Supplementary Information for

"Optimal Joint Deployment of Flow and Pressure Sensors for Leak Identification in Water Distribution Networks"

Ehsan Raei, Mohammad Reza Nikoo, Shokoufeh Pourshahabi, Mojtaba Sadegh

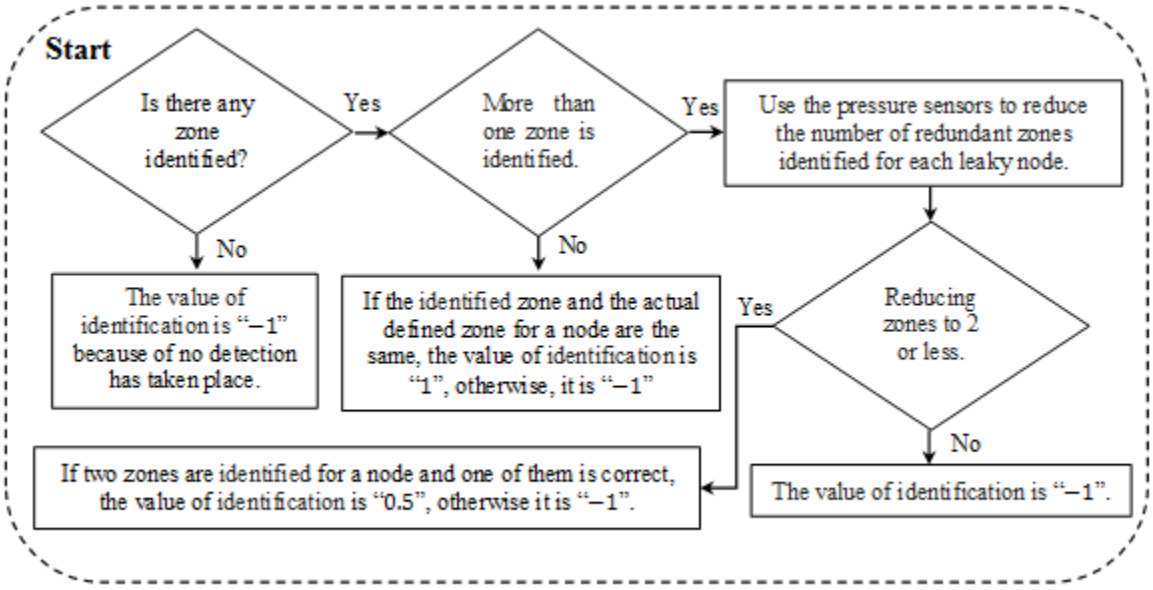
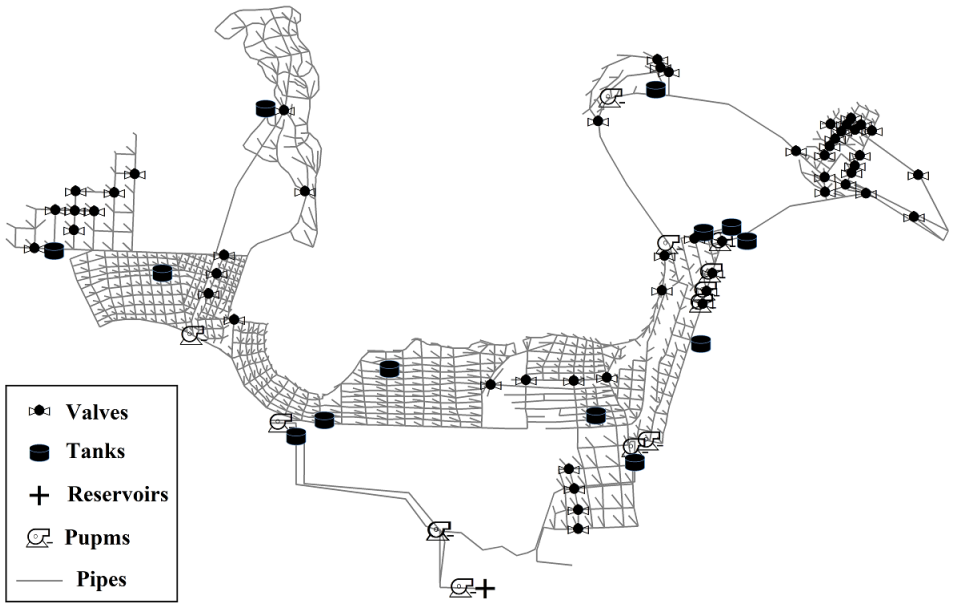


Fig. S1 A Sub-algorithm which should be performed in step 12 of the flowchart of the proposed methodology to specify the value of identification

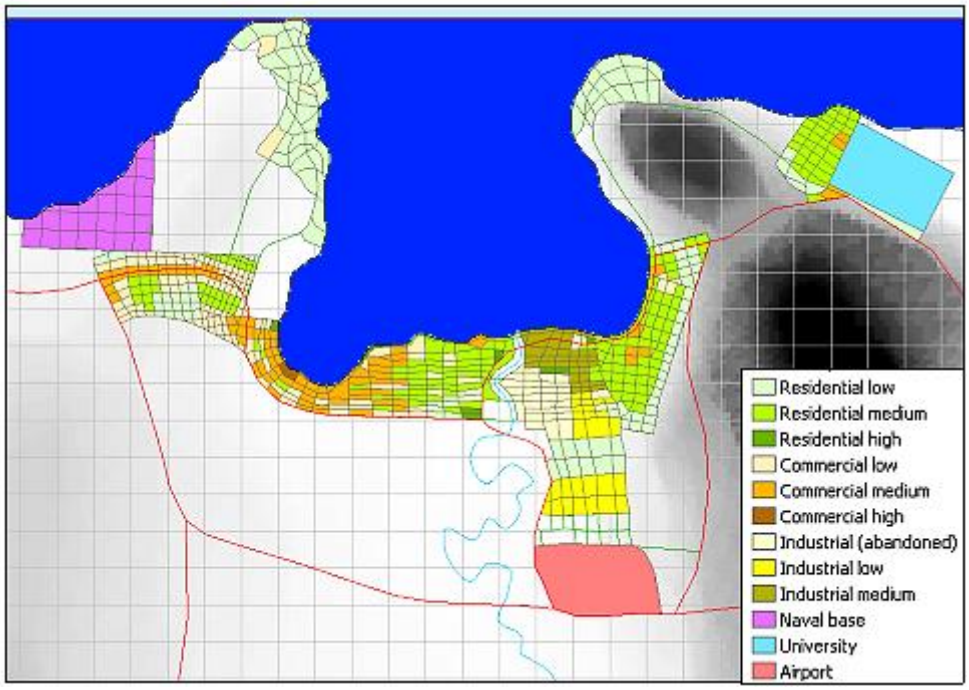
Input: Flow divergence matrix for each leakage scenario (M)
Output: $\text{best}_{\text{threshold}}$
$index_t$ = 0
For threshold $=$ 1 To max (M) **do**
    $index_t = index_t + 1$
    M (abs $(M) <$ threshold) $= 0$
    $u =$ cell of M is bigger than zero; specify row of cell; return the smallest row
    $index_s = 0$
    For $i = u$ to $u + 4$ **do**
        $index_s = index_s + 1$
        $\text{similarity}_{\text{node}}$ $(index_s)$ = Similarity between nodes (column(s) of M which is/are bigger than zero) of row i and row $i+1$
        num_{node} $(index_s)$ = number of node(s) in row i (column bigger than zero)
    end For
    $\text{data}_{\text{similaritynode}}$ $(index_t) =$ mean $(\text{similarity}_{\text{node}})$
    $\text{data}_{\text{numnode}}(index_t) = \text{mean}$ $(\text{num}_{\text{node}})$
end For
$\text{best}_{\text{threshold}} = \text{find}(\text{data}_{\text{similaritynode}} > 0.7 \text{ and } 15 \geq \text{data}_{\text{numnode}} \geq 5)$; return threshold(s)

Fig. S2 Pseudo code for specifying the tolerance threshold for flow divergence matrices.

Using Pseudo code presented in Fig. S2, the tolerance threshold is specified so that a limited number of pipes (between 5 to 15 pipes) show flow changes in a number of successive time steps when the leakage occurs in the node n_i. This process is performed for all leakage scenarios and the tolerance threshold that has been repeated the most among all leakage scenarios is selected.



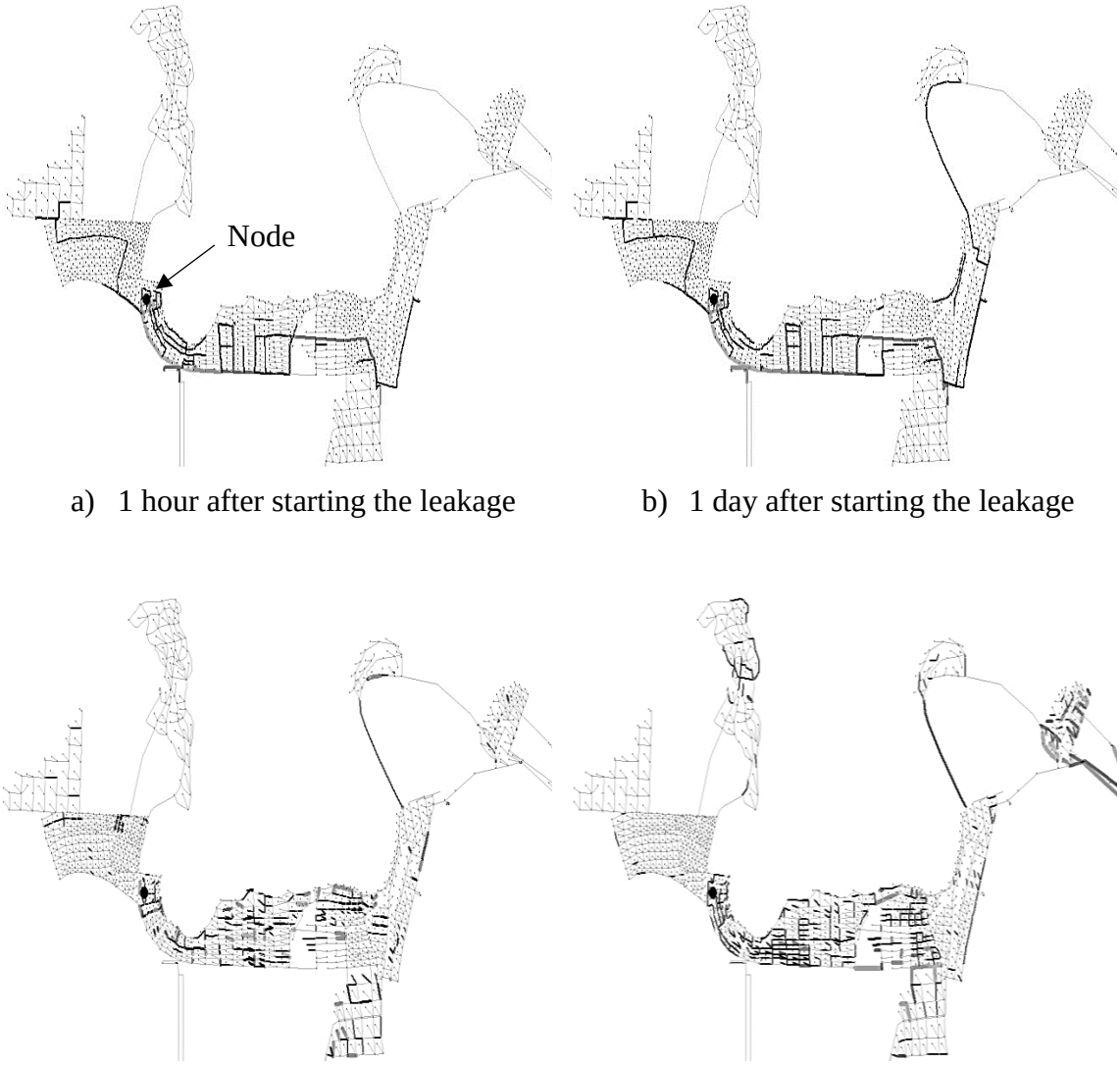
(a) Mesopolis Network equipment



(b) Land use map of Mesopolis (Johnston and Brumbelow 2008)

Fig. S3 Mesopolis water distribution network

Fig. S4 The elevation of the nodes in Mesopolis water distribution network



a) 1 hour after starting the leakage

b) 1 day after starting the leakage

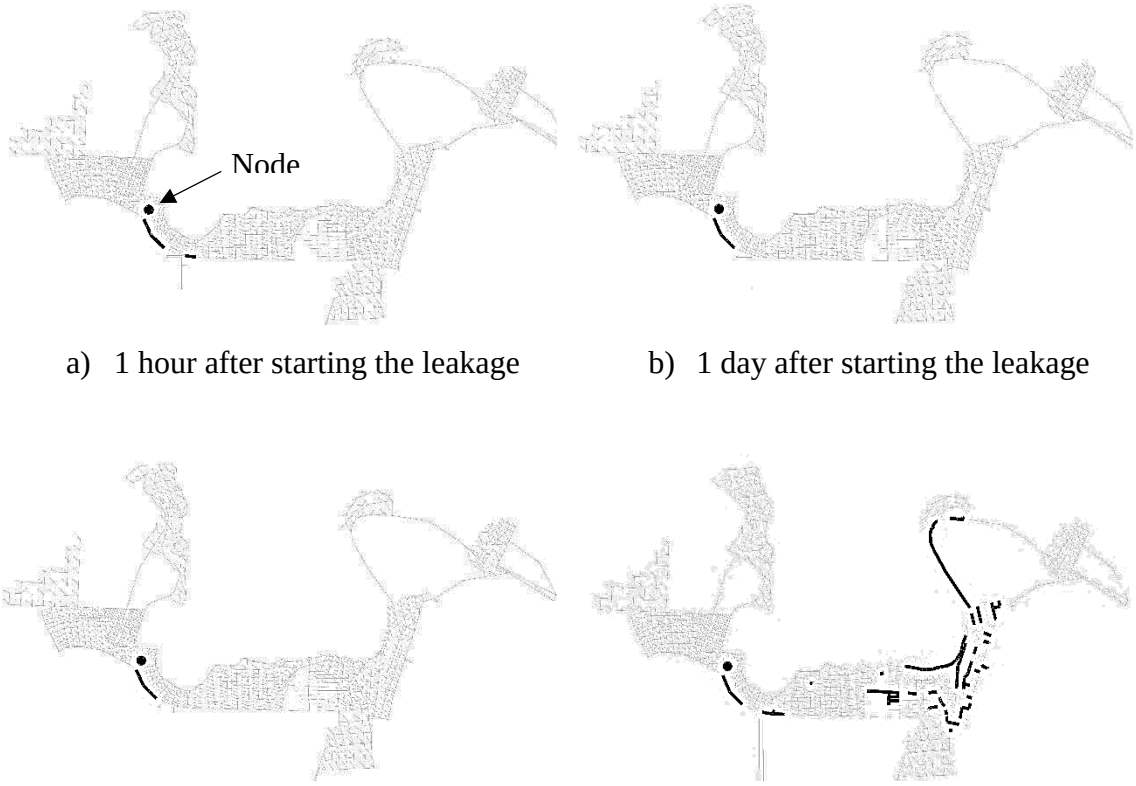
c) 4 days after starting the leakage

d) 8 days starting the leakage

Fig. S5 The pipes* in the WDN of Mesopolis that show the flow divergence from the control state of more than 0.1 gpm due to the leakage at node 1** at four different times

*Black pipes show flow divergence from the control state that are more than 0.1 gpm due to the leakage at node 1

**The node 1 is presented with a big black circle



a) 1 hour after starting the leakage

b) 1 day after starting the leakage

c) 4 days after starting the leakage

d) 8 days starting the leakage

Fig. S6 The pipes* in the WDN of Mesopolis that show the flow divergence from the control state of more than 5 gpm due to the leakage at node 1** at four different times

*Black pipes show flow divergence from the control state that are more than 5 gpm due to the leakage at node 1

**The node 1 is presented with a big black circle

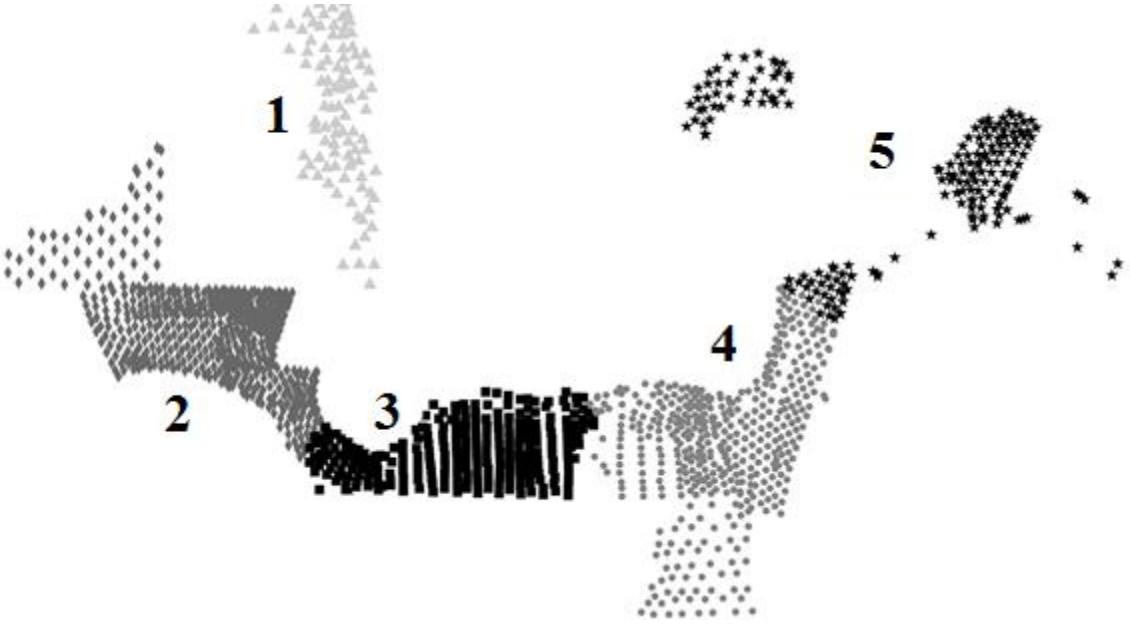


Fig. S7 Division of the study area into 5 zones using k-means clustering algorithm

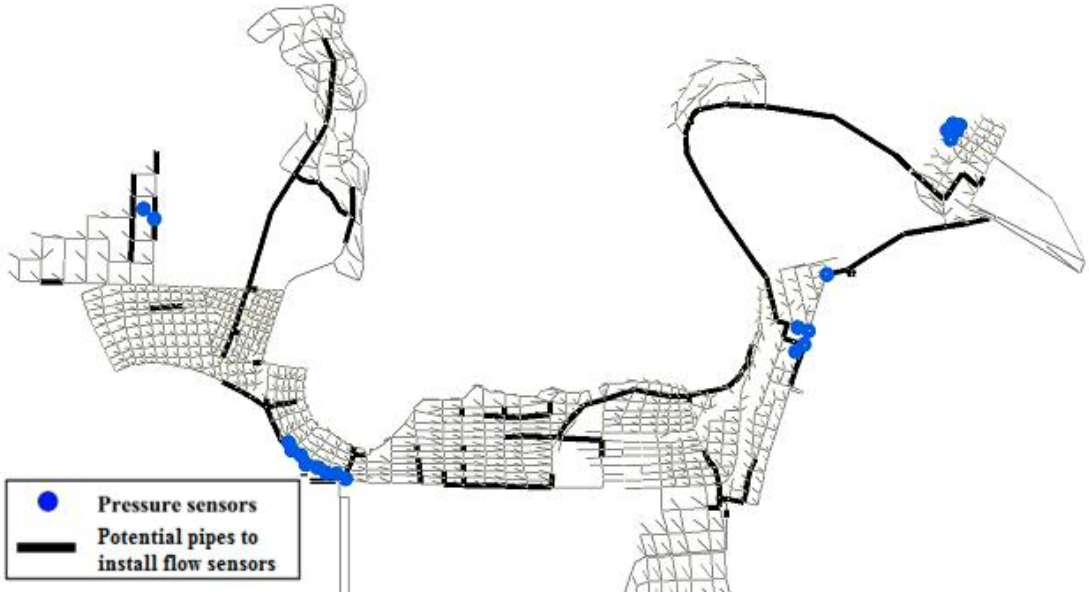


Fig. S8 Locations of potential flow and pressure sensors which are considered as decision variables in the NSGA-II multi-objective optimization algorithm

Fig. S9 Division of the study area into 10 zones using *k*-means clustering algorithm

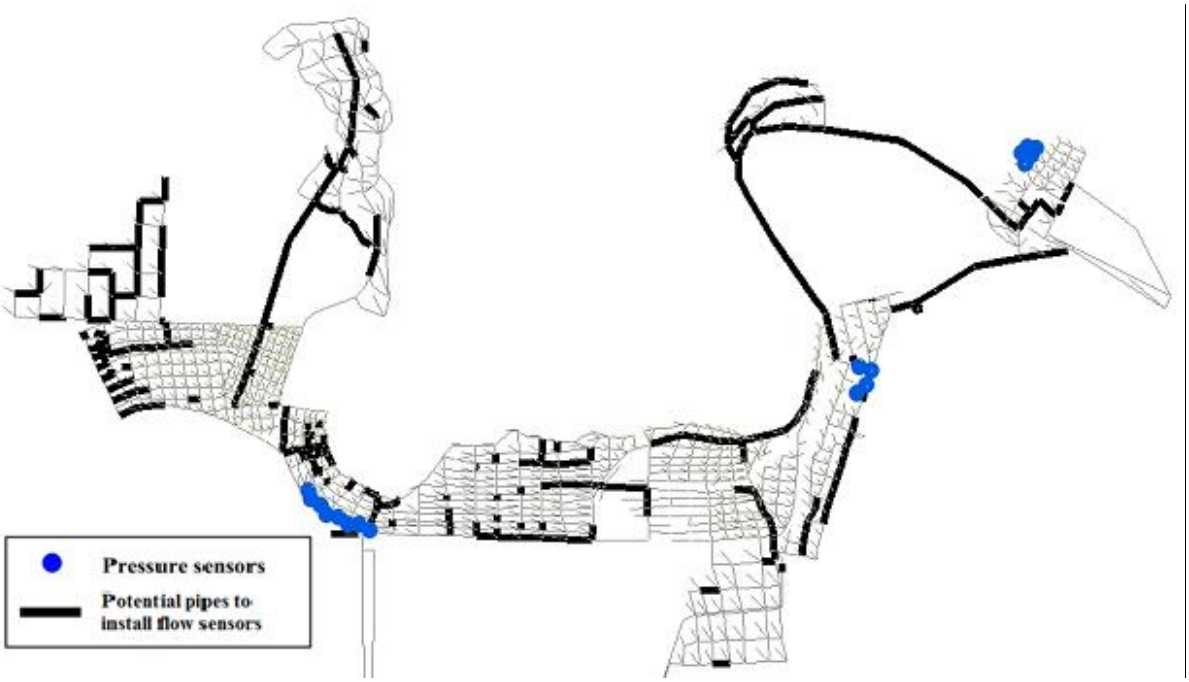


Fig. S10 Locations of potential flow and pressure sensors which are considered as decision variables in the NSGA-II optimization algorithm after dividing the study area into 10 zones

Fig. S11 Locations of optimized flow and pressure sensors in 10 zones in the WDN of Mesopolis

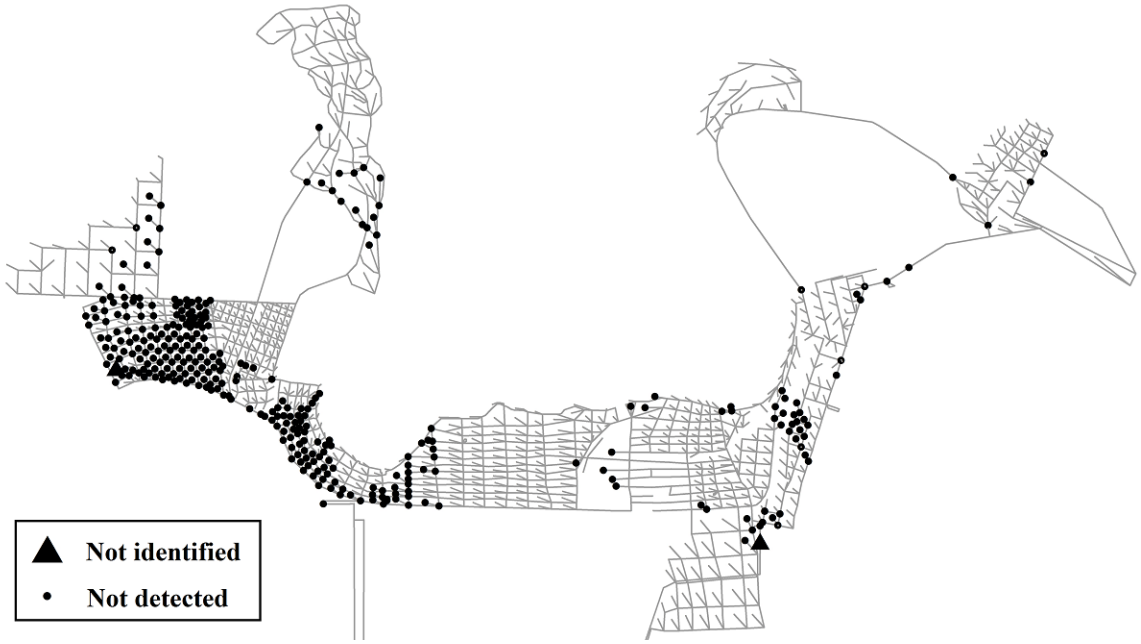


Fig. S12 Locations of the nodes where the leakage cannot be detected or the correct leakage zone cannot be identified by dividing the study area into 10 zones

K-means clustering

K-means clustering is an approach for clustering a number of observations into "k" zones, in which each observation is placed in the zone with the nearest mean. The k-means clustering algorithm consists of the following steps (Sharma et al. 2012):

1. Define k centroids in the data space, each centroid representing one cluster.
2. Place each observation in the zone with the nearest centroid according to Euclidean distance.
3. Recalculate k new centroids based on the points in each cluster.
4. Repeat Steps 2 and 3 until the locations of the centroids do not change any more.

K-means clustering algorithm has the following advantages over hierarchical clustering method (Sharma et al. 2012):

1. K-means clustering method may be faster than hierarchical clustering method in problems with a large number of variables.
2. K-means clustering method may result in tighter zones than hierarchical clustering method, especially when the zones are globular.

In this research, the following settings (as detailed in Table S1) are used:

Table S1. List of parameters and methods that are used in k-means clustering

Distance Measure	Sqeuclidean; $d(x,c)=(x-c)(x-c)'$
MaxIter	100
Replicates	5
Method for choosing initial cluster centroid positions	Plus; Select k seeds by implementing the k-means++ algorithm for cluster center initialization.

The main idea is to divide the space into a number of zones in which the nodes have at least a common feature and, most importantly, are close to each other (not scattered) in order to identify the leakage zone correctly. In Table S2 (Supplementary Information), total summation of distances among nodes of each zone are presented based on different combinations of feathers (latitude, longitude, elevation, and pressure) for dividing the WDN into different zones. As shown in Table S2 and Fig. S13-S16, using pressure or elevation in the combinations of features for clustering has no effect on the best value of total summation of distances among nodes. In other words, latitude and longitude are the two main features for clustering.

Table S2. Comparison of different combinations of feathers for k-means clustering

Features considered for k-means clustering	Replicate No.	Number of iterations	Total summation of distances among nodes	Best value of total summation of distances
Latitude and longitude, (Fig. S13)	1	21	1.85338e+11	1.85338e+11
	2	16	1.85338e+11	
	3	20	1.85338e+11	
	4	15	1.85338e+11	
	5	20	1.85338e+11	
Latitude, longitude, and elevation (Fig. S14)	1	14	1.85338e+11	1.85338e+11
	2	35	1.85338e+11	
	3	19	1.85338e+11	
	4	11	1.85338e+11	
	5	19	1.85338e+11	
Latitude, longitude, and pressure* (Fig. S15)	1	24	1.85333e+11	1.85338e+11
	2	12	2.24165e+11	
	3	15	1.85333e+11	
	4	14	2.19915e+11	
	5	19	2.24165e+11	
Latitude, longitude, elevation, and pressure*, (Fig. S16)	1	9	1.85338e+11	1.85338e+11
	2	13	2.28641e+11	
	3	41	1.85338e+11	
	4	13	1.85338e+11	
	5	7	1.85338e+11	

*Pressure at start time (t=0)

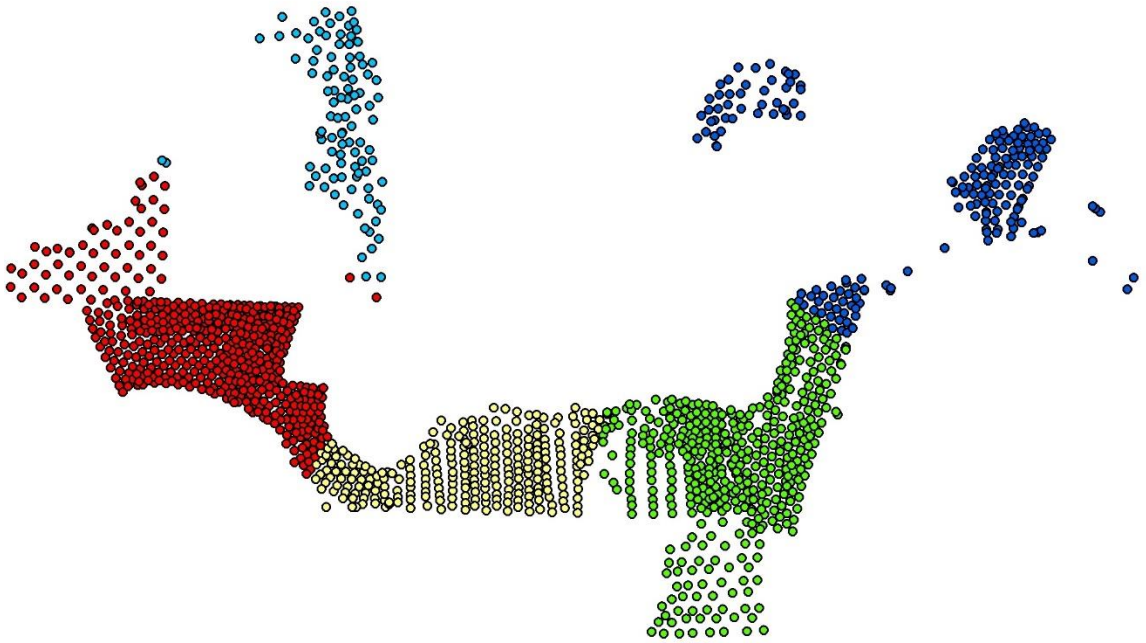


Fig. S13 Division of the study area into 5 zones using k-means clustering algorithm based on latitude and longitude

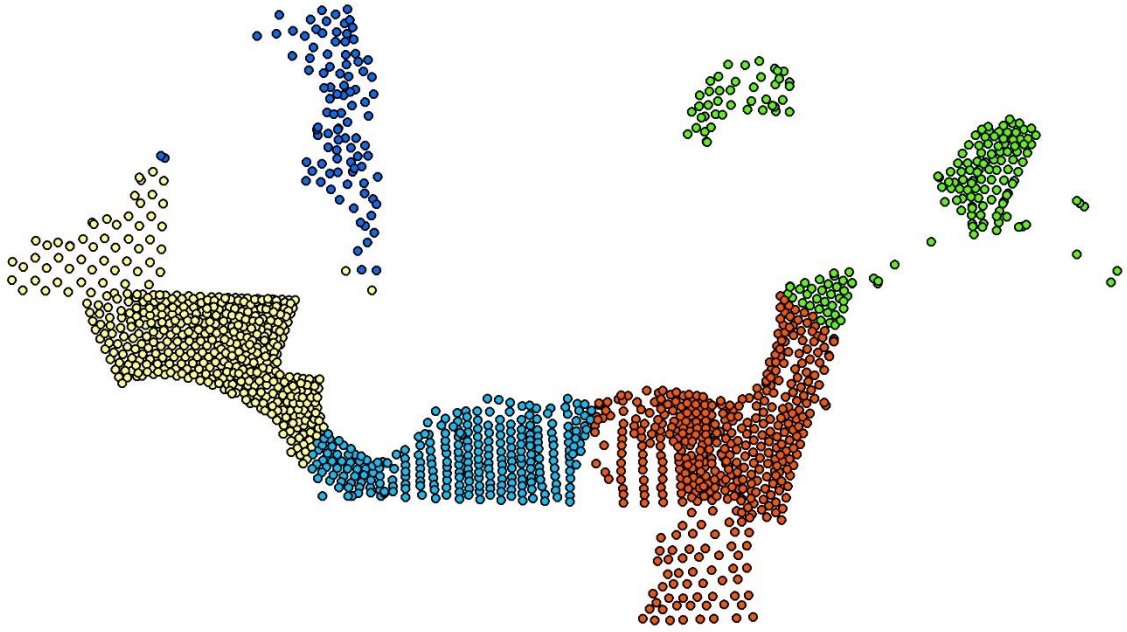


Fig. S14 Division of the study area into 5 zones using k-means clustering algorithm based on latitude, longitude, and elevation

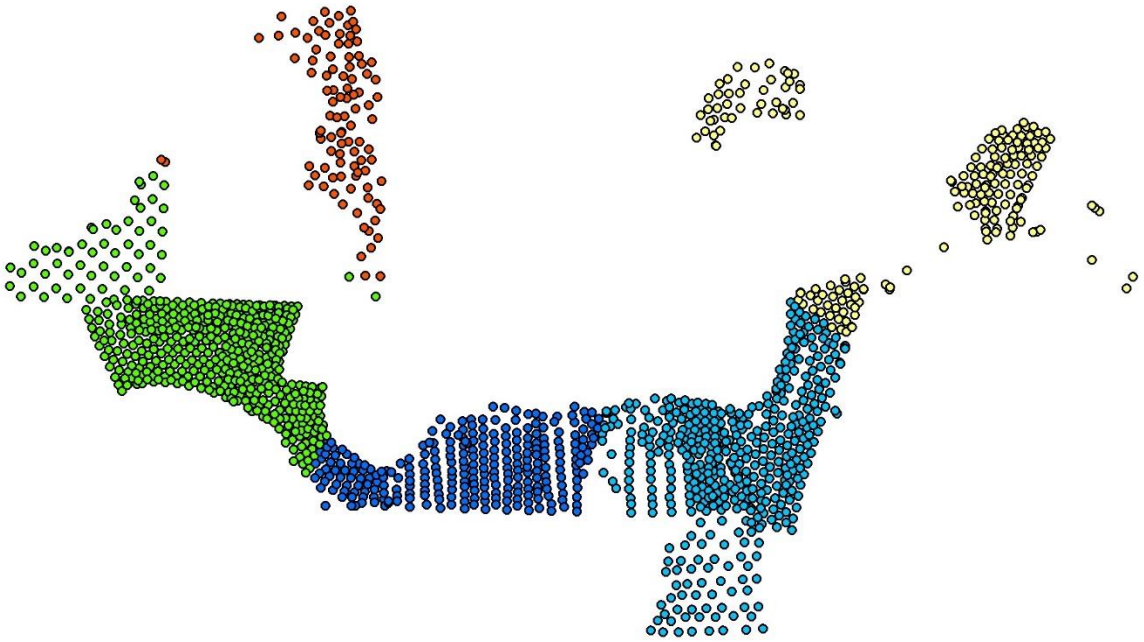


Fig. S15 Division of the study area into 5 zones using k-means clustering algorithm based on latitude, longitude, and pressure

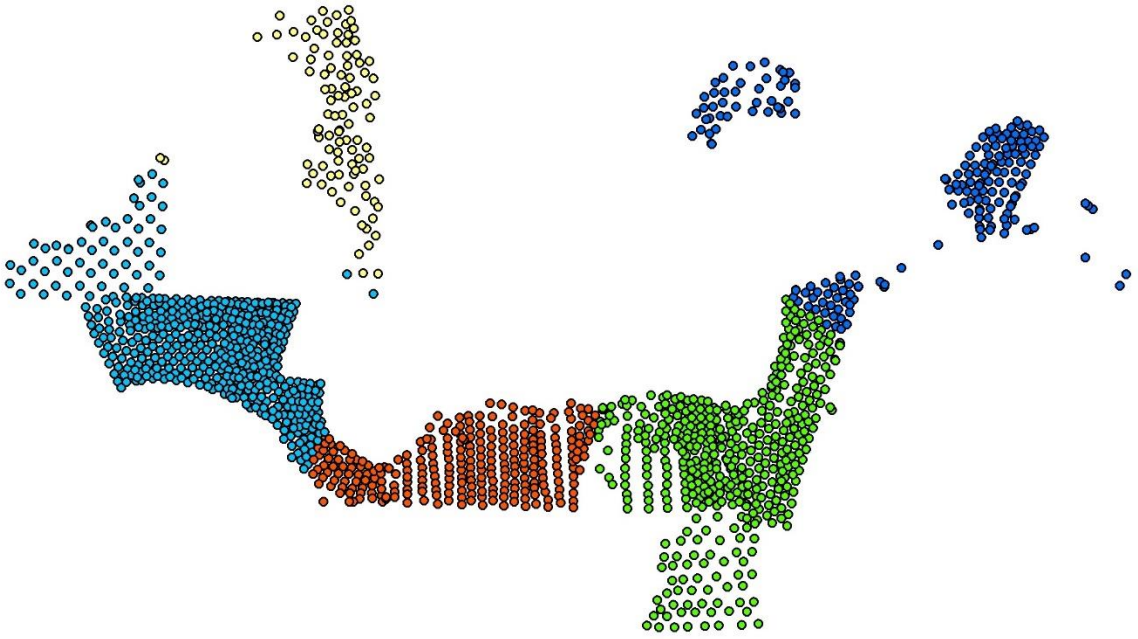


Fig. S16 Division of the study area into 5 zones using k-means clustering algorithm based on latitude, longitude, elevation, and pressure

Hierarchical clustering

The results of hierarchical clustering method based on the above features (latitude, longitude, elevation, and pressure) are shown in Table S3 (Supplementary Information). Solutions with closer value of Cophenetic correlation coefficient, "c", to 1, have better ranks. As presented in Table S3, latitude and longitude are two main features that lead to the highest quality solution for clustering.

Table S3. Comparison of different combinations of feathers for hierarchical clustering

Features considered for hierarchical clustering	Cophenetic correlation coefficient (c)
Latitude and longitude (Fig. S17)	0.7801
Latitude, longitude, and elevation (Fig. S18)	0.7802
Latitude, longitude, and pressure (Fig. S19)	0.7708
Latitude, longitude, elevation, and pressure* (Fig. S20)	0.7707

*Pressure at start time (t=0)

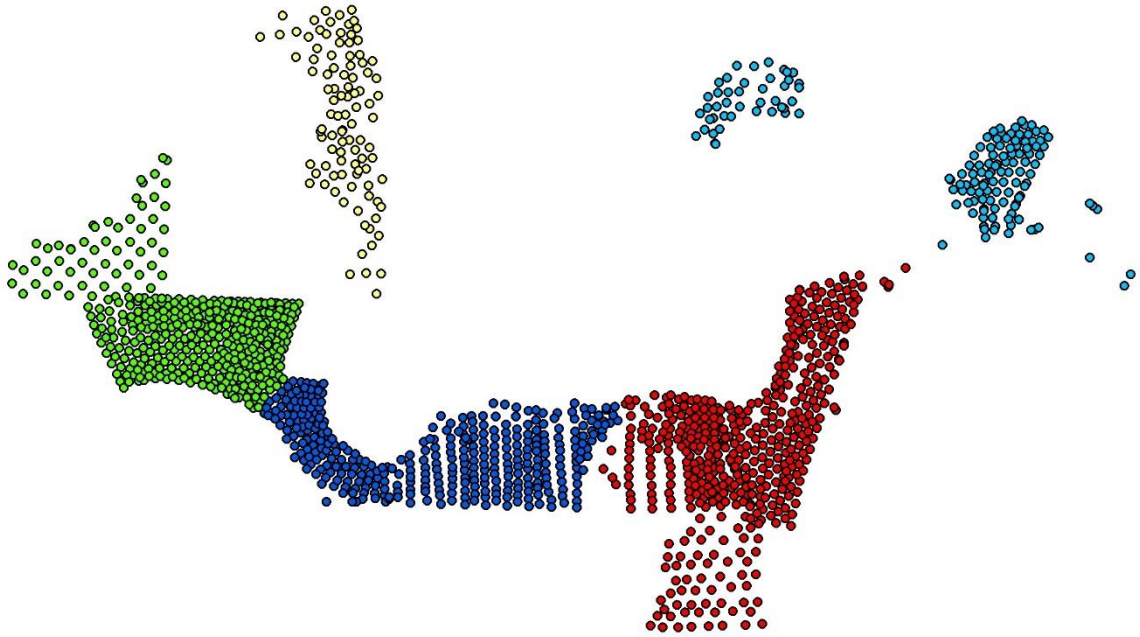


Fig. S17 Division of the study area into 5 zones using hierarchical clustering algorithm based on latitude and longitude

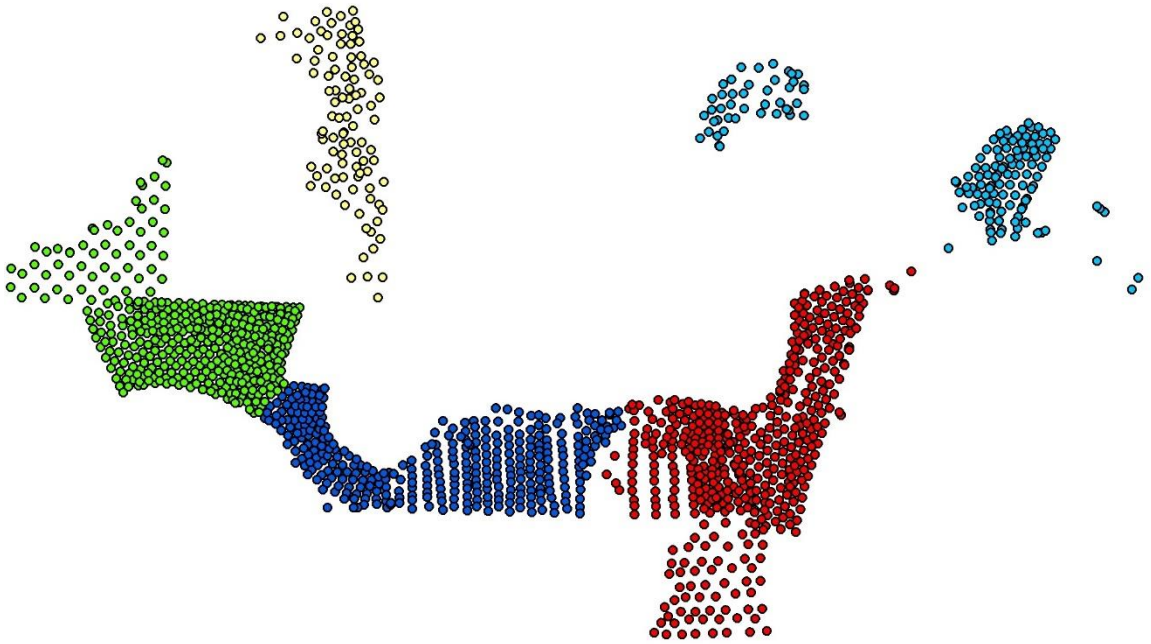


Fig. S18 Division of the study area into 5 zones using hierarchical clustering algorithm based on latitude, longitude, and elevation

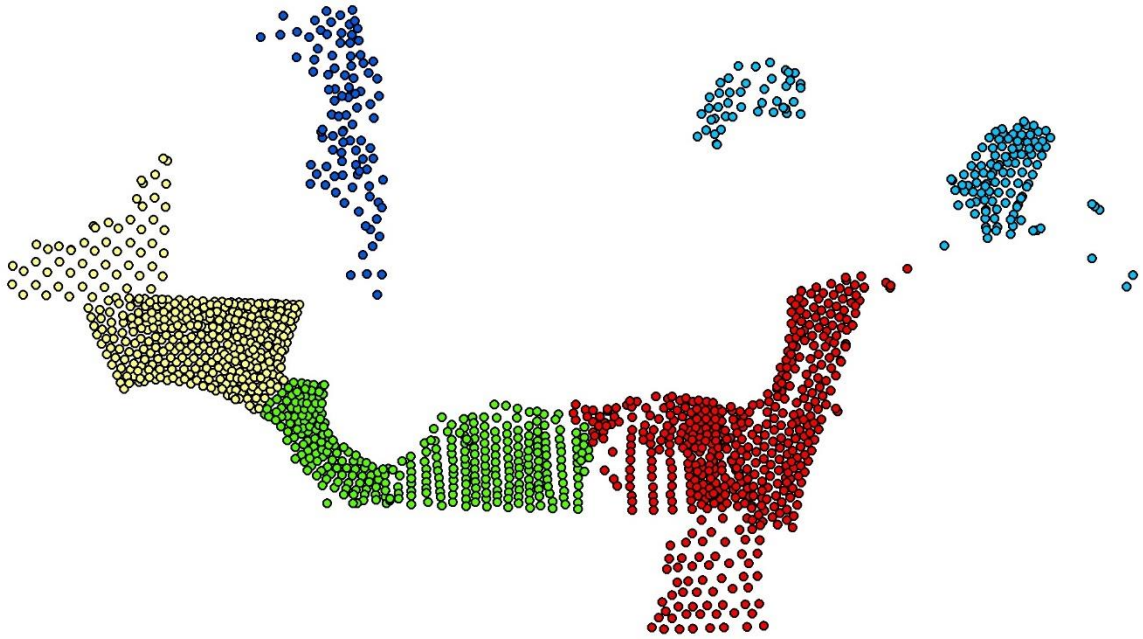


Fig. S19 Division of the study area into 5 zones using hierarchical clustering algorithm based on latitude, longitude, and pressure

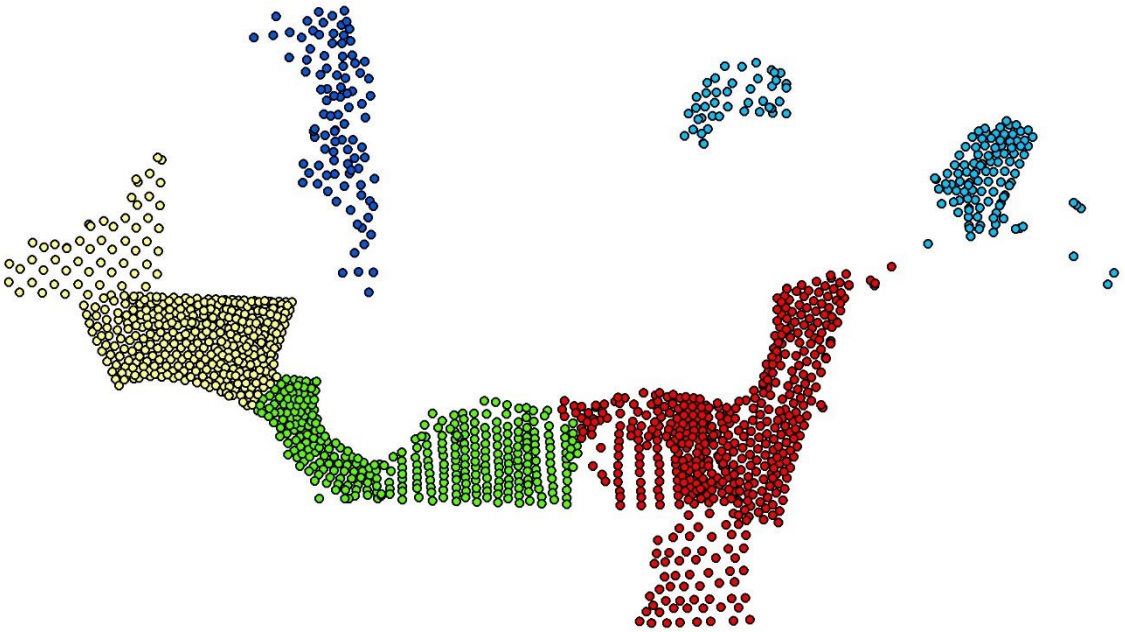


Fig. S20 Division of the study area into 5 zones using hierarchical clustering algorithm based on latitude, longitude, elevation, and pressure

Multi-objective optimization algorithm for C-town WDN

Here, we describe the multi-objective optimization used in this paper with a smaller network for simplicity. The step-by-step description of this algorithm can then be used for the much larger networks, as that of the main paper. Water distribution network of C-town includes 388 nodes, 444 pipes, 8 tanks, and 24 control valves. The duration of simulation is 96 hours. The WDN is divided into 5 zones. A set of potential pressure and flow sensors, as described in sections 2.3 and 2.4, are considered as decision variables in the NSGA-II multi-objective optimization algorithm. Fig. S21 shows the location of all potential flow and pressure sensors.

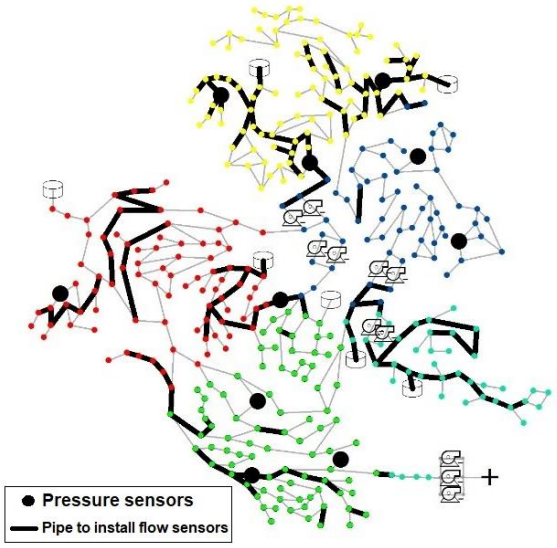


Fig. S21 C-town water distribution network and locations of potential flow and pressure sensors

The proposed multi-objective optimization algorithm is applied in order to select the optimal combination of pressure and flow sensors considering two objective functions:

1. Maximizing accuracy of identified leakage zone;
2. Minimizing number of sensors

At the first stage, the NSGA-II multi-objective optimization algorithm selects a number of flow sensors (K_f) among potential flow sensors (n_{sf}) for each chromosome. Initially, only flow sensors are optimized, while all potential pressure sensors are incorporated in the network.

Assume that a leakage has occurred in a single node (blue arrow in Fig. S22) and the multi-objective optimization algorithm has selected 26 flow sensors (K_f) among potential flow

sensors (n_{sf}) for one chromosome, while all of 10 potential pressure sensors are incorporated in the network.

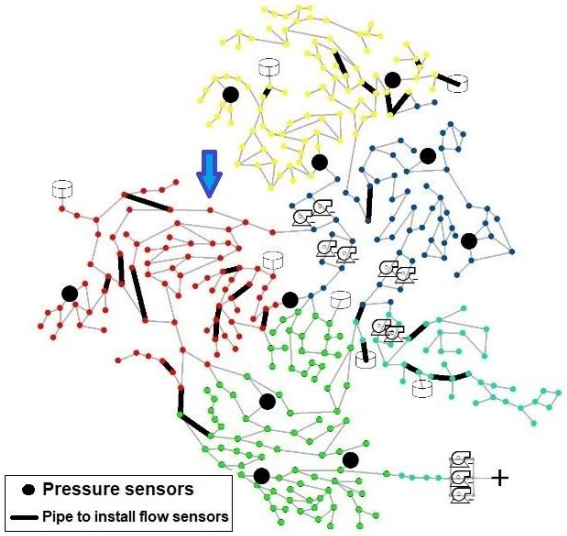


Fig. S22 Locations of selected flow sensors at the first stage for the above assumed chromosome, while all potential pressure sensors are incorporated in the network

There are 388 leakage scenarios corresponding to 388 nodes. As a result, there are 388 divergence matrices for flow and 388 divergence matrices for pressure sensors according to the 388 leakage models. For each leakage scenario, a matrix of size $\frac{T}{Time\ step} \times K$ is constructed for the selected sensors, where

T	97=96 hours+1, considering ($t = 0$)
$Time\ step$	1 hour
K	Number of sensors

As shown in Fig. S22, there are 10 pressure sensors and the NSGA-II multi-objective optimization algorithm selected 26 flow sensors among potential flow sensors. The flow and pressure divergence matrices are as follows (Table S4):

Table S4. Flow and pressure divergence matrices

Leakage scenario	1	2	3	…	388
Flow divergence matrices	97×26	97×26	97×26	…	97×26
Pressure divergence matrices	97×10	97×10	97×10	…	97×10

The optimization algorithm searches for flow divergence from the control state to be greater than the tolerance threshold in six successive time steps. Each flow sensor that satisfies this

condition is selected as an optimal one. In this stage, all potential pressure sensors with pressure divergence from the control state exceeding the tolerance threshold are also selected. This process is performed for all 388 leakage scenarios and as a result a matrix of size 388×5 is created. For the assumed leakage (blue arrow in Fig. S22) only 6 flow sensors out of 26 have responded to this leakage. These flow sensors and the corresponding zones are presented in Table S5.

Table S5. Flow sensors and the corresponding zones that have responded to the leakage shown by blue arrow in Fig. S22

Sensor	78	98	278	317	353	421
Zone	3	[1,5]	[1,3]	[1,3]	3	[1,3]

By dividing the total number of times that each leakage zone is identified to sum of the elements in each row, the probability of identification of each zone is determined for each leakage scenario (Table S6). Then, in each row, the zone(s) with identification probability greater than 80% of the maximum probability in the same row is/are selected.

Table S6. Probability of identification of each zone

Zone	1	2	3	4	5
probability	4/10	0/10	5/10	0/10	1/10

$0.8 \times 0.5 = 0.4 \Rightarrow$ zone 1 and zone 3 are selected

This can potentially select more than one zone for each node (each leakage scenario). In this case, pressure sensors are optimized to decrease the number of identified zones for each leaking node and increase accuracy of leakage zone identification. Therefore, the above optimization process of flow sensors is done for pressure sensors. Assume that two pressure sensors out of 10 have responded to the leakage in the above assumption. These pressure sensors and the corresponding zones are presented in Table S7.

Table S7. Two sensors out of 10 pressure sensors and the corresponding zones

Pressure sensor	158	282
Zone	3	5

Then identified leakage zone(s) by the flow sensors that is/are common with the identified leakage zone(s) by the pressure sensors is/are selected. If there is no common zone, the identified leakage zone(s) by flow sensors is/are chosen.

[1,3] ∩ [3,5] = [3] ⇒ zone 3 is selected

After determining the common identified zones by the combination of flow and pressure sensors for N nodes (N leakage models), each identified zone is compared with the corresponding actual defined zone for each node. If the identified zone and the actual defined zone for a node are identical, then the value of identification is assigned as "$d_n = 1$", otherwise "$d_n = -1$" (Eqs. 4 to 7). If two zones are identified for a node and one of them is correct, then the value of identification is "$d_n = 0.5$", otherwise "$d_n = -1$". If more than two zones are identified for a node, the value of identification is "$d_n = -1$". Finally, sum of all values of identification greater than zero is divided by N. The optimized flow and pressure sensors C-town WDN with identification probability greater than 80% for 5 zones are presented in Fig. S23.

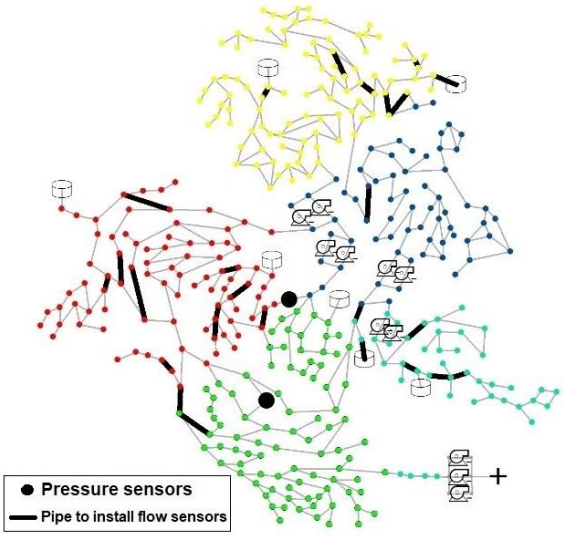


Fig. S23 Locations of optimized flow and pressure sensors for C-town WDN

Mesopolis WDN

The water distribution network of Mesopolis is a widely-used virtual WDN that is developed for research projects. Since real-world networks are not readily available due to security issues, a large number of studies have been conducted on this virtual network including Drake and Zechman (2012), Shafiee and Zechman (2013), and Rasekh and Brumbelow (2014 and 2015). We have modeled the Mesopolis WDN using the EPANET software that performs extended period simulation within pressurized pipe networks.

Demand Patterns

It is assumed that all nodes of the Mesopolis WDN follow 7 patterns of daily consumption and each pattern is the same for different days/seasons. As an example, two patterns are shown in Fig. S24.

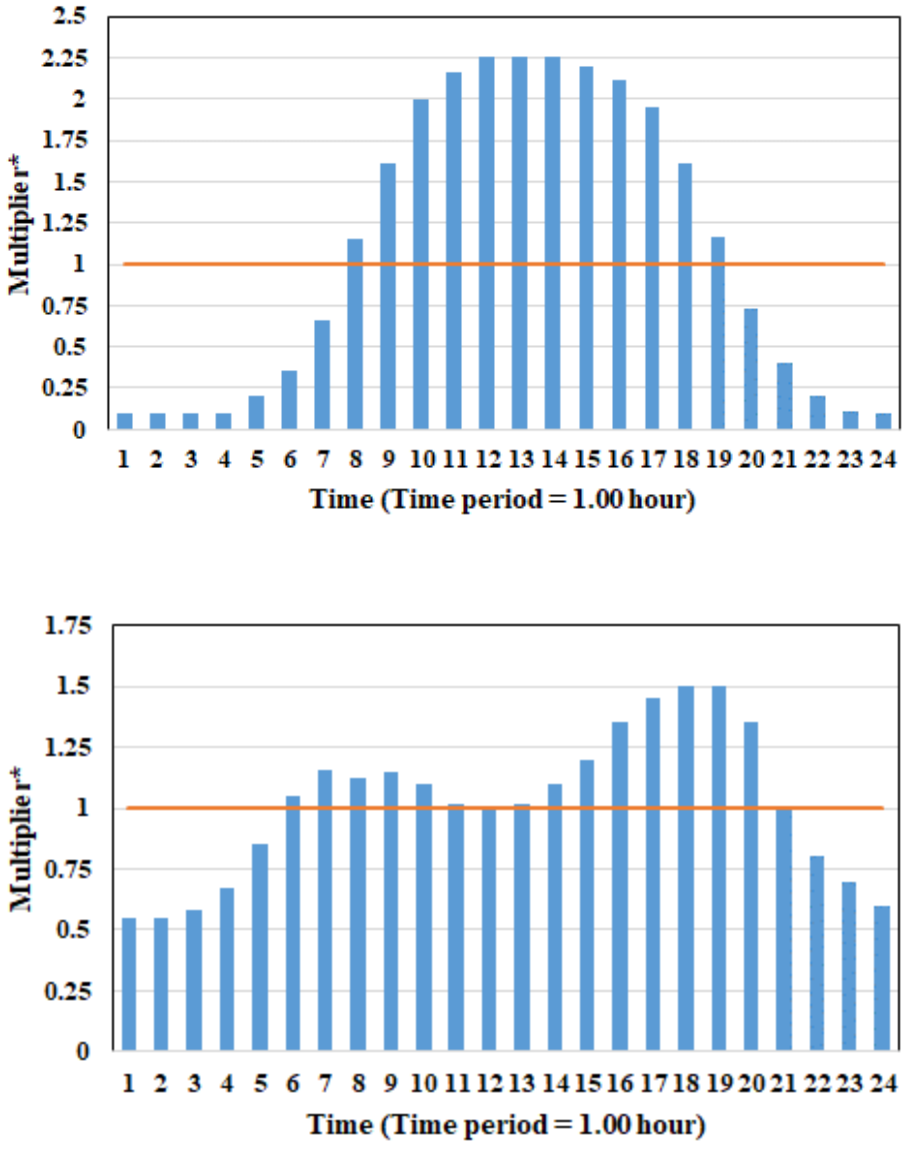


Fig. S24 Two patterns of daily consumption for nodes of the Mesopolis WDN
*The multipliers are used to modify the demand from its base level in each time period.

The pumps of Mesopolis WDN work based on 11 specified curves. As an example, two curves are shown in Fig. S25.

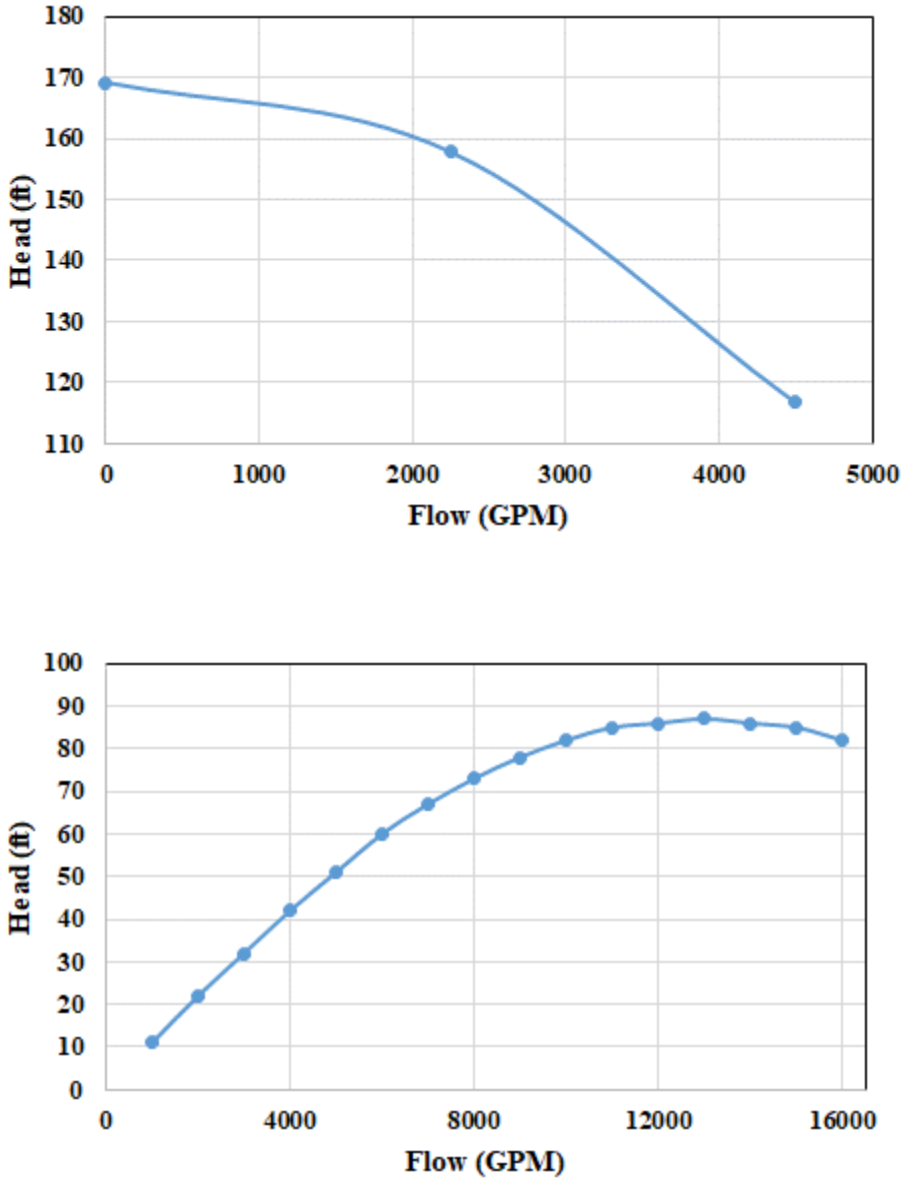


Fig. S25 Two samples of 11 pump curves of the Mesopolis WDN

Forty "If-Then" rules have been defined for the Mesopolis WDN to coordinate pumps, tanks, etc. with each other considering hours/days. As an example, the rule 2 is as follows (Fig. S26):

RULE 2

IF TANK wtpe-tank LEVEL BELOW 6

AND SYSTEM CLOCKTIME <= 8 AM

OR SYSTEM CLOCKTIME >= 8 PM

THEN PUMP intake1 STATUS IS OPEN

AND PUMP intake2 STATUS IS OPEN

AND PUMP intake3 STATUS IS OPEN

AND PUMP intake4 STATUS IS OPEN

AND PUMP intake5 STATUS IS OPEN

AND PUMP intake6 STATUS IS OPEN

AND PUMP intake7 STATUS IS OPEN

AND PUMP intake-B1 STATUS IS OPEN

AND PUMP intake-B2 STATUS IS OPEN

AND PUMP intake-B3 STATUS IS OPEN

AND PUMP intake-B4 STATUS IS OPEN

AND PUMP intake-B5 STATUS IS OPEN

ELSE PUMP intake1 STATUS IS CLOSED

AND PUMP intake2 STATUS IS CLOSED

AND PUMP intake3 STATUS IS CLOSED

AND PUMP intake4 STATUS IS CLOSED

AND PUMP intake5 STATUS IS CLOSED

AND PUMP intake6 STATUS IS CLOSED

AND PUMP intake7 STATUS IS CLOSED

AND PUMP intake-B1 STATUS IS CLOSED

AND PUMP intake-B2 STATUS IS CLOSED

AND PUMP intake-B3 STATUS IS CLOSED

AND PUMP intake-B4 STATUS IS CLOSED

AND PUMP intake-B5 STATUS IS CLOSED

Fig. S26 One of the "If-Then" rules for controlling the Mesopolis WDN

According to this rule, if the water level in the tank "wtpe" is less than 6 units and the time of simulation is between 8 pm and 8 am then the above mentioned pumps would be opened or closed.