these findings are important for the implementation of the method as they can guide the process of selecting the optimal number of sampling teams and required laboratory capacity.

In view of the practical application, the proposed MGSM can be used to regularly check water quality safety for WDSs with a low density of sensors as this is routine work in many water utilities. For instance, in China, many water utilities need to take water samples from hydrants or end users every month, with the number of samples depending on the scale of the WDS and importance level of the city. These water samples are comprehensively measured in the laboratory following the Water Quality Standard that has 106 parameters. Many water utilities collect grab samples from large WDSs at fixed locations based on specialists' engineering expertise. For example, a practitioner may collect grab samples from all established fixed locations (if say, 50 locations) and test for a combination (or all) of the specified water quality parameters in the laboratory. Such a strategy is time-consuming and expensive (labor and measurement costs). Therefore, the sampling strategy can be improved with the aid of the proposed MGSM in order to save the cost. It can be concluded that the MGSM is an alternative to the sensor-based detection methods.

The limitation of the proposed method is the potentially high cost and time required to identify the source(s) as all grab samples need to be collected manually (with technicians moving between different locations during multiple cycles) and processed in the laboratory. In addition, the pipes identified as the potential contamination sources need to be visited in the field to micro-locate the contamination source(s) via manual inspection or detection robots (Huang et al., 2020). This too requires time and has a cost associated with it. This, however, applies to most of the existing sensor-based methods as well. Another limitation is that the proposed MGSM can be only applicable to contamination events with continuous injections to the WDS conditioned on known pipe flow directions. Furthermore, when dealing with scenarios with pipe flow changes, there is likely that it would affect the utility of the proposed MGSM, which needs attention during practical implementation. While the application of the developed MGSM can be simple as it only requires flow direction information (Zhang et al., 2021), it should be also acknowledged the flow information can be challenging for some old undocumented areas due to system uncertainties.

Future studies along this research line should include (a) the application of the proposed method to large real WDSs; (b) the extension of the graph partitioning strategy within the proposed MGSM to account for both the pipe length and pipe velocity; and (c) the extension of the proposed MGSM to deal with contamination events with intermittent injections to the WDS.

Data Availability Statement

The data will eventually be deposited in the general repository Zenodo by the time the article is accepted, and the data are now available as Supporting Information for review purpose.

Acknowledgments
The corresponding author Professor Feifei Zheng was funded by the National Natural Science Foundation of China (Grant 51922096), and the Excellent Youth Natural Science Foundation of Zhejiang Province, China (LR19E080003). Dr. Weiwei Bi was funded by the National Natural Science Foundation of China (51808497). Dr. Huan-Feng Duan would like to appreciate the support from the Hong Kong Research Grants Council (RGC) (Project no. 15200719). Prof. Dragan Savic has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant 951424).

References

Asheri Arnon, T., Ezra, S., & Fishbain, B. (2019). Water characterization and early contamination detection in highly varying stochastic background water, based on Machine Learning methodology for processing real-time UV-Spectrophotometry. *Water Research*, *155*, 333–342. https://doi.org/10.1016/j.watres.2019.02.027

Bhatia, R., & Davis, C. (1995). A Cauchy-Schwarz inequality for operators with applications. *Linear Algebra and its Applications*, *223–224*, 119–129. https://doi.org/10.1016/0024-3795(94)00344-d

Bragalli, C., D'Ambrosio, C., Lee, J., Lodi, A., & Toth, P. (2012). On the optimal design of water distribution networks: Practical MINLP approach. *Optimization and Engineering*, *13*(2), 219–246. https://doi.org/10.1007/s11081-011-9141-7

Butera, I., Gómez-Hernández, J. J., & Nicotra, S. (2021). Contaminant-source detection in a water distribution system using the ensemble Kalman filter. *Journal of Water Resources Planning and Management*, *147*(7), 04021029. https://doi.org/10.1061/(asce)wr.1943-5452.0001383

ChinaNews. (2020). Retrieved from http://www.chinanews.com/sh/2020/07-30/9252169.shtml

Di Nardo, A., Giudicianni, C., Greco, R., Herrera, M., Santonastaso, G. F., & Scala, A. (2018). Sensor placement in water distribution networks based on spectral algorithms. *Paper presented at the 13th International Conference on Hydroinformatics (HIC2018)* (p. 7).

Giudicianni, C., Herrera, M., Nardo, A. D., Greco, R., Creaco, E., & Scala, A. (2020). Topological placement of quality sensors in water-distribution networks without the recourse to hydraulic modeling. *Journal of Water Resources Planning and Management*, *146*(6), 04020030. https://doi.org/10.1061/(asce)wr.1943-5452.0001210

Grbčić, L., Lučin, I., Kranjčević, L., & Družeta, S. (2020). Water supply network pollution source identification by random forest algorithm. *Journal of Hydroinformatics*, *22*(6), 1521–1535.

Hart, D., Rodriguez, J. S., Burkhardt, J., Borchers, B., Laird, C., & Murray, R. (2019). Quantifying hydraulic and water quality uncertainty to inform sampling of drinking water distribution systems. *Journal of Water Resources Planning and Management*, *145*(1), 04018084. https://doi.org/10.1061/(asce)wr.1943-5452.0001005

Hart, W. E., & Murray, R. (2010). Review of sensor placement strategies for contamination warning systems in drinking water distribution systems. *Journal of Water Resources Planning and Management*, *136*(6), 611–619. https://doi.org/10.1061/(asce)wr.1943-5452.0000081

He, G., Zhang, T., Zheng, F., Li, C., Zhang, Q., Dong, F., & Huang, Y. (2019). Reaction of with chlorine and chlorine dioxide in drinking water distribution systems: Kinetics, transformation mechanisms and toxicity evaluations. *Chemical Engineering Journal*, *374*, 1191–1203. https://doi.org/10.1016/j.cej.2019.06.022

He, G., Zhang, T., Zheng, F., & Zhang, Q. (2018). An efficient multi-objective optimization method for water quality sensor placement within water distribution systems considering contamination probability variations. *Water Research*, *143*, 165–175. https://doi.org/10.1016/j.watres.2018.06.041

Hu, C., Dai, L., Yan, X., Gong, W., Liu, X., & Wang, L. (2020). Modified NSGA-III for sensor placement in water distribution system. *Information Sciences*, *509*, 488–500. https://doi.org/10.1016/j.ins.2018.06.055

Hu, C., Ren, G., Liu, C., Li, M., & Jie, W. (2017). A Spark-based genetic algorithm for sensor placement in large scale drinking water distribution systems. *Cluster Computing*, *20*(2), 1089–1099. https://doi.org/10.1007/s10586-017-0838-z

Hu, C., Zhao, J., Yan, X., Zeng, D., & Guo, S. (2015). A MapReduce based Parallel Niche Genetic Algorithm for contaminant source identification in water distribution network. *Ad Hoc Networks*, *35*, 116–126. https://doi.org/10.1016/j.adhoc.2015.07.011

Huang, Y., Zheng, F., Kapelan, Z., Savic, D., Duan, H.-F., & Zhang, Q. (2020). Efficient leak localization in water distribution systems using multistage optimal valve operations and smart demand metering. *Water Resources Research*, *56*(10), e2020WR028285. https://doi.org/10.1029/2020wr028285

Jerez, D. J., Jensen, H. A., Beer, M., & Broggi, M. (2021). Contaminant source identification in water distribution networks: A Bayesian framework. *Mechanical Systems and Signal Processing*, *159*, 107834. https://doi.org/10.1016/j.ymssp.2021.107834

Jia, Y., Zheng, F., Maier, H. R., Ostfeld, A., Creaco, E., & Savic, D. (2021). Water quality modelling in sewer networks: Review and future research directions. *Water Research*, *202*, 117.

Jia, Y., Zheng, F., Zhang, Q., Duan, H.-F., Savic, D., & Kapelan, Z. (2021). Foul sewer model development using geotagged information and smart water meter data. *Water Research*, *204*, 117594. https://doi.org/10.1016/j.watres.2021.117594

Khorshidi, M. S., Nikoo, M. R., & Sadegh, M. (2018). Optimal and objective placement of sensors in water distribution systems using information theory. *Water Research*, *143*, 218–228. https://doi.org/10.1016/j.watres.2018.06.050

Li, C., Yang, R., Zhou, L., Zeng, S., Mavrovouniotis, M., & Yang, M. (2021). Adaptive multipopulation evolutionary algorithm for contamination source identification in water distribution systems. *Journal of Water Resources Planning and Management*, *147*(5), 04021014. https://doi.org/10.1061/(asce)wr.1943-5452.0001362

Maier, H. R., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L. S., & Cunha, M. C. (2014). Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions. *Environmental Modelling & Software*, *62*, 271–299. https://doi.org/10.1016/j.envsoft.2014.09.013

Mann, A. V., McKenna, S. A., Hart, W. E., & Laird, C. D. (2012). Real-time inversion in large-scale water networks using discrete measurements. *Computers & Chemical Engineering*, *37*, 143–151. https://doi.org/10.1016/j.compchemeng.2011.08.001

Naserizade, S. S., Nikoo, M. R., & Montaseri, H. (2018). A risk-based multi-objective model for optimal placement of sensors in water distribution system. *Journal of Hydrology*, *557*, 147–159. https://doi.org/10.1016/j.jhydrol.2017.12.028

Ohar, Z., Lahav, O., & Ostfeld, A. (2015). Optimal sensor placement for detecting organophosphate intrusions into water distribution systems. *Water Research*, *73*, 193–203. https://doi.org/10.1016/j.watres.2015.01.024

Ostfeld, A., Oliker, N., & Salomons, E. (2014). Multiobjective optimization for least cost design and resiliency of water distribution systems. *Environmental Modelling and Software*, *140*(12), 04014037. https://doi.org/10.1061/(asce)wr.1943-5452.0000407

Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., & Phillips, C. A. (2008). The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management*, *134*(6), 556–568. https://doi.org/10.1061/(asce)0733-9496(2008)134:6(556)

Preis, A., & Ostfeld, A. (2006). Contamination source identification in water systems: A hybrid model trees–linear programming scheme. *Journal of Water Resources Planning and Management*, *132*(4), 263–273. https://doi.org/10.1061/(asce)0733-9496(2006)132:4(263)

Preis, A., & Ostfeld, A. (2007). A contamination source identification model for water distribution system security. *Engineering Optimization*, *39*(8), 941–947. https://doi.org/10.1080/03052150701540670

Preis, A., & Ostfeld, A. (2008). Genetic algorithm for contaminant source characterization using imperfect sensors. *Engineering and Environmental Systems*, *25*(1), 29–39. https://doi.org/10.1080/10286600701695471

Qi, Z., Zheng, F., Guo, D., Maier, H. R., Zhang, T., Yu, T., & Shao, Y. (2018). Better understanding of the capacity of pressure sensor systems to detect pipe burst within water distribution networks. *Journal of Water Resources Planning and Management*, *144*(7), 04018035. https://doi.org/10.1061/(asce)wr.1943-5452.0000957

Rathi, S., & Gupta, R. (2014). Sensor placement methods for contamination detection in water distribution networks: A review. *Procedia Engineering*, *89*, 181–188. https://doi.org/10.1016/j.proeng.2014.11.175

Robertson, L., Gjerde, B., Hansen, E. F., & Stachurska-Hagen, T. (2008). A water contamination incident in Oslo, Norway during October 2007; A basis for discussion of boil-water notices and the potential for post-treatment contamination of drinking water supplies. *Journal of Water and Health*, *7*(1), 55–66. https://doi.org/10.2166/wh.2009.014

Rodriguez, J. S., Bynum, M., Laird, C., Hart, D. B., Klise, K. A., Burkhardt, J., & Haxton, T., (2021). Optimal sampling locations to reduce uncertainty in contamination extent in water distribution systems. *Journal of Infrastructure Systems*, *27*(3), 04021026. https://doi.org/10.1061/(asce)is.1943-555x.0000628

Sankary, N., & Ostfeld, A. (2018). Multiobjective optimization of inline mobile and fixed wireless sensor networks under conditions of demand uncertainty. *Journal of Water Resources Planning and Management*, *144*(8), 04018043. https://doi.org/10.1061/(asce)wr.1943-5452.0000930

Sankary, N., & Ostfeld, A. (2019). Bayesian localization of water distribution system contamination intrusion events using inline mobile sensor data. *Journal of Water Resources Planning and Management*, *145*(8), 04019029. https://doi.org/10.1061/(asce)wr.1943-5452.0001086

Sun, L., Yan, H., Xin, K., & Tao, T. (2019). Contamination source identification in water distribution networks using convolutional neural network. *Environmental Science and Pollution Research*, *26*(36), 36786–36797. https://doi.org/10.1007/s11356-019-06755-x

Ung, H., Piller, O., Gilbert, D., & Mortazavi, I. (2017). Accurate and optimal sensor placement for source identification of water distribution networks. *Journal of Water Resources Planning and Management*, *143*(8), 04017032. https://doi.org/10.1061/(asce)wr.1943-5452.0000777

Vrachimis, S. G., Lifshitz, R., Eliades, D. G., Polycarpou, M. M., & Ostfeld, A. (2020). Active contamination detection in water-distribution systems. *Journal of Water Resources Planning and Management*, *146*(4), 04020014. https://doi.org/10.1061/(asce)wr.1943-5452.0001176

Winter, C. D., Palleti, V. R., Worm, D., & Kooij, R. (2019). Optimal placement of imperfect water quality sensors in water distribution networks. *Computers & Chemical Engineering*, *121*, 200–211. https://doi.org/10.1016/j.compchemeng.2018.10.021

Wong, A., Young, J., & Laird, C. D. (2010). Optimal determination of grab sample location and source inversion in large-scale water distribution systems. *Water Distribution Systems Analysis*, *2010*, 412–425.

Yang, X., & Boccelli, D. L. (2014). Bayesian approach for real-time probabilistic contamination source identification. *Journal of Water Resources Planning and Management*, *140*(8), 04014019. https://doi.org/10.1061/(asce)wr.1943-5452.0000381

Yang, X., & Boccelli, D. L. (2016). Model-based event detection for contaminant warning systems. *Journal of Water Resources Planning and Management*, *142*(11), 04016048. https://doi.org/10.1061/(asce)wr.1943-5452.0000689

Zhang, Q., Zheng, F., Jia, Y., Savic, D., & Kapelan, Z. (2021). Real-time foul sewer hydraulic modelling driven by water consumption data from water distribution systems. *Water Research*, *188*, 116544. https://doi.org/10.1016/j.watres.2020.116544

Zhang, Q., Zheng, F., Kapelan, Z., Savic, D., He, G., & Ma, Y. (2020). Assessing the global resilience of water quality sensor placement strategies within water distribution systems. *Water Research*, *172*, 115527. https://doi.org/10.1016/j.watres.2020.115527

Zheng, F., Du, J., Diao, K., Zhang, T., Yu, T., & Shao, Y. (2018). Investigating effectiveness of sensor placement strategies in contamination detection within water distribution systems. *Journal of Water Resources Planning and Management*, *144*(4), 06018003. https://doi.org/10.1061/(asce)wr.1943-5452.0000919