Computers and Chemical Engineering 153 (2021) 107404

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Optimal sensor placement for contamination identification in water distribution system considering contamination probability variations

Zukang Hu[a], Wenlong Chen[b,c], Dingtao Shen[b,c], Beiqing Chen[b,c], Song Ye[b,c], Debao Tan[a,b,*]

[a] College of Computer and Information, Hohai University, Nanjing 210098, China
[b] Hubei Provincial Key Laboratory of River Basin Water Resources and Eco-environmental Sciences, Changjiang River Scientific Research Institute, Wuhan 430310, China
[c] Spatial Information Technology Application Department, Changjiang River Scientific Research Institute, Wuhan 430310, China

ARTICLE INFO

Article history:
Received 25 February 2021
Revised 4 May 2021
Accepted 4 June 2021
Available online 22 June 2021

Keywords:
Contaminant intrusion
Optimal sensor placement
Identifiability
Contamination probability

ABSTRACT

Optimal sensor placement is of great importance for water distribution networks (WDNs). Currently available studies have not considered variations in node contamination probabilities on the WDN. Therefore, this study proposes an optimal sensor placement method for WDNs that considers the variations in node contamination probabilities and the impact of contamination on the network. First, contamination events were selected and clustered according to risk assessment results. Then, based on a hierarchical algorithm, a set of initial sensor placement schemes that meet the identifiability objective of the sensor network was obtained. Subsequently, all placement schemes were compared based on the numbers of sensors and the impact caused by sensor interruptions to obtain an optimal scheme. Tests were performed based on the WDN example model *k1*. The test results showed that considerations of the variations in node contamination probabilities can ensure the detection and identification of contamination events with high contamination probabilities.

1. Introduction

The water distribution network (WDN), which distributes drinking water to residents through pipes, is an essential component of a city's infrastructure (Palleti et al., 2016). In WDNs, however, there are risks of contaminant intrusions that may directly threaten the lives of users (Cardoso et al., 2021). To alleviate the harm caused by such contaminant intrusions, the construction of a safe and efficient monitoring system is essential. This requires the placement of detection sensors in WDNs. Owing to the high installation and maintenance costs, it is impossible to place sensors at every WDN node. Therefore, it is necessary to optimize the location of sensors and minimize their numbers.

The optimal sensor placement for monitoring contaminant intrusions in WDNs has received wide attention; the sensor network's observability and identifiability are two particularly important reference indicators that have been studied intensely (Palleti et al., 2016). Observability refers to the sensor network's ability to detect contaminant intrusions—that is, at least one sensor should respond to a contaminant intrusion (Hooshmand et al., 2020). Researchers have proposed various methods to optimize the observability of sensor networks. Some optimization objectives—such as detection time (He et al., 2018; Cardoso et al., 2021), demand coverage (Rathi et al., 2016), or the impact of contamination (Ciaponi et al., 2019) have been widely adopted.

Identifiability refers to the sensor network's ability to identify the locations of contaminant intrusions—that is, when contaminant intrusions occur, sensors not only respond but also identify the corresponding locations (Hooshmand et al., 2020). Researchers have proposed several methods to achieve sensor network identifiability, such as establishing a bipartite graph between the contaminant intrusion locations and the nodes that might be affected by contaminations, and then converting it into a minimum set cover problem on the bipartite graph (Palleti et al., 2016; Winter et al., 2019). Based on the identifiability criteria, the number of contaminant intrusion nodes with the same warning mode can be minimized (Hooshmand et al., 2020). Only by identifying the contaminant intrusion nodes accurately, can corresponding measures then be taken to promptly restore the WDNs.

Based on the method proposed by Palleti et al. (2016), Winter et al. (2019) considered cases of sensor interruptions; their results showed that such interruptions could affect the selection of optimal sensor locations. Moreover, owing to the daily maintenance of sensors—for example, battery replacement or calibration—and

* Corresponding author.
E-mail address: tdebao@126.com (D. Tan).

https://doi.org/10.1016/j.compchemeng.2021.107404
0098-1354/

sensor communication interruptions, data monitoring will be interrupted (Sela and Amin, 2018). To ensure better stability of sensor placement schemes in practical applications, researchers have adopted various methods, such as quantifying the defects or failures of a single sensor (Comboul and Ghanem,2013; Jung and Kim, 2018), assuming each sensor fails at a given probability (Winter et al., 2019), simultaneously detecting each contamination event with more than one sensor (Preis and Ostfeld, 2008), considering the scenarios of sensor interruptions (Sela and Amin, 2018), and minimizing the loss of information entropy during sensor interruptions (Bertola et al., 2019).

Although the optimal sensor placement for identifying contaminant intrusion locations has been extensively studied, most of these studies assumed that the contamination probability at each contaminant intrusion location is equal; they did not consider variations in contamination probabilities. However, owing to variations in the demands at each node, the lengths of pipes directly connected to each node, the flow rates in pipes, and user attributes, the contamination probability of each node may be quite different (He et al., 2018). In cases of intentional contamination in WDNs, consideration of the variations of node contamination probabilities is conducive to the establishment of an effective monitoring system (Cardoso et al., 2021). Some studies have considered contamination probability variations, such as the contamination probability based on distances between public nodes and special nodes (Cardoso et al., 2021), user attributes (He et al., 2018), and the risk levels of contamination events (Rathi et al., 2016). Moreover, owing to the limited number of sensors, the selection of contamination events can reduce the calculation overhead and obtain more practical sensor placement schemes (Rathi et al., 2016). Some studies have selected contamination events when optimizing sensor placement. For example, if contaminant intrusions only occur at non-terminal nodes (Shen and Mcbean, 2011) or non-zero–demand nodes (Berry et al., 2006, 2009), only contamination events that affect more than 10% (Krause et al., 2008) of the nodes in the WDN or result in serious consequences need be considered (Carr et al., 2006; Weickgenannt et al., 2010; Perelman and Ostfeld, 2010). Although these studies have considered variations of contamination probabilities and selected contamination events, they have not been able to identify the locations of contaminant intrusions. Currently available studies on optimal sensor placement for identifying contaminant intrusion locations have also not considered variations of the probabilities of contamination events.

For the aforementioned reasons, this study proposes an optimal sensor placement method for identifying contaminant intrusion locations to identify contamination events that may have serious consequences, by considering variations in the contamination probabilities of contaminant intrusion nodes as well as the impact of contamination. Contamination events were selected and clustered; then, a risk assessment was performed. Based on a hierarchical algorithm, a set of sensor placement schemes that satisfied the identifiability criteria of sensor networks were obtained. Subsequently, based on the numbers of sensors and the impact of sensor interruptions on the sensor network, the various schemes were compared to obtain the optimal one; variations on contamination probabilities were also considered during the comparison.

2. Methods

2.1. Selection and clustering of contamination events

The detection and identification of all contamination events is a challenging task that is almost impossible to perform completely (Cardoso et al., 2021). At the same time, owing to the limited number of sensors, considerable attention is often paid to contamination events with serious consequences. Therefore, contamination events are selected based on risk assessment results in the current study. The purpose of a risk assessment or hazard vulnerability analysis is to determine the risks faced by WDNs; the various types of events are sorted or selected based on the risk assessment results (Berglund et al., 2020).

Suppose that contaminant intrusions occur at the WDN nodes, the number of which is N. Based on an EPANET simulation, the contamination detection matrix $L(i, j)$ of contaminant intrusions occurred at each node can be expressed by Eq. (1):

$$L(i,j)=\begin{pmatrix} l_{11} & \cdots & l_{1N} \\ \vdots & \ddots & \vdots \\ l_{N1} & \cdots & l_{NN} \end{pmatrix} \tag{1}$$

where l_{ij} is the impact of contaminant intrusion at node j on node i. If there is an impact (the contaminant concentration at node i is greater than 0), then $l_{ij}=1$; otherwise, $l_{ij}=0$.

After acquiring the contamination detection matrix, the impact caused by contamination at each node is assessed to obtain the corresponding contamination risk level. The contamination risk level of node j is

$$R(j)=\frac{\sum_{i=1}^{N}\sum_{t=\Delta t}^{A\ \Delta t} l_{ij}\cdot D_i(t)}{\sum_{i=1}^{N}\sum_{t=\Delta t}^{A\ \Delta t} D_i(t)} \tag{2}$$

where $R(j)$ is the risk level of the impact caused by contaminant intrusion at node j, $j=1,\cdots,\ N$ (N is the number of nodes in the WDN), $D_i(t)$ is the water demand of node i at time $t=\Delta t, 2\Delta t,\ \cdots,\ A\Delta t$, A is the total simulation period, and $\sum_{i=1}^{N}\sum_{t=\Delta t}^{A\ \Delta t} D_i(t)$ is the total water demand of all the WDN nodes during the simulation period.

The contamination events are then selected based on the contamination risk levels; contamination events with relatively low risk levels (no more than 2 nodes affected after contamination) are ignored. Moreover, there are some contamination events for which measures are taken that affect a number of WDN nodes (such as cutting off the water supply to a certain area of the WDN). Therefore, various contamination events are clustered based on their impact nodes, and the contamination events in one cluster are regarded as one type of event. The K-means clustering method is used to divide all contamination events into C clusters. First, C cluster centers are placed randomly. Then, the Euclidean distance between each contamination event and each cluster center is calculated (Steinley, 2006):

$$d\left(x_i,x_j\right)=\left(\left|x_{i1}-x_{j1}\right|^2+\cdots\left|x_{in}-x_{jn}\right|^2\cdots+\left|x_{iN}-x_{jN}\right|^2\right)^{1/2} \tag{3}$$

where $d(x_i,x_j)$ represents the Euclidean distance between the contamination event i and the cluster center j, x_{in} represents the impact of the contaminant intrusion at node i on node n, and N is the number of WDN nodes.

If the contamination event i and the cluster center j have the shortest Euclidean distance $d(x_i,x_j)$, the contamination event i is divided into cluster j. Subsequently, the average location of contamination events in the cluster is taken to be the new cluster center, and a new cluster is obtained via iteration. The aforementioned process is repeated until the center location no longer changes (Cheikh et al., 2010). The silhouette coefficient is used to determine the ideal number of clusters C. The silhouette coefficient of each data is a measure of its similarity to the data in the cluster. For the ith contamination event, the silhouette coefficient $S(k)$ is defined as follows (Kaufman and Rousseeuw, 2009):

$$S(k)=\frac{(b_i-a_i)}{max(a_i,b_i)} \tag{4}$$

where a_i is the ith contamination event to the other contamination event in the same cluster, b_i is the minimum average distance from

the ith contamination event to the other contamination event in the different cluster.

For a single cluster, if it contains more nodes and the contamination risks caused by these nodes are higher, the contamination risk of the cluster will be higher. The contamination risk level of a single cluster is

$$R(c) = \frac{\sum_{j=1}^{n_c} \sum_{i=1}^{N} \sum_{t=\Delta t}^{A\ \Delta t} l_{ij} \cdot D_i(t)}{\sum_{i=1}^{N} \sum_{t=\Delta t}^{A\ \Delta t} D_i(t)} \tag{5}$$

where $R(c)$ is the risk level of contaminant intrusion for cluster c, $D_i(t)$ is the water demand of node i at time t, and $i = 1, \cdots, N; j = 1, \cdots, n_c$ (n_c is the number of nodes in cluster c).

2.2. Contamination probabilities

In this study, four types of contamination probabilities are considered: (1) based on node demand; (2) based on pipe length; (3) based on risk level; (4) based on quantified risk.

2.2.1. Based on node demand

When defining the contamination probability of a node, the node demand is naturally considered to be a direct indicator (He et al., 2018). This is because nodes with high water demands are usually related to densely populated areas, which are more vulnerable to malicious attacks with potentially more serious consequences. Therefore, the contamination probabilities of nodes with high water demands are generally higher than those with low water demands. The contamination probability of each cluster is

$$P_D(c) = \frac{\sum_{i=1}^{n_c} \sum_{t=\Delta t}^{A\ \Delta t} D_i(t)}{\sum_{i=1}^{N} \sum_{t=\Delta t}^{A\ \Delta t} D_i(t)} \tag{6}$$

where $P_D(c)$ is the contamination probability of cluster c obtained based on node demand, $i = 1, 2, \cdots, n_c$, N is the total number of WDN nodes; $D_i(t)$ is the water demand of node i at time t, and $\sum_{i=1}^{N} \sum_{t=\Delta t}^{A\ \Delta t} D_i(t)$ is the total water demand of all the WDN nodes during the entire period.

2.2.2. Based on pipe length

In addition to node demand, the total length of pipes directly connected to a node can also reflect its contamination probability. When contamination events occur at nodes with relatively long pipe lengths, the affected space/area may be larger than at nodes with short pipe lengths, thereby resulting in greater social impact. Moreover, contaminant intrusions often occur at pipe-rupture points. The longer the pipe length—the higher the probability of pipe rupture—the higher the contamination probability of the pipe. The contamination probability of each cluster is

$$p_L(c) = \frac{\sum_{i=1}^{n_c} \frac{1}{2} \sum_{k=1}^{\Omega(i)} L_k}{\sum_{j=1}^{M} L_j} \tag{7}$$

where $p_L(c)$ is the contamination probability of cluster c obtained based on the length of pipes directly connected to the node, $i = 1, 2, \cdots, n_c$ (n_c is the number of nodes in cluster c), L_j is the pipe length, $j = 1, 2, \cdots, M$ (M is the total number of pipes in the WDN); and $\Omega(i)$ is the set of pipes for node i, including all the pipes that are directly connected to node i.

2.2.3. Based on risk level

Based on the contamination risk level of each node $R(j)$, all nodes are divided into three risk levels: high, medium, and low; and each risk level has an equal number of nodes. The higher the risk level of a node, the higher the occurrence probability of its contamination. The contamination probability $p_r(i)$ of the node at a high, medium, and low risk level is [0.7–0.9], [0.5–0.6], and [0.3–0.4], respectively. The contamination probability of each cluster is

$$p_R(c) = \frac{\sum_{i=1}^{n_c} p_r(i)}{n_c} \tag{8}$$

where $p_r(c)$ is the contamination probability of cluster c obtained based on the risk level, and $p_r(i)$ is the contamination probability of node i. As shown in Eq. (8), the contamination probability of cluster c is defined as the average contamination probability of all the nodes in cluster c.

2.2.4. Based on quantified risk

When the nodes' contamination probabilities are differentiated based on their contamination risk levels, the contamination probabilities of nodes at the same risk level are the same. Quantifying the contamination probability of a node directly based on its risk $R(j)$ can distinguish the contamination probabilities of nodes at the same risk level. The contamination probability of each cluster is

$$p_Q(c) = \frac{\sum_{c=1}^{n_c} \sum_{i=1}^{N} \sum_{t=\Delta t}^{A\ \Delta t} l_{ij} \cdot D_j(t)}{\sum_{i=1}^{N} \sum_{t=\Delta t}^{A\ \Delta t} D_i(t)} \tag{9}$$

where $p_Q(c)$ is the contamination occurrence probability of cluster c obtained based on the quantified risks of nodes. As shown in Eq. (9), the contamination probability of cluster c is defined as the ratio of the water demand affected by the contamination of cluster c to the water demand of the entire WDN.

2.2.5. Combined contamination probability

In order to better compare with the same pollution probability, we combine the four proposed pollution probabilities and get the combined pollution probability. The contamination probability of each cluster is

$$P_C(c) = \omega_D \cdot P_D(c) + \omega_L \cdot P_L(c) + \omega_R \cdot P_R(c) + \omega_Q \cdot P_Q(c) \tag{10}$$

where $P_C(c)$ is the combined contamination probability of cluster c obtained based on the four contamination probabilities. ω_D, ω_L, ω_R and ω_Q are weight coefficients, $\omega_D + \omega_L + \omega_R + \omega_Q = 1$.

2.3. Initial sensor placement schemes

In a WDN, a single sensor usually can only respond to the contamination events of some intrusion nodes. For sensor S_1 placed at a given WDN node, all the contamination events are divided into two intervals: [1] and [0], where "1″ indicates that the event can be detected, and "0″ indicates that the event cannot be detected. Combined with sensor S_2 at another node, all the contamination events are divided into four intervals: [1, 1], [1, 0], [0, 1], and [0, 0]. As the number of sensors continues to increase, the intervals of contamination events are continuously divided until all events are distributed in a single interval. To meet the detectability of the sensor network for contamination events, there should be no interval comprising only zeros.

Therefore, when selecting sensor locations, it is not only necessary to consider the detection of various contamination events by a single sensor, but also the combinations of various sensors. In the field of parameter identification, researchers have proposed a hierarchical algorithm based on joint information entropy for optimal sensor placement (Bertola et al., 2017). This method selects the sensor locations based on a greedy strategy, which maximizes the information gain of the obtained optimal sensor placement scheme, thereby ensuring the optimal identification ability of the sensor network. The joint information entropy of the initial locations of all sensors is (Bertola et al., 2019):

$$H(y_i) = -\sum_{j=1}^{N_{l,i}} P\left(y_{i,j}\right) \log_{10} P\left(y_{i,j}\right) \tag{11}$$

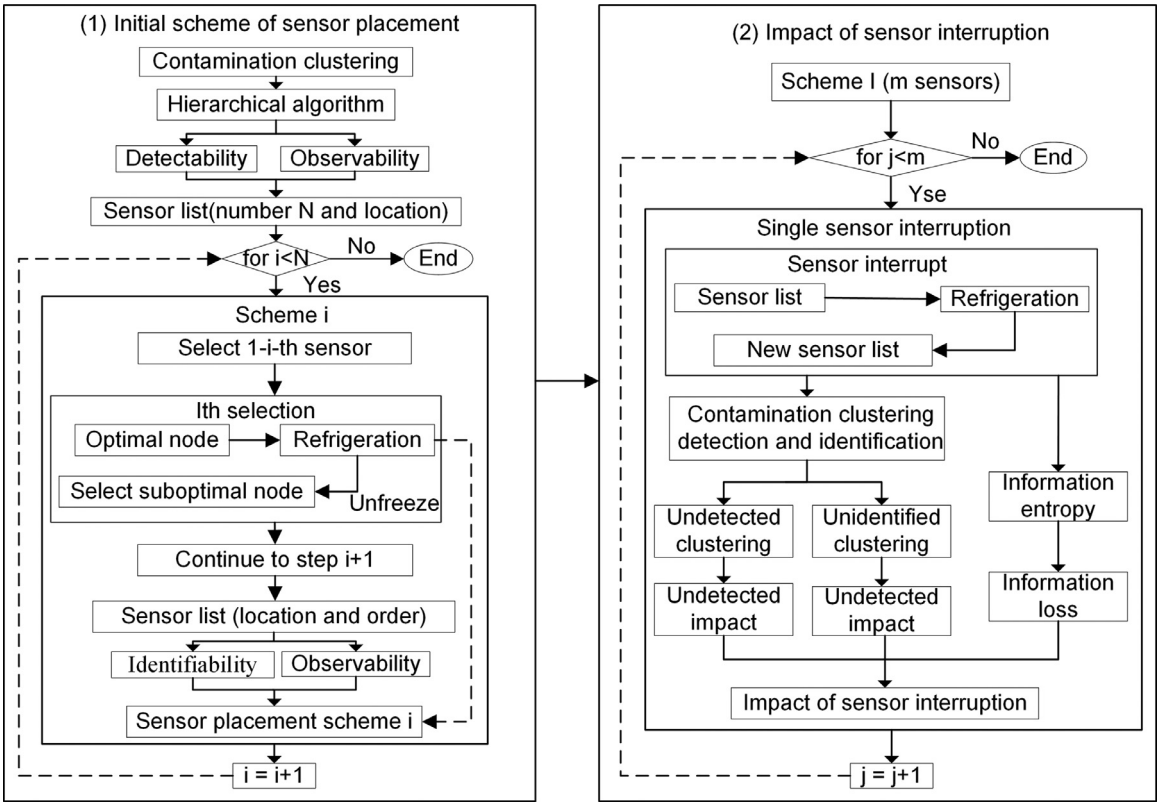

Fig. 1. Initial selection of sensor placement schemes and the impact of sensor interruptions.

where $H(y_i)$ is the sensor's information entropy at node i, $j = 1, \cdots, N_{I,i}$, $N_{I,i}$ is the number of intervals of contamination events divided by sensor S_i at node i, $P(y_{i,j}) = m_i/N_{I,i}$ is the probability of a contamination event being distributed in the jth interval, and m_i is the number of contamination events in the jth interval.

During sensor location selection, the node with the largest joint information entropy is selected each time. When selecting a new sensor location, the intervals obtained from the division of the existing sensors are further divided based on each node. The joint information entropy at each node is recalculated, and the node with the largest joint information entropy continues to be selected. New sensors are added, and the intervals are divided continuously until all contamination events are distributed in a single interval.

Considering cases of sensor interruptions, when implementing the optimal sensor placement, the sensors are selected sequentially based on the joint information entropy, thereby obtaining the sensor placement scheme under deterministic conditions. Subsequently, by considering sensor interruptions, each sensor is "refrigerated" and "unfreezed" in turn to obtain N types of improvement schemes. As shown in Fig. 1, there are two steps involved in obtaining the initial schemes: (1) obtain a set of initial sensor placement schemes; (2) obtain the impact of each sensor interruption.

2.4. Comparison and selection of sensor placement schemes

When comparing and selecting the various schemes, the main concerns are the number of sensors and the impact of sensor interruptions. The number of sensors affects the installation and maintenance costs and should be kept to a minimum. To minimize the impact of sensor interruptions, three functions are defined for the evaluation, including the contamination events "*unidentified impact*" and "*undetected impact*," as well as the impact of information loss. For optimal sensor placement scheme, this work considers four objective functions: minimizing contamination event-unidentified impact, contamination event "undetected impact", the loss of information entropy and the number of sensors.

2.4.1. Contamination event-unidentified impact

When there are no sensor interruptions, the sensor network can identify all contamination events. However, if a sensor is interrupted suddenly, some contamination events may not be identified. In this case, auxiliary positioning measures are needed to identify the locations of contamination events. The function of contamination event "*unidentified impact*" is as follows:

$$F_1 = min I_{Unidentified}(i) = min \frac{\sum_{s=1}^{n_i} \sum_{c=1}^{C} p(c) \cdot n_c}{n_i} \tag{12}$$

where $I_{Unidentified}(i)$ represents the contamination event "*unidentified impact*" when each sensor in scheme i is interrupted, $p(c)$ is the probability of contamination occurrence for cluster c, C is the number of clusters that cannot be identified when sensor s is interrupted, $s = 1, \cdots, n_i$ (n_i is the number of sensors in scheme i), and n_c is the number of nodes in cluster c.

2.4.2. Contamination event "undetected impact"

Compared with unidentified contamination events, undetected contamination events may result in more serious consequences. The impact is related to the number of undetected contamination events and their risk levels. The more undetected contamination events there are and the higher the risk levels they have, the greater the undetected impact. The contamination event "*undetected impact*" is as follows:

$$\begin{aligned} F_2 &= min I_{Undetected}(i) \\ &= min \frac{\sum_{s=1}^{n_i} \sum_{c=1}^{C} p(c) \cdot \sum_{j=1}^{n_c} \sum_{i=1}^{N} \sum_{t=\Delta t}^{A} {}^{\Delta t} l_{ij} \cdot D_i(t)}{n_i \cdot \sum_{i=1}^{N} \sum_{t=\Delta t}^{A} {}^{\Delta t} D_i(t)} \end{aligned} \tag{13}$$

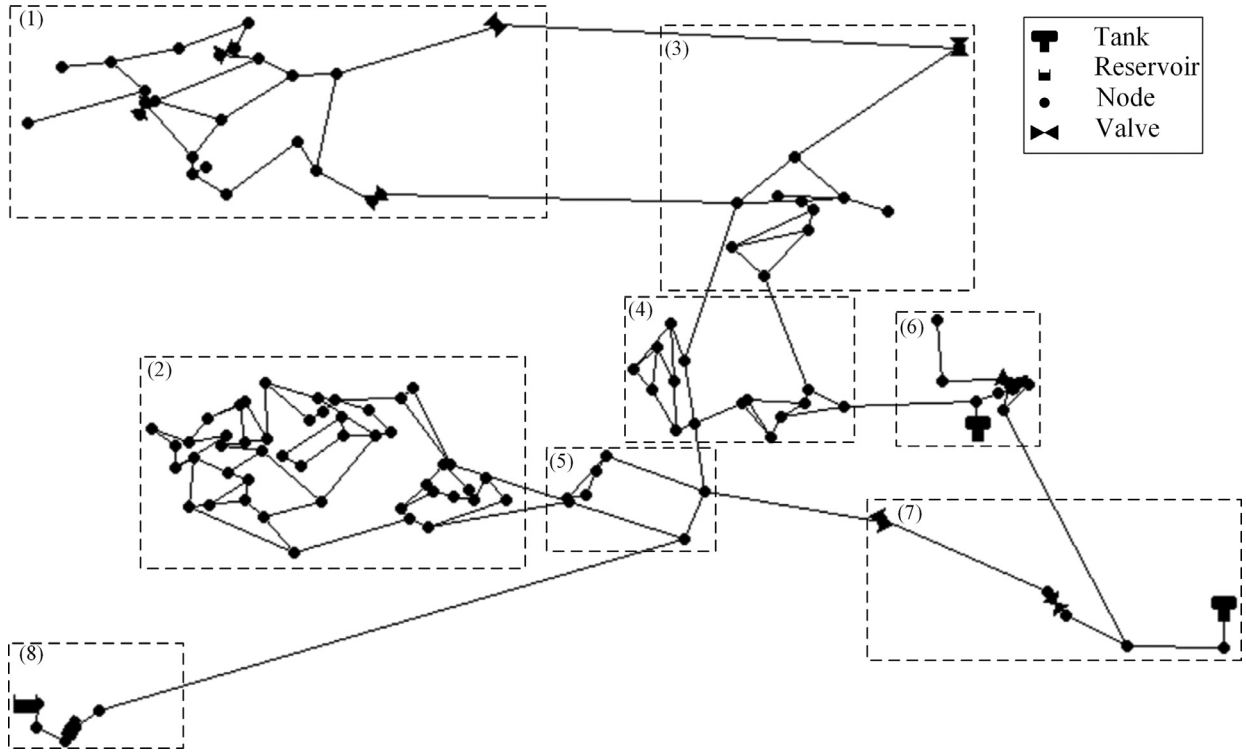

Fig. 2. Topological structure of the example WDN model *k1*.

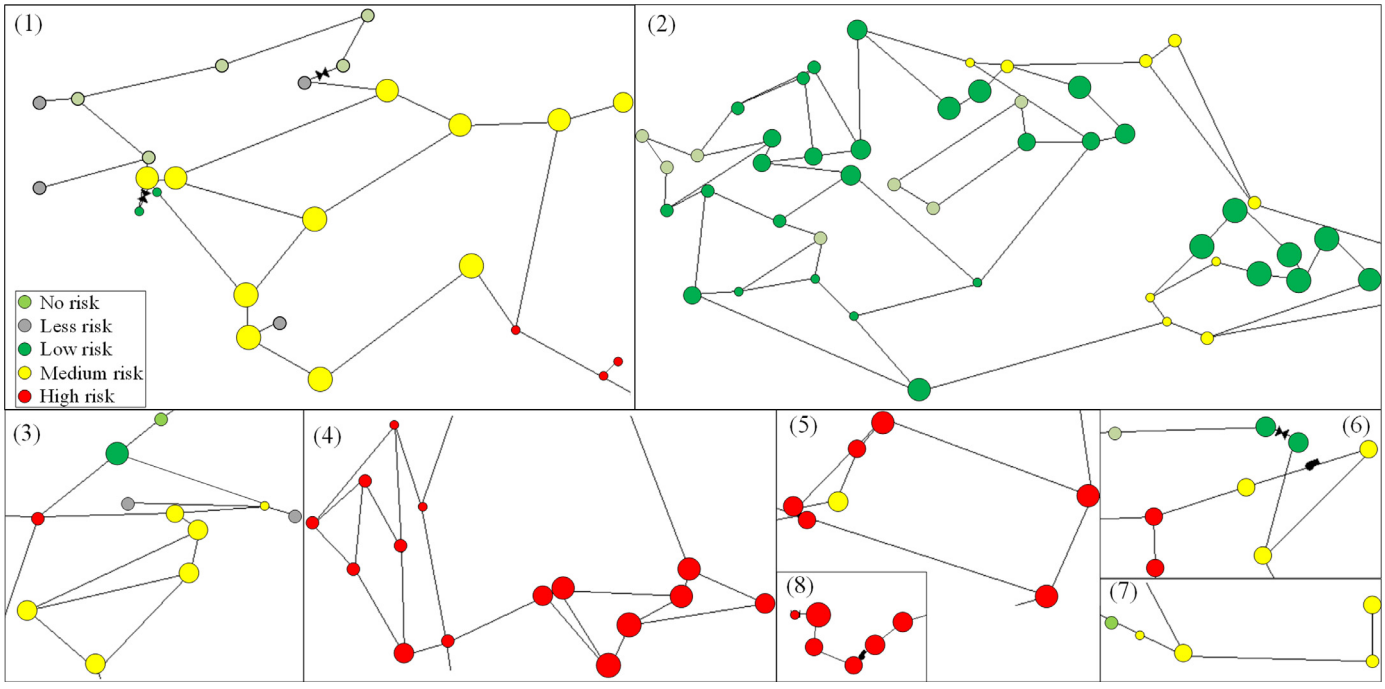

Fig. 3. Ranking of contamination risk levels of each node and the selection of contamination events.

where $I_{Undetected}(i)$ represents the contamination event-undetected impact when each sensor is interrupted in scheme i, and $p(c)$ is the probability of contamination occurrence for cluster c.

2.4.3. Impact of information loss

Information loss means the loss of information entropy of the sensor network during sensor interruptions; a smaller value indicates a smaller impact of sensor interruptions. The impact of information loss is

$$F_3 = minI_{Information\ loss}(i) = min\frac{n_i \cdot HSN - \sum_{s=1}^{n_i} HSN_s}{n_i \cdot HSN} \quad (14)$$

where $I_{Information\ loss}(i)$ represents the impact of information loss when each sensor in scheme i is interrupted, HSN is the joint information entropy of the sensor network when there are no sensor interruptions, HSN_s is the information loss of the sensor network when sensor s is interrupted.

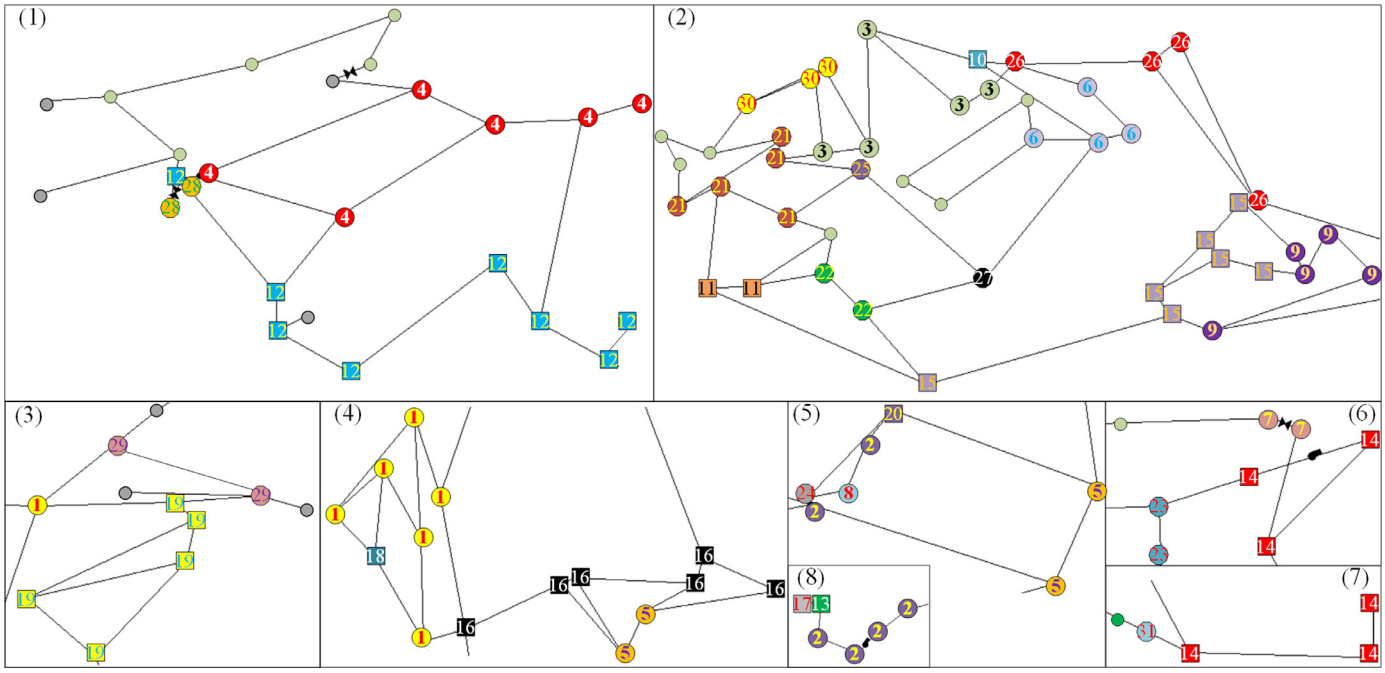

Fig. 4. Clustering results of various contamination events.

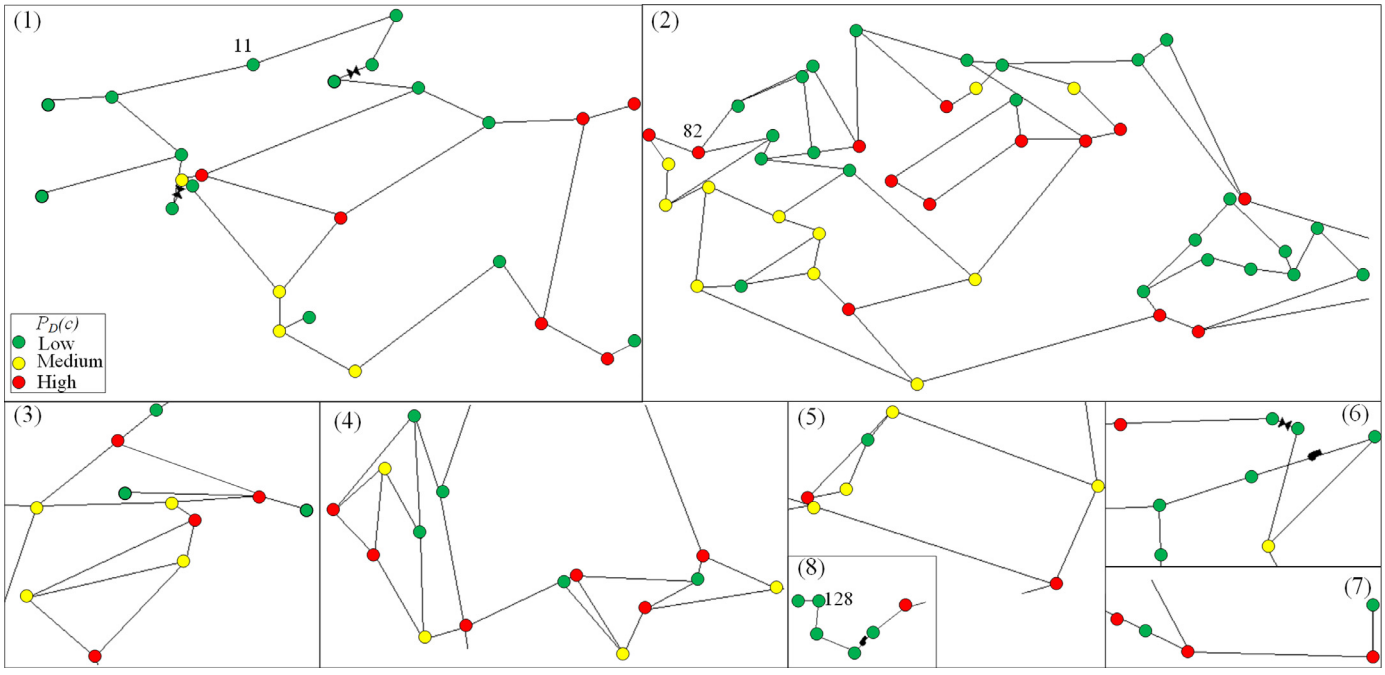

Fig. 5. Contamination probability distribution of each node obtained based on the node demand.

2.4.4. The number of sensors

To find the optimal sensor placement scheme, function F_4 is minimized.

$$F_4 = \min n_i \tag{15}$$

where n_i is the sensor number of scheme i.

According to the four types of performance evaluation criteria (contamination event-unidentified impact F_1, contamination event "undetected impact" F_2, impact of information loss F_3 and the number of sensors F_4), the various schemes are compared based on PROMETHEE to obtain the optimal scheme. PROMETHEE is a prioritization ranking method based on principal component comparison (Brans and Vincke, 1985), which is widely used in multi-criteria decision-making methods (Arcidiacono et al., 2018).

When given a set of candidate schemes A, the criteria $F = \{F_1, F_2, F_3, F_4\}$ are used for its evaluation. For each criterion $F_j \in F$, where $F_j : A \to R, j \in J = \{1, 2, 3, 4\}$; for each $F_j \in F$, $P_j(a, b)$ represents the preference for scheme a versus b based on the criterion F_j. The preference function is defined by Eq. (16):

$$P_j(a,b) = \begin{cases} 1 & if \quad F_j(a) - F_j(b) \geq p_j \\ \frac{[F_j(a) - F_j(b)] - q_j}{p_j - q_j} & if \quad F_j < F_j(a) - F_j(b) < p_j \\ 0 & if \quad F_j(a) - F_j(b) \leq q_j \end{cases} \tag{16}$$

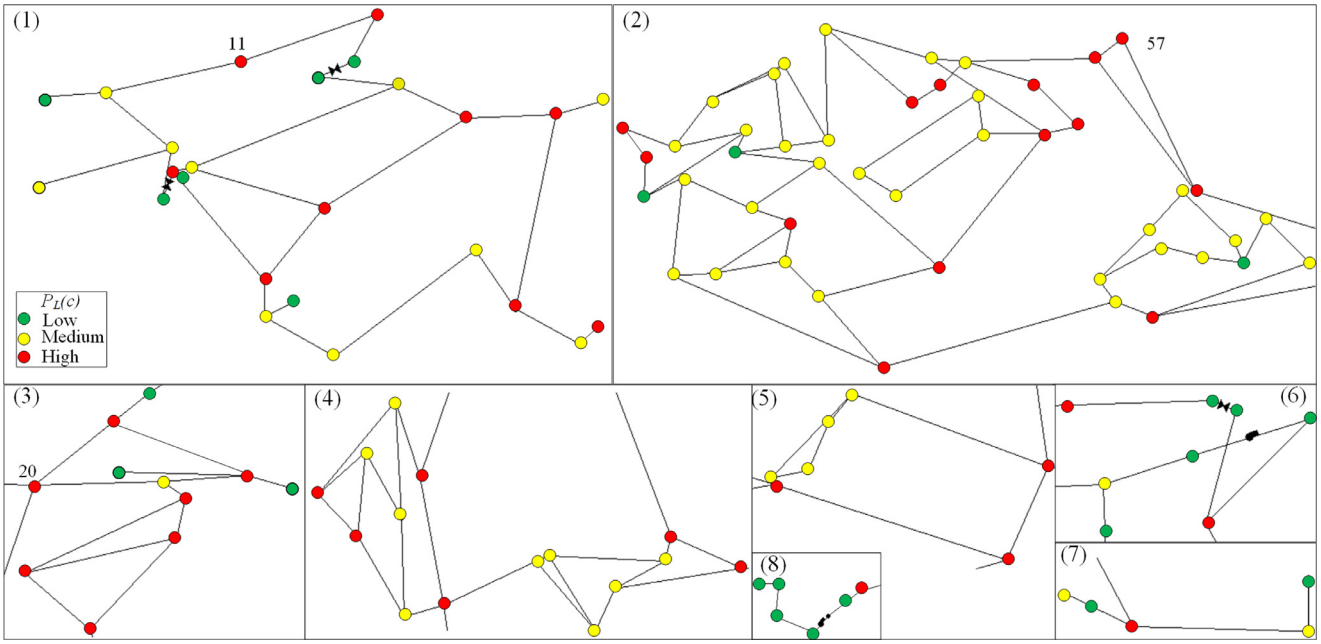

Fig. 6. Contamination probability distribution of each node obtained based on the pipe length.

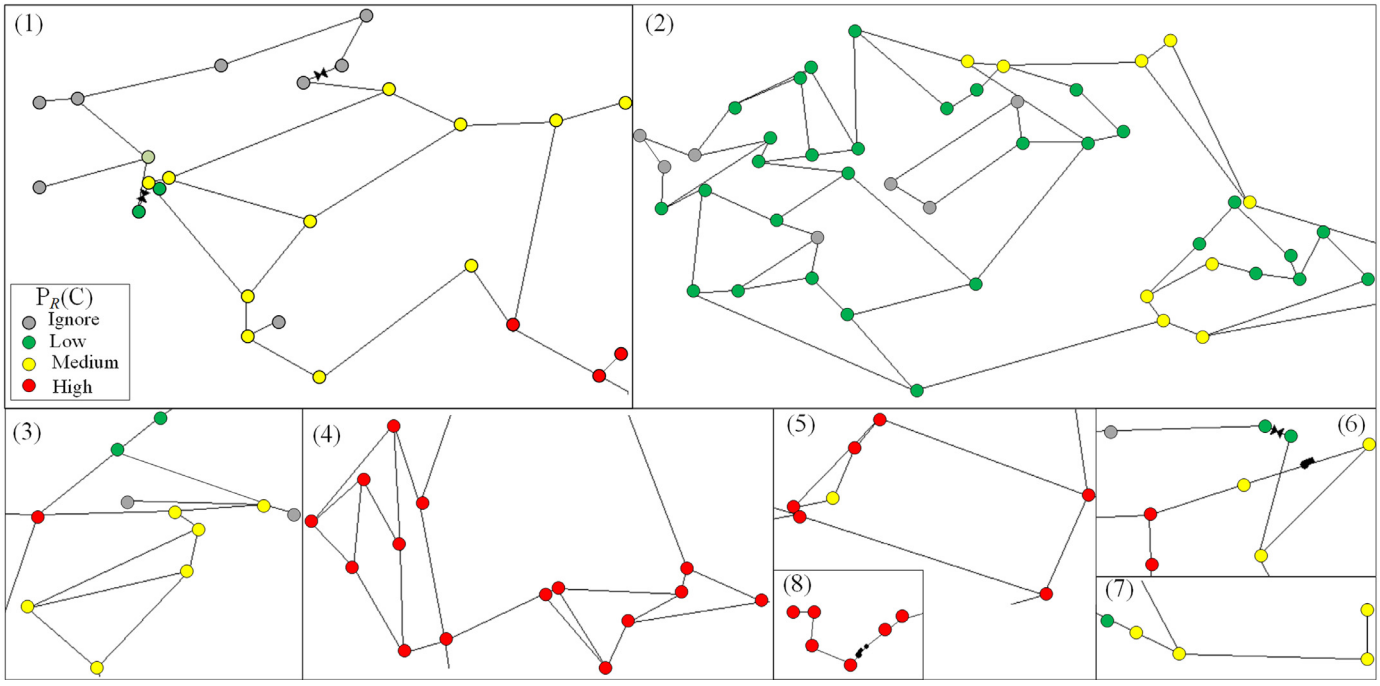

Fig. 7. Contamination probability distribution of each node obtained based on the contamination risk level.

where q_j and p_j represents the difference threshold and preference threshold of each criterion F_j, respectively. A weight w_j is assigned to each criterion, where $w_j > 0$ and $\sum_{j=1}^{4} w_j = 1$. For each pair of candidate schemes $(a, b) \in A \times A$, while considering all the criteria $g_j \in g$, the PROMETHEE method calculates the priority of a over b, as shown by Eq. (17):

$$\pi(a,b) = w_1 \cdot P_1(a,b) + w_2 \cdot P_2(a,b) + w_3 \cdot P_3(a,b) + w_4 P_4(a,b) \tag{17}$$

For each scheme $a \in A$, the positive flow $\Phi^+(a)$, negative flow $\Phi^-(a)$, and net flow $\Phi(a)$ can be calculated. $\Phi^+(a)$ represents the average preference degree for scheme a compared with the other schemes; $\Phi^-(a)$ represents the average preference degree for the other schemes compared with scheme a. $\Phi(a)$ represents the balance between positive flow and negative flow, and the strength of scheme a in the entire scheme set.

$$\Phi^+(a) = \frac{1}{|A|-1} \sum_{b \in A \setminus \{a\}} \pi(a,b) \tag{18}$$

$$\Phi^-(a) = \frac{1}{|A|-1} \sum_{b \in A \setminus \{a\}} \pi(b,a) \tag{19}$$

$$\Phi(a) = \Phi^+(a) - \Phi^-(a) \tag{20}$$

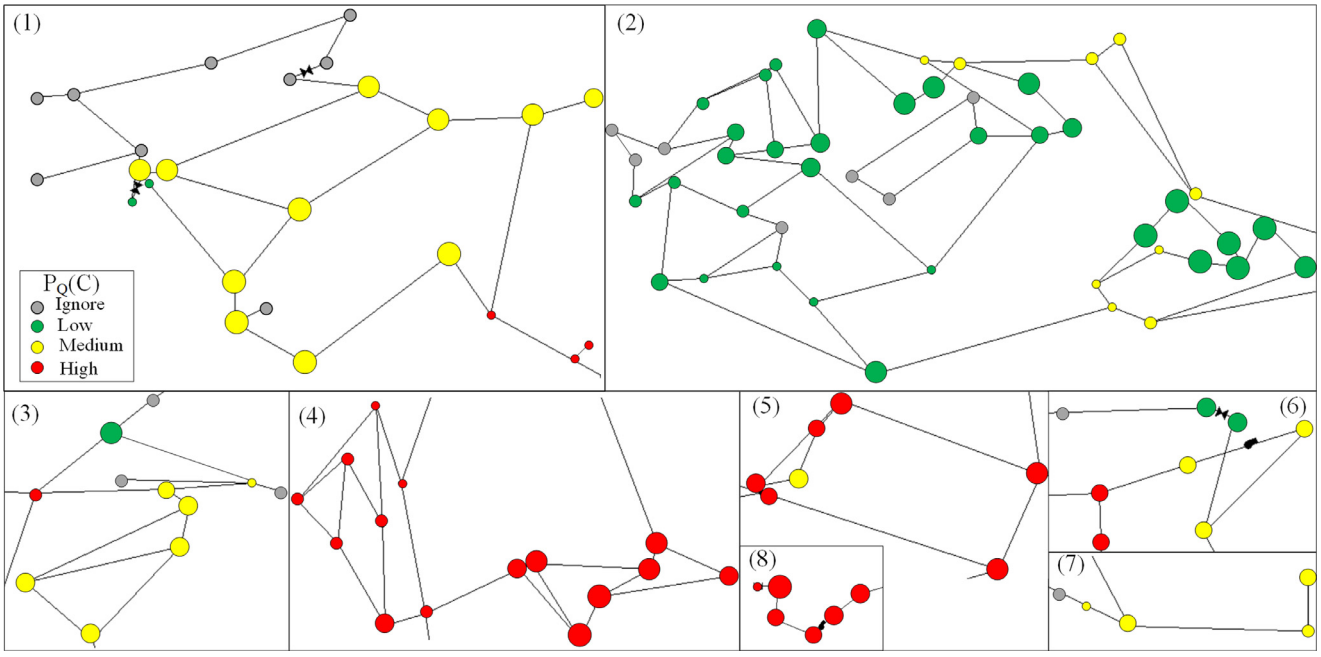

Fig. 8. Contamination probability distribution of each node obtained based on the quantified risk.

3. Case study

In this paper, the example WDN model *k1* was used to perform a case study, as shown in Fig. 2. The data were obtained from the Battle of the Water Sensors Network1 (BWSN1) competition (Ostfeld et al., 2008). The example WDN model contains a total of 126 nodes, 168 pipes, a constant headwater source, two tanks, two pumps, and eight valves. The simulation parameters were set as follows: the total simulation period was 96 h, the hydraulic time step was 10 min, the water quality time step was 5 min, and the mode time step was 10 min. The candidate sensor locations were the WDN nodes, so there were a total of 126 initial candidate locations. The contaminations occurred at the WDN nodes and the total number of contamination events was 129. To better illustrate the analysis results, in this study, the WDN topology was divided into eight parts.

4. Results and discussion

4.1. Contamination events selection and clustering

Based on the risk assessment results, the risk level of each node when contaminant intrusion occurred was obtained, as shown in Fig. 3. The nodes covered by black and light green circles in the figure represent nodes with extremely low contamination risk levels. When contaminant intrusions occurred at these nodes, they had little to no impact on the other nodes in the WDN. Thus, the cases when contaminant intrusions occurred at these nodes were ignored when selecting contamination events, and the total number of selected contamination events was 103. The events were ranked at three risk levels—namely high, medium, and low. A larger circle indicates a higher contamination risk.

Fig. 4 shows the clustering results of contamination events. The selected contamination events were divided into 31 clusters, where the contamination events at the same risk levels were usually grouped into the same cluster. When assessing the risk levels of contamination events, two factors—namely nodes affected by contamination events and node demand—were considered; however, the clustering only considered nodes affected by contamination events. Moreover, for contaminated nodes in the same cluster, their contaminations had a similar impact on the entire WDN. Therefore, it was a reasonable approach to cluster together the contaminated nodes.

4.2. Occurrence probabilities of contamination events

Based on the contamination probability function, the occurrence probability of contamination at each node could be obtained, as shown in Fig. 5 (node demand), Fig. 6 (node length), Fig. 7 (risk level), and Fig. 8 (quantified risk), respectively. Some contamination events with relatively high-risk levels (such as node 128 in Fig. 5, 5) had lower contamination probabilities due to the relatively low node demands or short lengths of pipes connected to them. Some contamination events with relatively low risk levels had higher contamination probabilities due to their relatively high node demands (such as node 82 in Fig. 5, 2) or long lengths of pipes connected to them (such as node 11 in Fig. 6, 1). The probabilities of various contamination events obtained based on node demands and pipe lengths were also different. For some low-demand nodes with long pipe lengths (such as node 57 in Fig. 6, 2), or some high-demand nodes with short pipe lengths (such as node 20 in Fig. 6, 3), their contamination probabilities exhibited different characteristics. When selecting contamination events, some were ignored based on their contamination risk levels. Due to higher node demands (such as node 82 in Fig. 5, 2) or longer connected pipe lengths (such as node 11 in Fig. 5, 1), the contamination probabilities were relatively higher. Within the same risk level, the occurrence probabilities of various contamination events obtained based on the risk level were the same; however, those obtained based on the quantified risk were different.

Fig. 9 shows the probabilities of the same contamination clusters under different contamination probabilities and impacts. The blue dotted line and the yellow histogram represent the contamination probability and contamination impact of each contamination cluster under an equal probability, respectively. The contamination probabilities of some contamination clusters obtained based on the node demand were 0 (such as cluster 7), and their contamination impacts were relatively small; however, the contamination

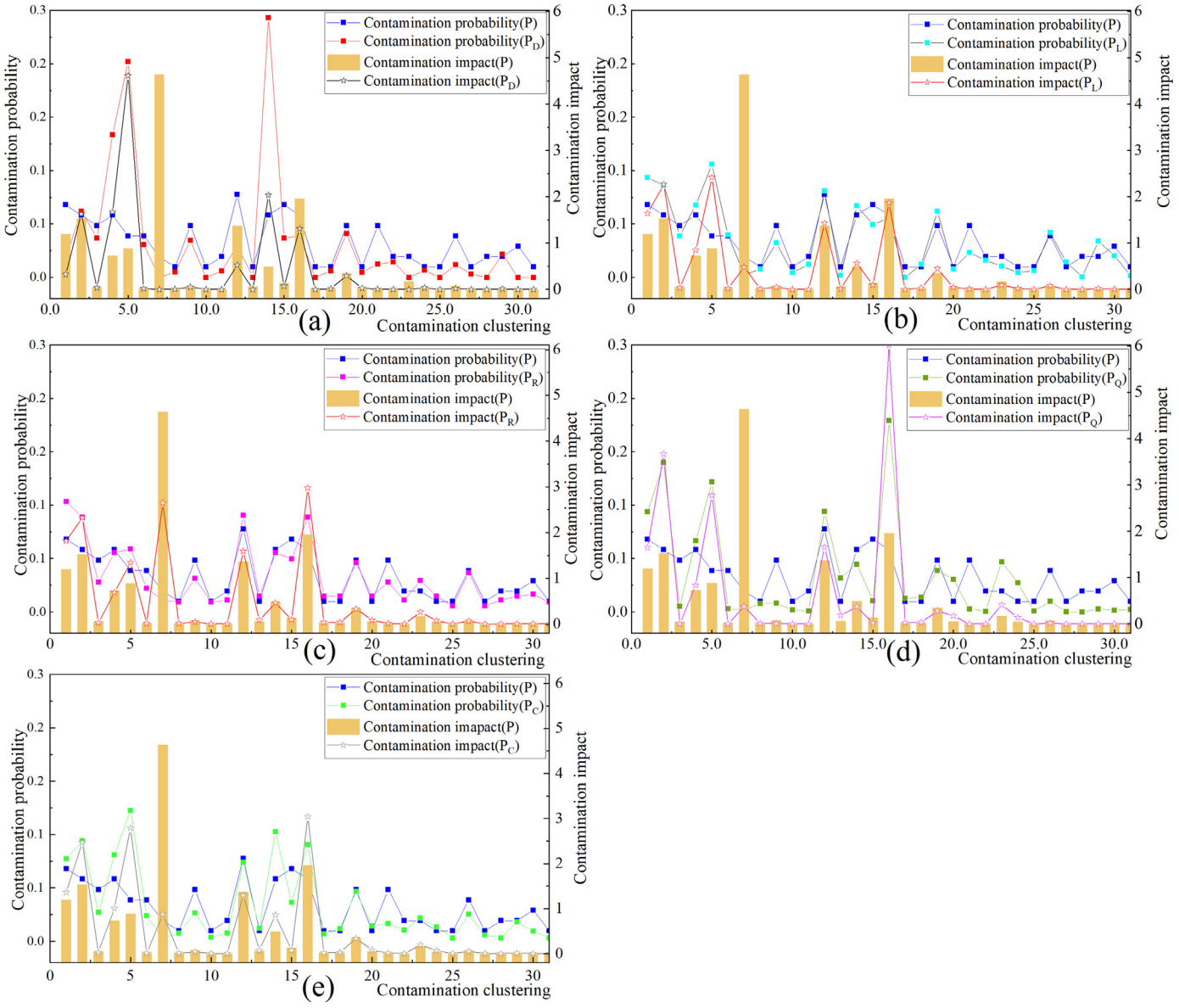

Fig. 9. Contamination probability and contamination impact of each contamination cluster: (a) based on node demand; (b) based on pipe length; (c) based on risk level; (d) based on quantified risk; (e) combined probability.

impacts of these clusters were relatively high under an equal probability and the other three types of changing probabilities. Some connecting nodes in the WDN had no user demand, but these nodes were at important locations in the pipeline, and their contaminations would impose a relatively large impact on the WDN. Therefore, it was unreasonable to ignore the zero-demand nodes in previous studies. When considering the variations of the contamination probability of each node, the contamination probabilities of some clusters were higher than the case where each node has an equal contamination probability. The change trends of cluster contamination probabilities and contamination impacts obtained based on node demand and pipe length were different; however, the cluster contamination probabilities and contamination impacts obtained based on node contamination risk level and quantified risk exhibited the same changing pattern.

4.3. Initial sensor placement scheme

The locations and numbers of sensors were selected based on a hierarchical algorithm; a total of 15 sensor placement schemes were obtained, as shown in Table 1. Scheme 1 did not consider cases with sensor interruptions, whereas schemes 2–15 considered sensor interruption scenarios. When there were no sensor interruptions, the 15 schemes could detect and identify all contamination events.

Fig. 10 shows the number of sensors in each scheme and the loss of information entropy during sensor interruptions. Scheme 1 did not consider cases with sensor interruptions; hence, it had the least number of sensors. However, it had the highest loss of information entropy during sensor interruptions than the other 14 schemes.

Fig. 11 shows the comparison of contamination event "*undetected*" and "*unidentified impacts*" during sensor interruptions under an equal contamination probability as well as under different contamination probabilities. Among them, the $p_R(c)$ of contamination events at three risk levels were high (0.8), medium (0.5), and low (0.3), respectively. When deploying the same scheme, the performances of the sensor network were different under different contamination probabilities. Under an equal contamination probability, the contamination event "*undetected impact*" was higher than

Table 1
Comparison of sensor placement schemes.

Scheme	Sensor placement nod	Scheme ranking					
		$p(c)$	$P_D(c)$	$p_L(c)$	$p_R(c)$	$p_Q(c)$	$p_C(c)$
1	[81,17, 45, 82, 76, 12, 126, 86, 72, 37, 121, 92, 62,98]	10	10	3	10	10	10
2	[81,79, 117, 75, 12, 42, 86, 84, 82, 45, 121, 14, 62, 72, 92, 76,98]	11	13	11	12	14	12
3	[81, 17,37, 45, 126, 76, 82, 12, 86, 72, 121, 0, 92, 62,98]	6	5	2	6	5	6
4	[81, 17, 121, 82, 76, 12, 46, 35, 72, 86, 4, 62, 92, 98,45,117]	1	3	6	1	2	2
5	[81, 17, 45, 76, 126, 83, 12, 86, 37, 121, 62, 72, 92, 79,82,98]	4	4	7	4	4	4
6	[81, 17, 45, 82, 12, 86, 74, 126, 37, 77, 121, 62, 92, 72,76,98]	5	8	5	8	9	9
7	[81, 17, 45, 82, 76, 14, 86, 37, 72, 121, 10, 62, 92, 98,12,117]	8	6	1	5	6	5
8	[81, 17, 45, 82, 76, 12, 35, 72, 86, 117, 121, 4, 92, 62,126,98]	2	1	4	2	1	1
9	[81, 17, 45, 82, 76, 12, 126, 72, 98, 35, 104, 4, 92, 62,86,121]	3	2	13	3	3	3
10	[81, 17, 45, 82, 76, 12, 126, 86, 74, 37, 121, 92, 62,72,98]	13	14	10	13	12	14
11	[81, 17, 45, 82, 76, 12, 126, 86, 72, 39, 121, 92, 62,98,37]	12	11	12	11	11	11
12	[81, 17, 45, 82, 76, 12, 126, 86, 72, 37, 122, 92, 62,121,98]	14	12	14	14	13	13
13	[81, 17, 45, 82, 76, 12, 126, 86, 72, 37, 121, 93, 62,92,98]	15	15	15	15	15	15
14	[81, 17, 45, 82, 76, 12, 126, 86, 72, 37, 121, 92, 63,62,98]	7	7	8	7	7	7
15	[81, 17, 45, 82, 76, 12, 126, 86, 72, 37, 121, 92, 62,98,99]	9	9	9	9	8	8

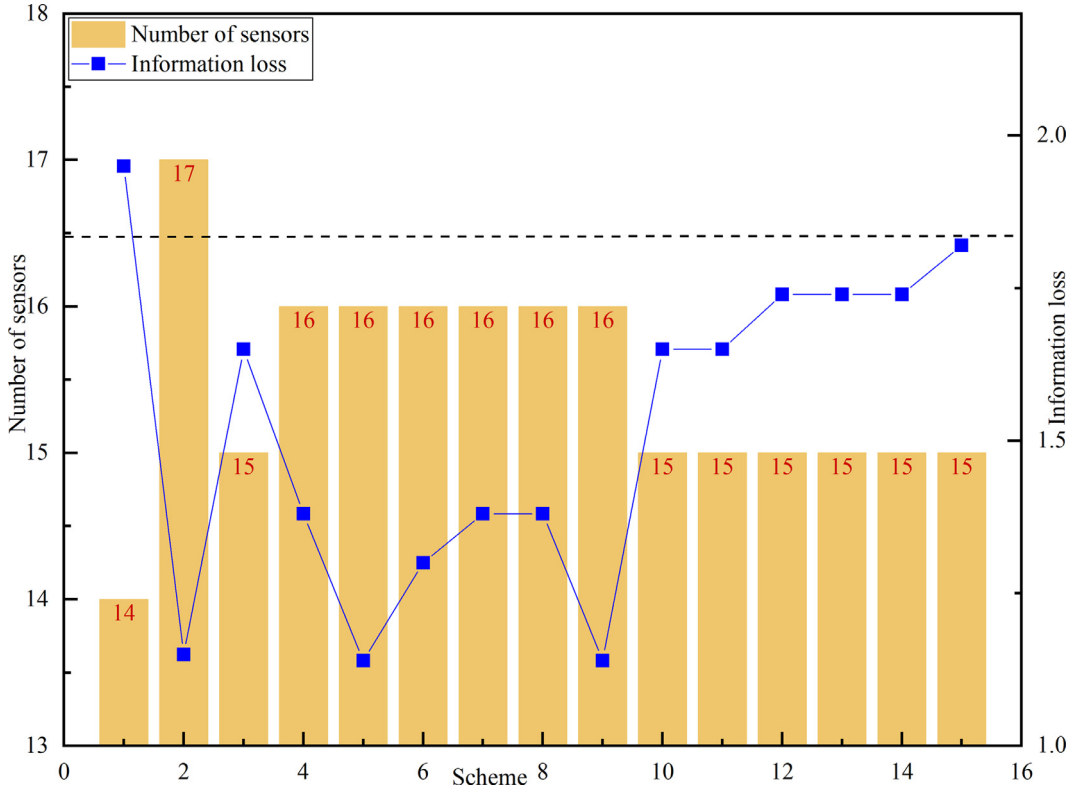

Fig. 10. Number of sensors and loss of information entropy for different sensor placement schemes.

that under different contamination probabilities. This is because the greater the impact of a contamination event, the more the number of WDN nodes were affected by the contamination event. These events were also more likely to be detected by multiple sensors simultaneously. Thus, the interruption of a single sensor had a smaller impact on the detection of contamination events with a greater contamination impact. As for the contamination event "*unidentified impact*," the same scheme exhibited the same trend under different contamination probabilities, but the impact levels were different. Therefore, the variations of contamination probabilities had to be considered when selecting the sensor placement scheme.

4.4. Selection of sensor placement scheme

The various schemes were compared using the PROMETHEE method to obtain the optimal scheme. According to the four optimization criteria—namely the number of sensors, the information loss during sensor interruptions, the contamination event "*undetected*" and "*unidentified impacts*" in each scheme—the ranking of the optimal sensor placement schemes under different contamination probabilities was obtained. The same weight was assigned to each criterion, as shown in Table 1. As the occurrence probabilities of contamination events varied, the ranking of optimal sensor placement schemes was also different. When each node had equal contamination probability, schemes 4, 7, and 8 were the best ac-

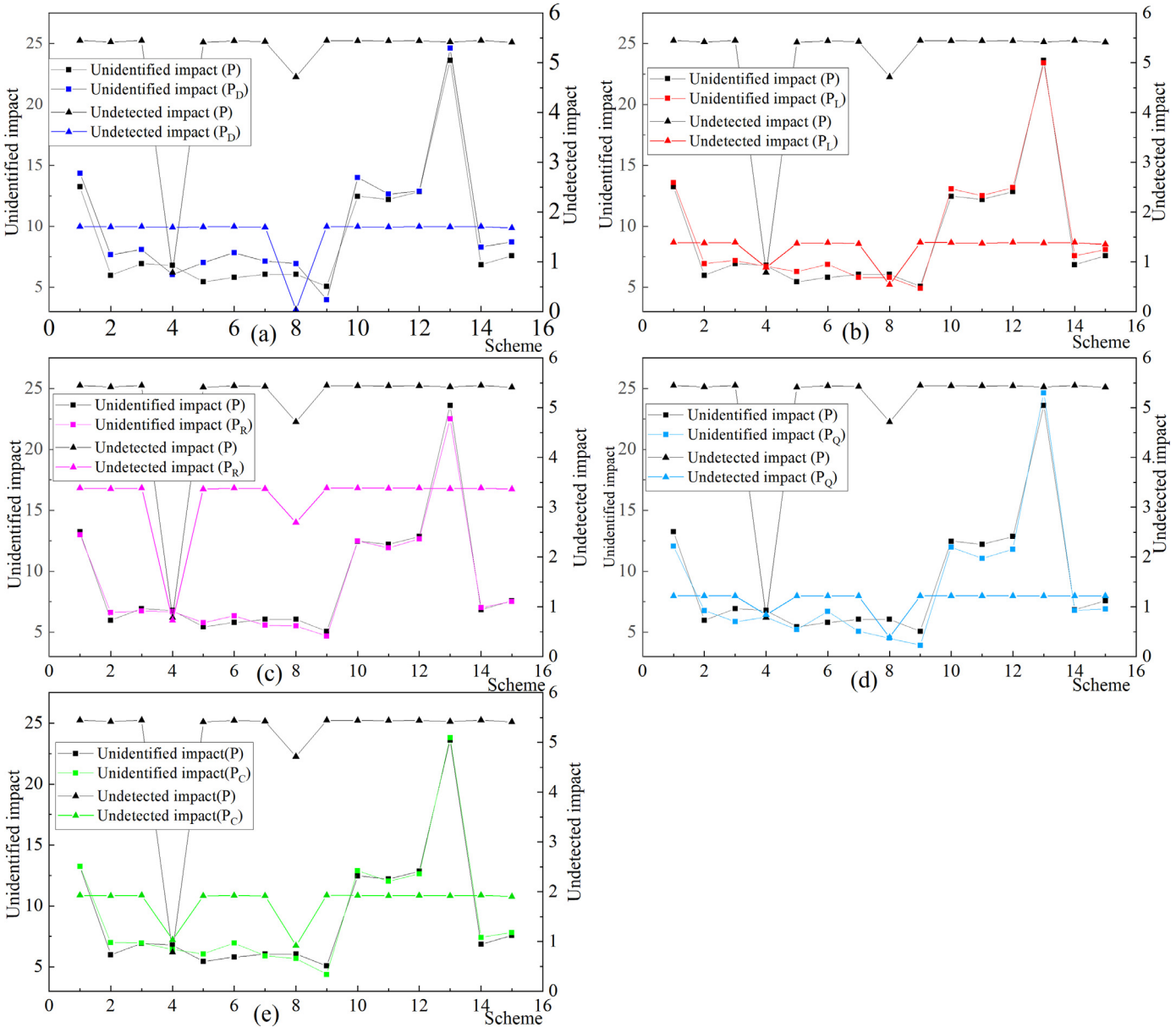

Fig. 11. Contamination event "undetected" and "unidentified impacts" under equal and different contamination probabilities: (a) based on node demand; (b) based on pipe length; (c) based on risk level; (d) based on quantified risk; (e) combined probability.

Table 2
Comparison of contaminant intrusion detection and identification performances of various schemes under different contamination probabilities.

Scheme	n_i	$I_{Information\ losss}$	$I_{Unidentified}(i)$						$I_{Undetected}(i)$					
			$p(c)$	$P_D(c)$	$p_L(c)$	$p_R(c)$	$p_Q(c)$	$p_C(c)$	$p(c)$	$P_D(c)$	$p_L(c)$	$p_R(c)$	$p_Q(c)$	$p_C(c)$
1	14	1.95	13.23	14.35	13.59	13	12.07	13.25	5.45	1.71	1.39	3.39	1.22	1.93
4	16	1.38	6.79	6.05	6.71	6.7	6.23	6.42	0.78	1.69	0.89	0.73	0.84	1.04
7	16	1.38	6.05	7.13	5.8	5.59	5.08	5.90	5.42	1.69	1.37	3.37	1.22	1.91
8	16	1.38	6.06	6.94	5.81	5.53	4.5	5.70	4.72	0.04	0.54	2.69	0.38	0.91

cording to the node contamination probability obtained based on the risk level, pipe length, and the combined of four probabilities, node demand and quantified risk, respectively. The various optimal sensor placement schemes are shown in Fig. 12.

Table 2 shows the contamination event "*undetected*" and "*unidentified impacts*" for each scheme under different contamination probabilities. In each scheme, the contamination event "*undetected*" and "*unidentified impacts*" based on the combined probability are in the middle position. Under the various contamination probabilities, the $I_{Unidentified}(i)$, $I_{Undetected}(i)$, and $I_{Information\ loss}(i)$ of scheme 1 during sensor interruptions were all higher than the other three schemes. To achieve better performance of the sensor network in practice, the interruption of sensors should be considered when placing the sensors.

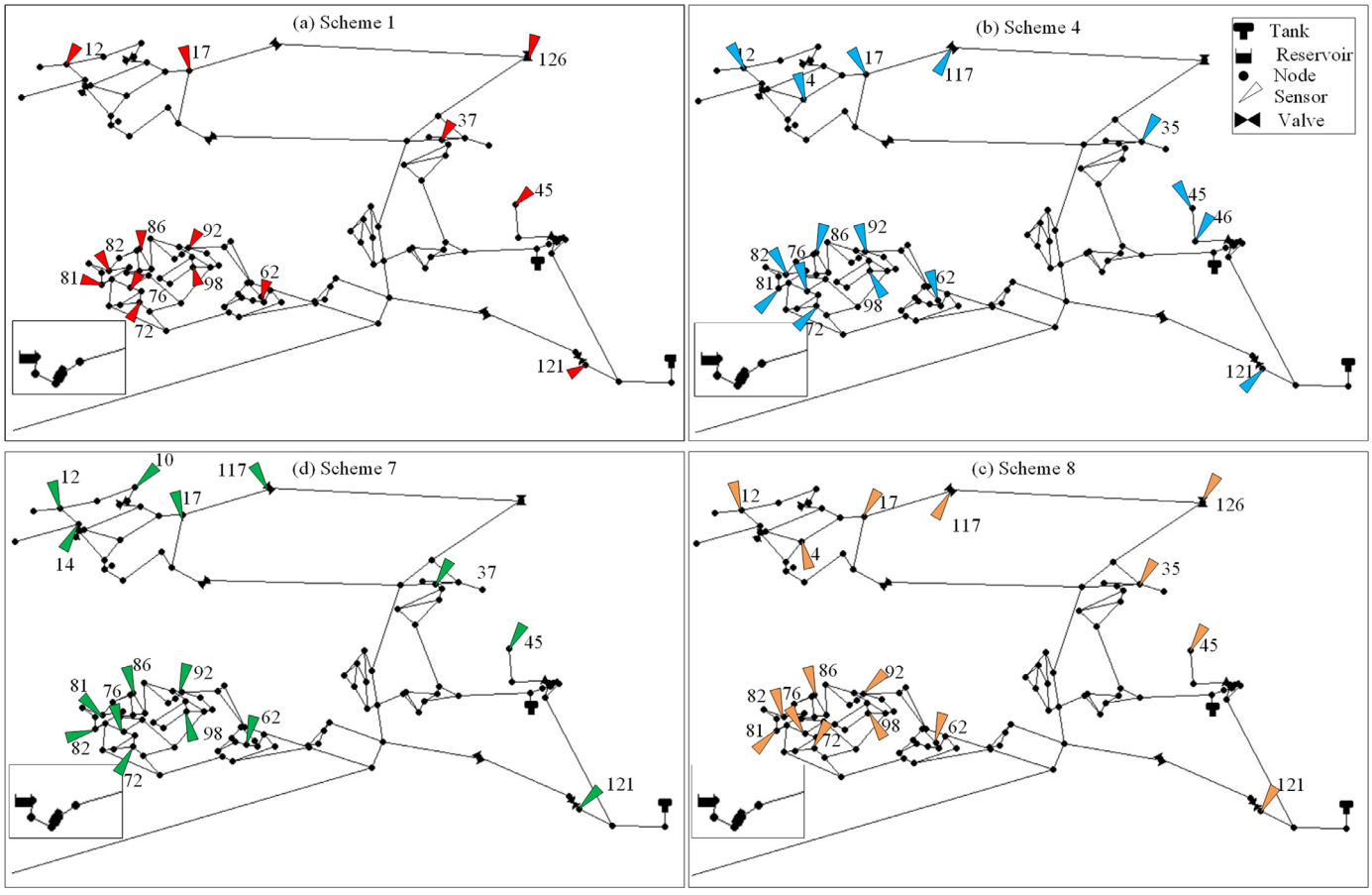

Fig. 12. Optimal sensor placement schemes.

In the case of intentional contamination in WDNs, considerations of variations of node contamination probabilities can help to establish an effective monitoring system (Cardoso et al., 2021). The risk level and quantified risk consider the impacts of contamination events on the WDN. The greater the contamination impact, the higher the contamination probability. The node demand and pipe length considered the influence of nodes' characteristics on the contamination probability. Some connection points of the WDN (non-user nodes) had a zero demand. However, the intrusions of contaminants at these nodes also affected most of the nodes. Although the contamination probabilities obtained based on the risk level distinguished the contamination probability of each node, most nodes had an equal contamination probability.

According to both $p(c)$ and $p_R(c)$, scheme 4 was the best. Under an equal contamination probability for each node, scheme 4 could ensure the detection of most contamination events in the case of sensor interruptions, thereby minimizing undetected impact. Owing to the limited number of sensors, effective monitoring of contamination events with higher risk levels and higher occurrence probabilities was necessary. As an important mitigation measure, mitigating the actual or potential consequences is a long-term and permanent goal (Berglund et al., 2020). Compared with scheme 4, scheme 8 ensured the detection and identification of events with high contamination probabilities (ignoring some events with low contamination probabilities). Under $p_Q(c)$, the "*unidentified*" and "*undetected impacts*" of scheme 4 were 6.23 and 0.84, respectively; and the "*unidentified* "and "*undetected impacts*" of scheme 8 were 4.50 and 0.38, respectively.

5. Conclusions

The optimal sensor placement is of great significance to the monitoring of contamination intrusions in WDNs. The identification of contaminant intrusions using a sensor network is an important optimization objective, which can ensure that it responds to the intrusions of contaminants and identifies the corresponding locations. Previous studies on the identification of contaminant intrusion locations did not consider the variations of contamination probabilities of each intrusion point. However, owing to the different characteristics of each node or the risk levels of contamination events, the occurrence probabilities of different contamination events are different. Such differences will affect the selection of sensor locations, thereby affecting the detection of contaminant intrusions.

This study proposes a sensor placement optimization method that considers different contamination probabilities. The results showed that different contamination probabilities of the various contamination events resulted in different optimal sensor placement schemes. Optimal placement schemes obtained under an equal contamination probability could ensure the detection and identification of various contamination events; whereas schemes obtained by considering the variations of contamination probabilities were more inclined toward detecting and identifying events with high contamination probabilities.

With a limited number of sensors, a more practical scheme could be obtained by considering the variations of contamination probabilities of contamination events. Future work needs to further consider different contamination probabilities and other optimiza-

tion objectives for sensor placement, such as the minimum detection time of contamination events and the minimum population affected by them.

Author contributions

Zukang Hu conducted the primary experiments, cartography, and analyzed the results. Debao Tan provided the original idea for this paper. Wenlong Chen, Dingtao Shen, Beiqing Chen and Song Ye actively participated throughout the research process and offered data support for this work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests

Acknowledgement

This research was funded by The National Key R&D Program of China (grant number 2019YFC0408805) and Key Technology Application and Demonstration of Water Conservation Society Innovation Pilot in Jinhua, Zhejiang (grant number SF-201801).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2021.107404.

References

Arcidiacono, S.G., Corrente, S., Greco, S., 2018. GAIA-SMAA-PROMETHEE for a hierarchy of interacting criteria. Eur. J. Oper. Res. 270 (2), 606–624. doi:10.1016/j.ejor.2018.03.038.

Berglund, E.Z., Pesantez, J.E., Rasekh, A., Shafiee, M.E., Sela, L., Haxton, T., 2020. Review of modeling methodologies for managing water distribution security. J. Water Resour. Plann. Manage. 146 (8), 03120001. doi:10.1061/(asce)wr.1943-5452.0001265.

Berry, J., Hart, W.E., Phillips, C.A., Uber, J.G., Watson, J.-P., 2006. Sensor placement in municipal water networks with temporal integer programming models. J. Water Resour. Plann. Manage. 132 (4), 218–224. doi:10.1061/(asce)0733-9496(2006)132:4(218).

Bertola, N.J., Papadopoulou, M., Vernay, D., Smith, I.F.C., 2017. Optimal multi-type sensor placement for structural identification by static-load testing. Sensors (Switzerland) 17 (12), 2904. doi:10.3390/s17122904.

Bertola, N.J., Cinelli, M., Casset, S., Corrente, S., Smith, I.F.C., 2019. A multi-criteria decision framework to support measurement-system design for bridge load testing. Adv. Eng. Inf. 39, 186–202. doi:10.1016/j.aei.2019.01.004, December 2018.

Brans, J.P., Vincke, P., 1985. Note—a preference ranking organisation method. Manage. Sci. 31 (6), 647–656. doi:10.1287/mnsc.31.6.647.

Cardoso, S.M., Barros, D.B., Oliveira, E., Brentan, B., Ribeiro, L., 2021. Optimal sensor placement for contamination detection: a multi-objective and probabilistic approach. Environ. Modell. Softw. 135, 104896. doi:10.1016/j.envsoft.2020.104896, October 2020.

Carr, R.D., Greenberg, H.J., Hart, W.E., Konjevod, G., Lauer, E., Lin, H., Morrison, T., Phillips, C.A., 2006. Robust optimization of contaminant sensor placement for community water systems. Math. Program. 107, 337–356. doi:10.1007/s10107-005-0689-x.

Cheikh, M., Jarboui, B., Loukil, T., Siarry, P., 2010. A method for selecting Pareto optimal solutions in multiobjective optimization. J. Informatics Math. Sci. 2, 51–62.

Ciaponi, C., Creaco, E., Di Nardo, A., Di Natale, M., Giudicianni, C., Musmarra, D., Santonastaso, G.F., 2019. Reducing impacts of contamination in water distribution networks: a combined strategy based on network partitioning and installation of water quality sensors. Water (Switzerland) 11 (6), 1–16. doi:10.3390/w11061315.

Comboul, M., Ghanem, R., 2013. Value of information in the design of resilient water distribution sensor networks. J. Water Resour. Plann. Manage. 139 (4), 449–455. doi:10.1061/(asce)wr.1943-5452.0000259.

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 Res. 143 (2018), 165–175. doi:10.1016/j.watres.2018.06.041.

Hooshmand, F., Amerehi, F., MirHassani, S.A., 2020. Logic-based benders decomposition algorithm for contamination detection problem in water networks. Comput. Oper. Res. 115, 104840. doi:10.1016/j.cor.2019.104840.

Jung, D., Kim, J.H., 2018. Using mechanical reliability in multiobjective optimal meter placement for pipe burst detection. J. Water Resour. Plann. Manage. 144 (7), 04018031. doi:10.1061/(asce)wr.1943-5452.0000953.

Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J., Faloutsos, C., 2008. Efficient sensor placement optimization for securing large water distribution networks. J. Water Resour. Plann. Manage. 134 (6), 516–526. doi:10.1061/(asce)0733-9496(2008)134:6(516).

Ostfeld, A., Uber, J.G., Salomons, E., Berry, J.W., Hart, W.E., Phillips, C.A., Watson, J.-.P., Dorini, G., Jonkergouw, P., Kapelan, Z., di Pierro, F., Khu, S.-.T., Savic, D., Eliades, D., Polycarpou, M., Ghimire, S.R., Barkdoll, B.D., Gueli, R., Huang, J.J., Walski, T., 2008. The Battle of the Water Sensor Networks (BWSN): a design challenge for engineers and algorithms. J. Water Resour. Plann. Manage. 134 (6), 556–568. doi:10.1061/(asce)0733-9496(2008)134:6(556).

Palleti, V.R., Narasimhan, S., Rengaswamy, R., Teja, R., Murty Bhallamudi, S., 2016. Sensor network design for contaminant detection and identification in water distribution networks. Comput. Chem. Eng. 87, 246–256. doi:10.1016/j.compchemeng.2015.12.022.

Perelman, L., Ostfeld, A., 2010. Extreme impact contamination events sampling for water distribution systems security. J. Water Resour. Plann. Manage. 136 (1), 80–86. doi:10.1061/(asce)0733-9496(2010)136:1(80).

Preis, A., Ostfeld, A., 2008. Multiobjective contaminant response modeling for water distribution systems security. J. Hydroinf. 10 (4), 267–274. doi:10.2166/hydro.2008.061.

Rathi, S., Gupta, R., Kamble, S., Sargaonkar, A., 2016. Risk based analysis for contamination event selection and optimal sensor placement for intermittent water distribution network security. Water Resour. Manage. 30 (8), 2671–2685. doi:10.1007/s11269-016-1309-7.

Sela, L., Amin, S., 2018. Robust sensor placement for pipeline monitoring: mixed integer and greedy optimization. Adv. Eng. Inf. 36, 55–63. doi:10.1016/j.aei.2018.02.004, January.

Shen, H., McBean, E., 2011. Pareto optimality for sensor placements in a water distribution system. J. Water Resour. Plann. Manage. 137 (3), 243–248. doi:10.1061/(asce)wr.1943-5452.0000111.

Steinley, D., 2006. K-means clustering: a half-century synthesis. Br. J. Math. Stat. Psychol. 59 (1), 1–34. doi:10.1348/000711005X48266.

Weickgenannt, M., Kapelan, Z., Blokker, M., Savic, D.A., 2010. Risk based sensor placements for contaminant detection in water distribution systems. J. Water Resour. Plann. Manage. 136 (6), 629–636.

de Winter, C., Palleti, V.R., Worm, D., Kooij, R., 2019. Optimal placement of imperfect water quality sensors in water distribution networks. Comput. Chem. Eng. 121, 200–211. doi:10.1016/j.compchemeng. 2018.10.021.