Water Supply

Water Supply Vol 22 No 1, 1020 doi: 10.2166/ws.2021.219

Leakage localization using pressure sensors and spatial clustering in water distribution systems

Xin Li, Shipeng Chu, Tuqiao Zhang, Tingchao Yu and Yu Shao *
College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China
*Corresponding author. E-mail: shaoyu1979@zju.edu.cn

YS, 0000-0003-2435-5618

ABSTRACT

Leakages in water distribution systems (WDSs) are a worldwide problem, which can result in an intolerable burden in satisfying the water demands of the consumers. There is an urgent demand to develop technologies that can detect and localize the leakage in a timely and efficient manner. The monitoring data of the WDS is a typical time series, and there is a certain spatiotemporal correlation between the data provided by the devices distributed at different locations of the WDS. This paper proposes a novel model-based method for WDS leakage localization. The method is characterized by (1) developing the dominant sensor sequence for each candidate leakage node to improve the localization accuracy based on the spatial correlation analysis; (2) utilizing multiple time steps of the measurements which are temporal varying correlated; (3) ranking leakage regions and nodes by their possibility to contain the true leakage. A realistic WDS is used to evaluate the performance of the method. Results show that the method can accurately and efficiently localize the leakage.

Key words: leakage localization, model-based method, spatial clustering, water distribution system

HIGHLIGHTS

- A dominant sensor sequence is developed for each node based on the spatial correlation between multi-sensors to enrich the information.
- A strategy that uses the pressure data from multiple time steps is adopted to locate the leakage region.
- Give the detection order in the leakage region.
- Candidate node analysis in the detection region.
- Improve the accuracy and time efficiency of leakage localization.

INTRODUCTION

The water distribution system (WDS) is one of the most important infrastructures that deliver drinking water to various consumers. The safety and reliability of water distribution systems (WDSs) are crucial for cities (Duan *et al.* 2020). Leakages in WDS can damage the infrastructure, leading to an intolerable burden in a world struggling with satisfying the water demands of a growing population. In some cities, water losses caused by pipe leakages account for 30% of the total amount of drinking water in the WDSs (Puust *et al.* 2010). The water losses, as well as the cost of repairing the failed pipes, can result in significant economic costs. Locating and repairing leakages in a timely manner is extremely urgent to the water utility for economic, environmental, and reputational reasons.

Generally, leakage localization is realized by the use of advanced devices to monitor system behaviors. The acoustic equipment uses the acoustic device to localize the leakages by monitoring the abnormal behaviors at the potential leakage locations in the WDS (Wu *et al.* 2016; Zhou *et al.* 2019). However, being time-consuming is the main drawback for this method as detection of all potential leakage pipes is a heavy burden and usually takes a lot of time. This shortage prevents the acoustic equipment from being widely used in real problems. Alternatively, methods that utilize measured data, typically nodal pressure and pipe flow data provided by the sensor networks distributed in the WDS, are developed for leakage detection and localization. Compared with flow meters, pressure meters are easily installed and less expensive. The methods that use pressure data to locate the leakage can efficiently reduce the investments (Wu & Liu 2017). Therefore, pressure-based methods have come to be more and more popular for leakage detection and localization in WDS. The main mechanism

of the pressure-based methods is that the pressure under normal working conditions fluctuates in a certain range, while the leakage can lead to pressure drop and make the pressure fluctuate outside the normal range. Analysis of the deviation between the real-time pressure data and the normal ranges can efficiently detect and locate the leakages (Wu *et al.* 2018b; Zhou *et al.* 2019).

The state-of-the-art for leakage localization in WDSs is filled with contributions of different methods (Pérez *et al.* 2014a; Sanz *et al.* 2016; Sun *et al.* 2016; Zhang *et al.* 2016; Moser *et al.* 2018; Salguero *et al.* 2018; Wu *et al.* 2018a; Manzi *et al.* 2019; Sun *et al.* 2019). Typically, these methods can be classified as model-based methods, transient-based methods, and data-driven-based methods. The model-based method uses the hydraulic model to predict the normal range of nodal pressure and a comparison between the measured value and predicted value can be used to detect the leakages. For example, Farley *et al.* (2013) proposed a model-based method in which a sensitive matrix of different pressure measurements is adopted to quantify the relationship between the leakage rate and pressure fluctuation. Moser *et al.* (2018) used an explicit representation of the uncertainty distribution of modeling and measurement at each location. The threshold limit for falsifying model instances in error domain modeling is calculated to determine the candidate leak nodes. In addition, the pipe materials would affect the applicability and accuracy/efficiency of the model-based method, since previous studies (Duan *et al.* 2010, 2012) show that the plastic pipe may give very different responses (pressure and energy) to the system operation (steady or unsteady flows and also noises). Transient-based methods determine the location of leakage through time-domain or frequency-domain analysis of pressure signal observations (Duan *et al.* 2020). Meniconi *et al.* (2021) confirm the potential of Transient Test-Based Techniques (TTBTs) for fault detection in a long transmission main of real WDS. According to the characteristics of the transient tests and the investigated system, only a part of the system can be explored properly from a given generation point. The proposed test procedure allows overcoming the negative effect of a change in the initial and boundary conditions. Huang *et al.* (2020) propose a multistage method for burst leak localization through valve operations (VOs) and smart demand metering in district meter areas (DMAs) of WDSs. Each stage includes partitioning of the DMA into two subregions using VOs and identification of potentially leaking pipes within the subregions through water balance analysis based on smart demand meters. Such a process is performed repeatedly (multiple stages) to narrow down the spatial range for pinpointing leak locations. In recent years, data-driven-based methods that focus on the knowledge mined from pressure data have prevailed (Soldevila *et al.* 2016; Wu & Liu 2017; Soldevila *et al.* 2018; Abokifa *et al.* 2019; Sun *et al.* 2019; Zhou *et al.* 2019). Soldevila *et al.* (2016) proposed a method that combines the hydraulic model and data-driven model to locate the leakages. In this method, a classifier is applied to the residuals to determine the leakage localization. The classifier is trained with data generated by simulation of the WDS under different leakage scenarios and uncertainty conditions. More recently, Soldevila *et al.* (2018) presented a data-driven method using the Kriging spatial (Kleijnen 2017) to interpolate the pressure in nodes without sensor information.

Despite the above methods having been widely reported in many technical papers, their applications are limited. The main challenge is that the performance of these methods heavily depends on data quality. The pressure data logged by the sensors distributed over the network are inevitably polluted by the background noises, resulting in data quality degradation. Besides, due to intermittent sensor failures, bandwidth constraint, packet size limitation, or limited battery energy, data missing and outliers come to be a common feature of the sensor network. These uncertainties severely limit the practical application of the above methods. There is an impetus to develop a more robust algorithm that can detect and locate the leakage in the presence of these uncertainties.

The measured pressure data is a typical time series, and there is a certain spatiotemporal correlation between the measurements provided by the sensors distributed at the WDS. Assuming that leakage occurs at a node, multiple pressure sensors will respond at the same time, leading to synchronous pressure changes. Thus the pressure data from different sensors are generally spatially correlated. Considering that pipe leakage is not immediately repaired, the sensor response will continue for some time. In this situation, the time series of a given sensor is auto-correlated. Using the spatiotemporal correlation of multiple sensors can significantly reduce the uncertainty of a single sensor and thus improve the robustness of the algorithm. Besides, such spatiotemporal correlations are typically determined by the time of leakage occurrence, leakage rate, and leakage location. This is because the sensor location, pipe topology, and leakage localization can greatly affect the response of sensors and thus cause different spatiotemporal correlations in the measurements. A method that can improve the accuracy of leakage localization is to extract the spatiotemporal correlation information of measurements for leakage localization.

This paper proposes a novel model-based method for leakage localization in water distribution systems. The main contributions are: (1) a dominant sensor sequence is developed for each node based on the spatial correlation between multiple

sensors to enrich the information from the sensors; (2) a strategy that uses the pressure data from multiple time steps is adopted to locate the leakage region; (3) an approach that gives the priority detection order in the leakage region is developed to improve the efficiency of the leakage localization. The paper is expected to improve the accuracy and time efficiency of leakage localization in WDS.

The rest of the paper is organized as follows. The Methodology section describes the principle of this methodology in detail. In the Results and Discussion section, a case study is presented using the WDS of J City of China to evaluate the performance of the leakage localization method, and a discussion of the relevant results is given. The Conclusions section draws the main conclusions of the paper and introduces some potential extensions.

METHODOLOGY

The framework of the leakage localization methodology

Figure 1 shows the framework of the model-based leakage localization method. As shown in Figure 1, this method consists of four parts, namely, leakage detection, leakage scenario simulation, indicator calculation, and leakage localization analysis. The leakage detection algorithm (Shao *et al.* 2019) is used to determine whether the leakage occurs. If the detection alarm is not triggered, the network status is classified as a no-fault state. Conversely, if the detection alarm is triggered, the leakage should be located in the following time steps. Since the leakage detection algorithm has been investigated by (Shao *et al.* 2019), we focus on the leakage localization once the detection alarm is triggered (the remaining three parts).

Different from the existing approaches that locate the leakage immediately once the detection alarm is triggered, the developed approach locates the leakage only after the leakage has persisted for some time steps (DT). The reason is that the use of the pressure data from multiple time steps can utilize the temporal varying correlation of the measurements and thus efficiently improve the localization accuracy. The leakage localization process consists of three parts: leakage scenario simulation, indicator calculation, and leakage localization analysis. In the leakage scenario simulation, the concept of a dominant sensor sequence is developed based on sensitivity analysis. The dominant sensor sequence utilizes the spatial correlation between sensor and leakage and only the sensors that are highly correlated to the leakages are used for the leakage localization. Then a leakage scenario simulation is adopted to simulate the leakage at different nodes with different intensities. By calculating the similarity between the real-time measurements and the simulated leakage scenarios, the leakage will be located in a certain region. To achieve a comprehensive evaluation of the similarity, seven different metrics are used to evaluate the similarity between real-time measurement and simulated scenarios. Finally, the leakage localization analysis to determine the operation priority of the candidate leakage region and nodes is proposed to help the operator to inspect the

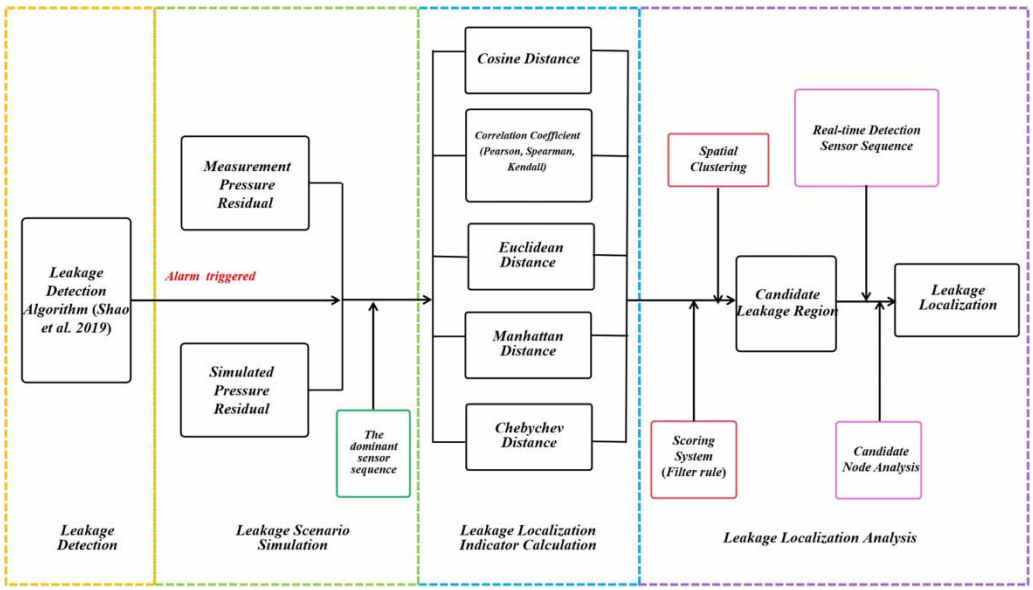

Figure 1 | The framework of the leakage localization methodology.

actual leakage location in the field. This paper has the following assumptions: (i) there is only one leakage that appears at one time, and (ii) pressure sensors in the WDS are placed in the selected nodes and are working well.

The dominant sensor sequence

In previous studies (Perez *et al.* 2014b; Kang *et al.* 2018; Shao *et al.* 2019), the measurements from all the sensors are used for leakage detection and localization. However, the sensors that are far from the leakage will only have a small pressure drop, which may be much smaller than the pressure fluctuation caused by the measurement uncertainties (noise, outliers, etc.). In this situation, the measurements used for detection and localization analysis are polluted by the measurement uncertainties, leading to poor robustness of the leakage detection and localization. To address this problem, Qi *et al.* (2018) have established the coverage region of a single sensor to detect and locate the leakage. This method reduces the adverse effects of insensitive sensors to a leakage. However, only a single sensor is used for localization, resulting in low localization accuracy in their study.

To compensate for the above shortcomings, the concept of the dominant sensor sequence is developed. The dominant sensor sequence is only a part of all the sensors. Each leakage node corresponds to a specific number of sensors, which must be sensitive to the leakage node. The generation of the dominant sensor sequence consists of two steps: sensitivity analysis, and sequence reconstruction.

The sensitivity matrix represents the sensitivity relationship between nodal demands and nodal pressures, which measures the degree of the effect of demand variation at one node on pressure variation at another node. In this paper, the sensor sensitivity matrix is obtained by WDS model hydraulic simulation (Perez *et al.* 2011; Blesa *et al.* 2012). A detailed process for the calculation of sensor sensitivity matrix ($\boldsymbol{S}^t$) at a given time step t can be found at (Blesa *et al.* 2014; Blesa *et al.* 2015; Steffelbauer & Fuchs-Hanusch 2016). As mentioned previously, the developed approach locates the leakage only after the leakage has persisted for DT time steps. Therefore, the sensor weight matrix $\bar{S}$ is constructed by averaging $\mathbf{S}^t$ at DT time steps, as shown in Equation (2).

$$S^t = \begin{bmatrix} \frac{\partial H_{S_1}}{\partial Q_{N_1}} & \cdots & \frac{\partial H_{S_1}}{\partial Q_{N_k}} & \cdots & \frac{\partial H_{S_1}}{\partial Q_{N_{nn}}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial H_{S_k}}{\partial Q_{N_1}} & \cdots & \frac{\partial H_{S_k}}{\partial Q_{N_k}} & \cdots & \frac{\partial H_{S_k}}{\partial Q_{N_{nn}}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial H_{S_{ns}}}{\partial Q_{N_1}} & \cdots & \frac{\partial H_{S_{ns}}}{\partial Q_{N_k}} & \cdots & \frac{\partial H_{S_{ns}}}{\partial Q_{N_{nn}}} \end{bmatrix} \tag{1}$$

$$\bar{S} = \frac{\sum_{t=1}^{DT} S^t}{DT} \tag{2}$$

where $S^t \in R^{n_s \times n_n}$ is the sensor sensitivity matrix at time step t; H_{S_k} and Q_{N_k} are the nodal pressure and water demand at the sensor S_k and node N_k, $\frac{\partial H_{S_k}}{\partial Q_{N_k}}$ indicates the effect of pressure change on the sensor S_k due to leakage at node N_k; $\bar{S}$ is the sensor weight matrix; n_s is the number of sensors and n_n is the number of nodes.

The number of elements in the original sensor sequence for each node is equal to the total number of pressure sensors arranged in WDS. Sequence reconstruction is based on sensor sensitivity weight matrix $\bar{S}$. The k^{th} column of the $\bar{S}$ matrix can be used to quantify the sensitivity of S_k sensors to the nodal demand at N_k node. The dominant sensors for the nodal demand at N_k node are selected based on the sensitivity values, corresponding to N_{sd} elements with big value in the k^{th} column of the $\bar{S}$ matrix. The original sensor sequence is reconstructed into the dominant sensor sequence. The number of dominant sensors N_{sd} ($N_{sd} \leq n_s$) is a hyper-parameter that can be determined by the engineering experience.

Leakage scenario simulation

Water demand prediction can be found in many studies (Kang & Lansey 2009; Arandia *et al.* 2016; Xie *et al.* 2017). In this paper, a prediction function is adopted to predict the nodal water demand at the current time step t from the historical

demand information.

$$x_t = f(x_{t-1,\,t-2,\ldots}) \tag{3}$$

where x_t is the nodal water demand vector at the current time step t; $f(\ )$ is the nodal water demand prediction function; $x_{t-1,\,t-2,\ldots}$ is the historical demand information.

Then the predicted nodal water demand at the current time step is used as input of the WDS hydraulic model to get the predicted pressure vector $\hat{\boldsymbol{p}}_f(t)$.

$$\hat{\boldsymbol{p}}_f(t) = [\hat{\boldsymbol{p}}_{f1}(t), \ldots \hat{\boldsymbol{p}}_{fi}(t), \ldots \hat{\boldsymbol{p}}_{fn_s}(t)]^T \tag{4}$$

$$\hat{\boldsymbol{p}}_{fi}(t) = \boldsymbol{g}_i(x_t) \tag{5}$$

where $\hat{\boldsymbol{p}}_{fi}(t)$ is the predicted pressure value of the i^{th} sensor at the current time step; g_i is the output WDS hydraulic model corresponding to the i^{th} sensor; and n_s is the number of pressure sensors in the network.

Equation (5) gives the predicted pressure value ($\hat{\boldsymbol{p}}_f(t)$) under the normal work condition. Comparing $\hat{\boldsymbol{p}}_f(t)$ with the pressure under the leakage scenarios can help to detect the leakages. Here, the leakage scenario is simulated by adding an extra demand to the predicted normal nodal water demand. For a given node, the extra demands with different values are added to the predicted demand at the node. After the hydraulic simulation, the pressures at the leakage scenarios are acquired. The above is processed node by node. Then the leakage scenarios at nodes with different leakage intensities are generated.

$$\hat{\boldsymbol{p}}_k(t) = [\hat{\boldsymbol{p}}_{k1}(t), \ldots \hat{\boldsymbol{p}}_{ki}(t), \ldots \hat{\boldsymbol{p}}_{kn_s}(t)]^T, k = 1, 2, \ldots, \mathrm{n} \tag{6}$$

where $\hat{\boldsymbol{p}}_{ki}(t)$ is the estimated pressure value of the i^{th} sensor for the k^{th} leakage scenario; n is the total number of simulated leakage scenarios.

The pressure residual vector $\hat{r}_k(t)$ for the k^{th} leakage scenario can be obtained by subtracting $\hat{\boldsymbol{p}}(t)$ from the predicted pressure vector $\hat{\boldsymbol{p}}_f(t)$.

$$\hat{\boldsymbol{r}}_k(t) = \hat{\boldsymbol{p}}_k(t) - \hat{\boldsymbol{p}}_f(t), k = 1, 2, \ldots, \mathrm{n} \tag{7}$$

At time step t, the n_s pressure sensors will upload a set of measured nodal pressure data, which constitute a measured pressure vector (Equation (8)).

$$\boldsymbol{p}(t) = [\boldsymbol{p}_1(t), \ldots \boldsymbol{p}_i(t), \ldots \boldsymbol{p}_{n_s}(t)]^{\mathrm{T}} \tag{8}$$

Then the measurement pressure residual vector $\mathrm{r}(\mathrm{t})$ can be calculated by Equation (9).

$$\boldsymbol{r}(t) = \boldsymbol{p}(t) - \hat{\boldsymbol{p}}_f(t) \tag{9}$$

Candidate leakage region ranking

Leakage localization is based on the analysis of simulated and measured pressure residuals ($\hat{\boldsymbol{r}}_k$ and $\boldsymbol{r}$). The leakage scenario, of which the simulated residual $\hat{\boldsymbol{r}}_k$ is most similar to the measured residual $\boldsymbol{r}$, is considered to be the representation of the actual leakage. Therefore, the location and leakage intensity value of the scenario is treated as the actual leakage location and intensity value. Some metrics are used to measure the similarity between $\hat{\boldsymbol{r}}_k$ and $\boldsymbol{r}(t)$, namely Manhattan distance, Euclidean distance, Chebyshev distance coefficient, cosine similarity, Pearson correlation coefficient, Spearman rank correlation coefficient, Kendall τ correlation coefficient (Perez *et al.* 2014b; Ponce *et al.* 2014). These metrics assess the similarity of the two vectors from different perspectives. For example, Manhattan distance is sensitive to changes in the average value, whereas the Pearson correlation coefficient considers the degree of linear correlation. Therefore, the above seven metrics

are combined to obtain more reliable localization results. Taking the Pearson correlation coefficient as an example, we explain below.

As mentioned previously, only the elements corresponding to the dominant sensors in the vector $\hat{\boldsymbol{r}}_k$ and $\boldsymbol{r}$ are adopted in the leakage localization process. Denoting the two residuals of the dominant sensors as $\hat{\boldsymbol{r}}'_k$ and $\boldsymbol{r}'$, the Pearson correlation coefficient has the following form,

$$C_t^k = \frac{\operatorname{cov}(\hat{\boldsymbol{r}}'_k, \boldsymbol{r}')}{\sqrt{\operatorname{cov}(\hat{\boldsymbol{r}}'_k, \hat{\boldsymbol{r}}'_k)\operatorname{cov}(\boldsymbol{r}', \boldsymbol{r}')}}, k = 1, 2, \ldots, n \tag{10}$$

where $\operatorname{cov}(\hat{\boldsymbol{r}}'_k, \boldsymbol{r}')$ is the covariance between $\hat{\boldsymbol{r}}'_k$ and $\boldsymbol{r}'$. n is the total number of leakage scenarios, C_k^t is the Pearson correlation coefficient that is used as an indicator for the k^{th} leakage scenario. Since the developed approach locates the leakage has persisted for DT time steps, the average correlation coefficient has the following form,

$$C_k = \frac{\sum_{t=1}^{DT} C_k^t}{DT}, k = 1, 2, \ldots, n \tag{11}$$

The seven metrics can be calculated similarly and the averaged indicators calculated, which are used to determine the most similar leakage scenario. Generally speaking, a higher correlation coefficient indicates that this scenario has a higher probability to be considered as the actual leakage. Each scenario is ranked and scored based on the correlation coefficient (C_k). The scenarios ranked in the top 5% are given a score of 2, while the scenarios ranked in 5%–20% are given a score of 1, and the rest of the scenarios are given a score of 0.

As mentioned previously, seven metrics are used as indicators for leakage localization. Therefore, there are 7 scores for each scenario, and the sum of the 7 scores is used as the total score of the scenario (SS). The scenarios of a higher score are treated as the candidate leakage scenarios. The scenarios in which the scores are higher than a score-threshold λ are selected as candidate leakage scenarios. A candidate leakage scenario corresponds to a leakage node and a leakage intensity. The above strategy allows that the scenarios with the same location but different intensities may be selected as candidate leakages. In this situation, the number of times (NS) that the scenario with the same location is selected is recorded.

Here, the locations (nodes) of the candidate scenarios constitute a candidate node set $\{C_{node}\}$. Based on the spatial distance, these nodes can be distributed into different candidate regions and each region has a clustering center, using the hierarchical clustering algorithm and K-means clustering algorithm (MacQueen 1967; MacKay 2004; Sarrate *et al.* 2014). Previous studies using clustering for leakage localization have classified the nodes first and then matched the localization to regions (Soldevila *et al.* 2016; Zhang *et al.* 2016). This method first generates the set of candidate nodes and then characterizes the probability that the candidate regions contain the actual leakage by classification and indicator. In the case of noise in pressure measurements, the localization performance is significantly improved. This is because the clustering can handle more efficiently the dispersion produced by noise in measurements. The same occurs when demand uncertainty is considered (Soldevila *et al.* 2016). As shown in Figure 2, the candidate nodes set consists of 10 nodes and these nodes are distributed to three regions based on the clustering algorithm. Three candidate regions and three candidate region centers are formed by spatial clustering.

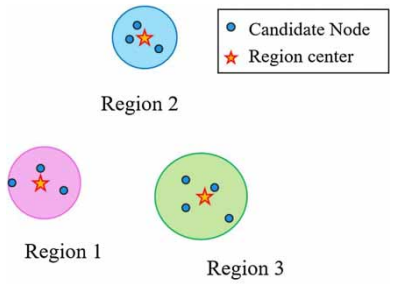

Figure 2 | Spatial Clustering.

Update to now, the leakages are probably located in the candidate regions. The acoustic methods can be used to precisely locate the leakage in the candidate regions. A key problem is to determine the order of leakage detection for these candidate regions. For a WDS network, the pressure drop will be large if the sensors are closer to the leakage location. The average pressure drop of all sensors can be calculated by Equation (12).

$$\bar{r} = \frac{\sum_{t=1}^{DT} r(t)}{DT} \tag{12}$$

The sensors are sorted according to the elements in the vector $\bar{r}$, the top n_g sensors of which, are chosen to determine the detection order of the candidate regions. The average distance from the regional center which is formed by spatial clustering to these sensors can be used to rank the order of the candidate regions.

Candidate nodes ranking

The previous section gives the detection order of the candidate region, and the next step is to determine the detection order of the nodes in the region to be detected. Assuming that the candidate region consists of y candidate nodes. For every node, it may be related to several candidate leakage scenarios with the same leakage location and different leakage intensities. Each scenario has a score (SS) as mentioned previously. Therefore, the Sum score (Sn_i) for the node n_i can be the sum of these scores. Similarly, the repeated number (NS) of candidate leakage scenarios with a certain candidate node is counted, the Cumulative ratio (P_i) for nodes n_i is obtained by dividing the NS by the total number of candidate scenarios. The pattern of pressure fluctuations at a certain leakage node is similar regardless of leakage intensity, the higher the cumulative ratio is, the more likely the pressure fluctuation caused by actual leakage is matched to the certain node.

Based on the magnitude of the characteristic parameter (Sum score, Cumulative ratio), the nodes in the detection region are ranked. The higher the characteristic parameters of the node, the higher the probability that the node is a true leakage node.

Detection evaluation indicator

The performance of the developed method is evaluated by two indicators, namely, Geographical distance (Topological distance) and Pipe distance. The geographic distance intuitively shows the distance from the leakage candidate nodes or the candidate region center to the actual leakage node, which is directly related to the localization accuracy.

The pipeline distance between two nodes refers to the shortest hydraulic path of the WDS connecting the two nodes; that is, the minimum value of the sum of the lengths of the pipes connecting the two nodes. This distance can help to assess the use of acoustic methods that can locate precisely the leakage if it is within a determined pipe distance.

RESULTS AND DISCUSSION

WDS description

The method is applied to a realistic WDS hydraulic model with synthetic data. The network is located in a city in Zhejiang Province, China. It consists of 509 pipes, 491 nodes, and three water sources, as shown in Figure 3. A total of 20 pressure sensors are installed in the network. The model of this network is created using the software EPANET 2.0 (Rossman 2000).

Parameter settings

Modeling error and measurement error are two uncertainties considered in this paper. This study focuses on using the spatio-temporal correlation of multiple sensors, which is closely related to the measurement error. The measurements in the normal working condition (without leakage) are synthesized by adding random noise N (0, σ_p) to the real pressure. For comparison, two data sets with different precision are generated with the standard deviation (STD) $\sigma_p = 0.1$ m and 0.3 m respectively.

As mentioned in the methodology, by comparing the similarity between the real-time measurements and the values of the simulated scenarios, the leakage will be located in a certain region. For the leakage scenarios simulation, The leakage intensity ranges from 20 m^3/h to 350 m^3/h, with an increment of 1 m^3/h for intervals of 20 m^3/h to 50 m^3/h, and 5 m^3/h for intervals of 50 m^3/h to 350 m^3/h. It contains a total of 91 discrete values. The total number of simulated scenarios n is 44,681 (491×91).

The parameters are as follows: N_n is the total number of nodes in WDS ($N_n = 491$), n_s is the total number of pressure sensors ($n_s = 20$). The prediction time-domain covers nt ($nt = 24$, one hour step) consecutive time steps, and DT is set to 3. The

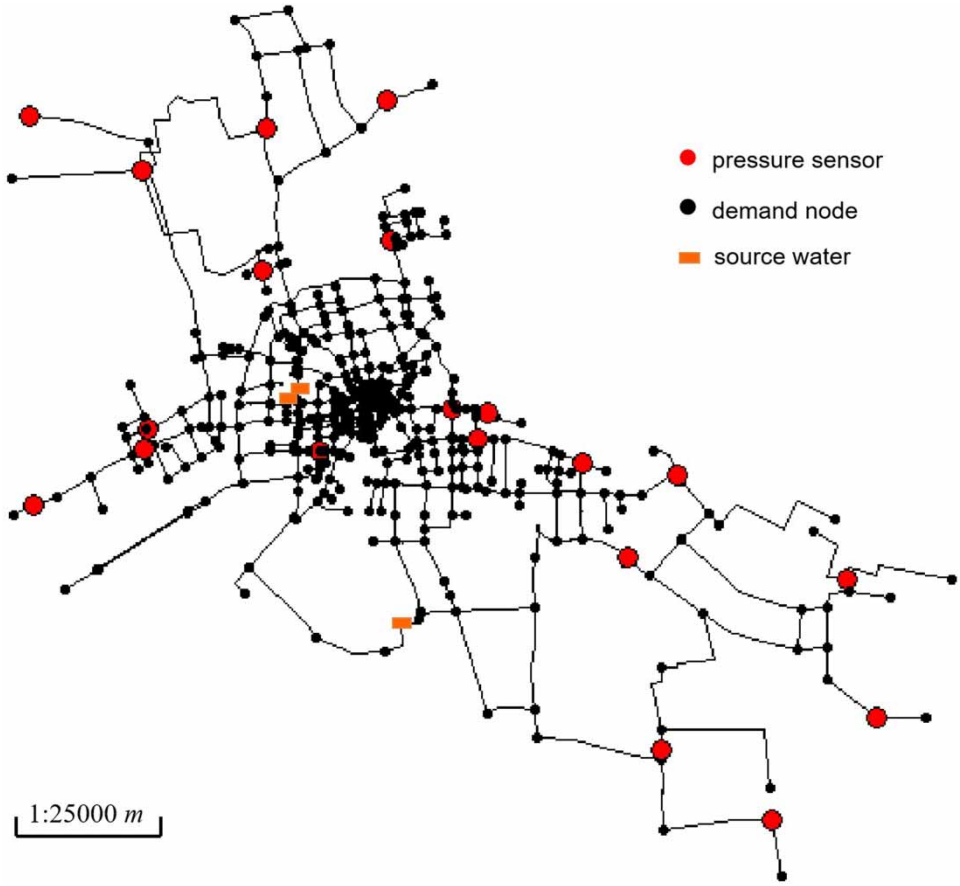

Figure 3 | Schematic representation of the water distribution network studied.

number of selected sensors (n_g) to determine the region detection order is set to 2. The score-threshold (λ) is set in a way that the probability of the total score of the scenario (SS) exceeding 5%.

To test the performance of the approach in leakage localization, several leakage scenarios are generated to represent the actual leakage events. The leakage intensity for samples tested is divided into six flow rate intervals: 20~30 m^3/h, 30~40 m^3/h, 40~50 m^3/h, 50~100 m^3/h, 100~200 m^3/h, 200~350 m^3/h. The leakage flow rate at the node is generated by randomly sampling in the corresponding interval. The samples contain new leakage that occurs at any time in the predicted time domain. The sample traverses all nodes, and each node simulates 30 different leakage intensities in the corresponding interval. The number of samples for each flow rate interval is 14,730 (491×30).

The number of dominant sensor sequence

Figure 4 shows the values of the sensor sensitivity weight $\bar{S}$ for the case study. Nodes with a small topological distance or pipeline distance to the sensor nodes have a big weight value, quantifying the hydraulic correlation between the sensors and the nodes. The weight peaks occur overwhelmingly when the X-axis and Y-axis coordinates correspond to the same node index, indicating that the most relevant leakage is the one on the node itself. As shown in Table 1, the sensor index 1 corresponds to node index 368, and the weight value is 6.269 when the coordinate is (368,1). The dominant sensor sequence for Node 368 is constructed based on the first column of the Table 1. For example, the dominant sensor sequence is (1, 16, 10, 12, 6) when the sequence length is set to 5. Some sensors are much more sensitive to leakage at a certain node than others. Thus, the dominant sensor sequence is developed for each leakage node to enrich the positive spatial information to the sensors.

The dominant sensor sequence is a container covered by multiple dominant sensors. To explore the reasonable number of dominant sensors (N_{sd}), the number of dominant sensors N_{sd} is taken from the integer interval [3, 20] ($N_{sd.} = 3,4,\ldots,20$). Two

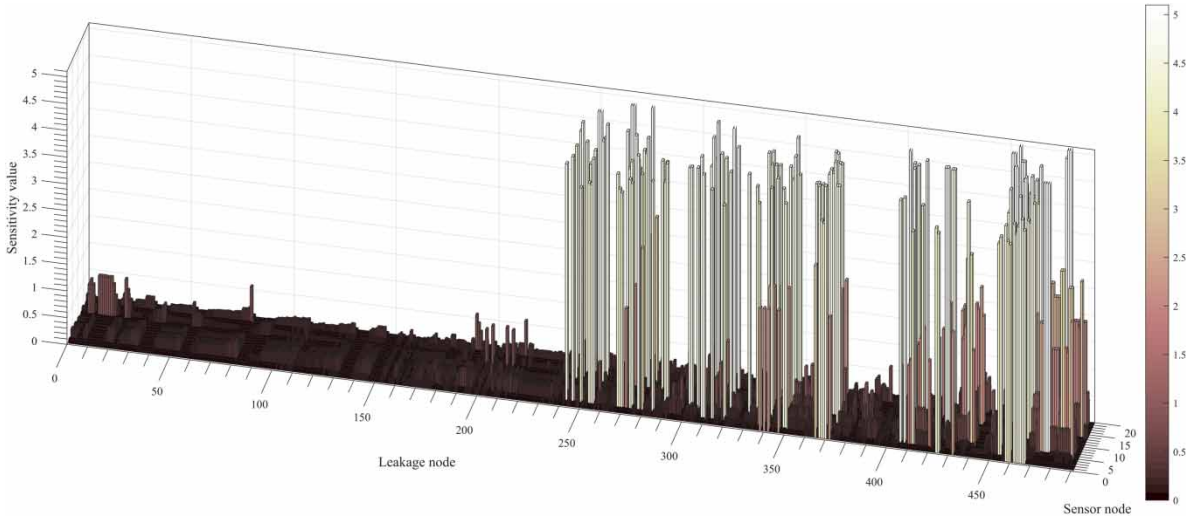

Figure 4 | Sensor sensitivity weight matrix.

Table 1 | Partial weight peaks of the sensitivity matrix

	Sensitivity									
	Node Index									
Sensor Index	368	350	390	258	476	290	436	403	318	459
1 (368)	6.269	0.062	0.076	0.059	0.013	0.151	0.019	0.029	0.013	5.181
2 (350)	0.062	8.330	0.076	0.018	0.004	0.041	0.006	0.009	0.013	0.062
3 (390)	0.076	0.023	0.076	0.258	0.055	0.170	0.080	0.124	0.013	0.076
4 (258)	0.059	0.018	0.076	0.744	0.078	0.133	0.117	0.186	0.013	0.059
5 (476)	0.013	0.004	0.076	0.078	45.29	0.029	0.531	0.320	0.013	0.013
6 (290)	0.151	0.041	0.076	0.133	0.029	4.525	0.042	0.064	0.013	0.151
7 (436)	0.019	0.006	0.076	0.117	0.531	0.042	3.402	0.622	0.013	0.019
8 (403)	0.029	0.009	0.076	0.186	0.320	0.064	0.622	1.197	0.013	0.029
9 (318)	0.042	0.013	0.076	0.308	0.149	0.094	0.245	0.414	0.013	0.042
10 (459)	5.181	0.062	0.076	0.059	0.013	0.151	0.019	0.029	0.013	10.51
11 (255)	0.063	7.706	0.076	0.018	0.004	0.041	0.006	0.009	0.013	0.063
12 (340)	2.285	0.062	0.076	0.059	0.013	0.151	0.019	0.029	0.013	2.285
13 (15)	0.036	0.039	0.076	0.047	0.010	0.051	0.015	0.023	0.013	0.036
14 (488)	0.014	0.004	0.076	0.083	2.565	0.030	0.588	0.351	0.013	0.014
15 (455)	0.062	8.248	0.076	0.018	0.004	0.041	0.006	0.009	0.013	0.062
16 (424)	5.852	0.062	0.076	0.059	0.013	0.151	0.019	0.029	0.013	4.912
17 (215)	0.076	0.023	0.076	0.258	0.055	0.170	0.080	0.124	0.013	0.076
18 (456)	0.093	0.237	0.076	0.025	0.005	0.060	0.008	0.012	0.013	0.093
19 (479)	0.018	0.006	0.076	0.115	0.541	0.041	1.917	0.599	0.013	0.018
20 (431)	0.013	0.004	0.076	0.081	0.392	0.030	0.447	0.313	0.013	0.013

clustering algorithms, namely hierarchical clustering algorithm and K-means clustering algorithm, are used for the nodes spatial clustering. The localization performance is evaluated by the geographic distance between the actual leakage node and the top-ranked candidate region cenr.

Figure 5 gives the localization accuracy for the different number of dominant sensors. The trend of geographic distance over the number of dominant sensors can be divided into three stages. In the first stage ($3 \leq N_{sd} \leq 6$), the localization accuracy increases rapidly as the number of dominant sensors increases, indicating that more information from sensors that are sensitive to leakage is used, leading to a rapid increase in localization accuracy. In the second stage ($6 \leq N_{sd} \leq 10$), the accuracy does not change much as the number of dominant sensors increases. This is because the sensitivity of these newly adopted sensors gradually decreases whereas the impact of sensor noises increases. In the third stage ($11 \leq N_{sd} \leq 20$), the accuracy gradually decreases as the number of dominant sensors increases. The main reason is that the newly used sensors are not sensitive enough to the leakage and the pressure fluctuations are mainly caused by the noise, resulting in misleading the leakage localization. Therefore, this is the trade-off between information enrichment caused by the increase in the number of dominant sensors and deterioration caused by the noises. The localization accuracy will be reduced while the number of dominant sensors N_{sd} is too small or too large. A reasonable number of dominant sensors can achieve a more accurate leakage localization. In this case study, the number of dominant sensors is set to 6 ($N_{sd} = 6$).

Compared with the traditional sensor sequence ($N_{sd} = 20$) which is utilized by Shao *et al.* (2019), this proposed method using the dominant sensor sequence can improve the localization accuracy. It is worth noting that the use of dominant sensors can greatly improve the localization accuracy especially when big noise exists in the measurements. As shown in Figure 5, the geographic distance for $\sigma_p = 0.1m$ is reduced from 2,000 m to 1,000 m when N_{sd} is reduced from 20 to 6. This accuracy improvement of using dominant sensors is more significant in the case of greater noise than smaller noise. The geographic distance is reduced from 6,000 m to 3,000 m when $\sigma_p = 0.3\ m$. Using the dominant sensor sequences can improve localization accuracy by about 50% compared with the method that does not use the dominant sensor sequence. The localization accuracy is deteriorated with the noise standard deviation increasing from 0.1 m to 0.3 m. Similar phenomena can be found in (Blesa *et al.* 2014; Pérez *et al.* 2014a). This indicates that accurate monitoring equipment will help improve localization accuracy, allowing more sensors to participate in leakage localization.

Candidate region detection priority

A set of candidate regions can be obtained based on the method described in the section Candidate Leakage Region Ranking. Then the detection priority of these regions should be ranked. As mentioned previously, a total of 14,730 leakage events are simulated to test the performance of the developed method. Figure 6 shows the candidate regions and their detection priority for one of these leakage events. Three candidate regions are obtained and the number '1' indicates that the detection order is first and the leakage is most likely to occur in this region, the area of which is about 390,625 square meters. The actual

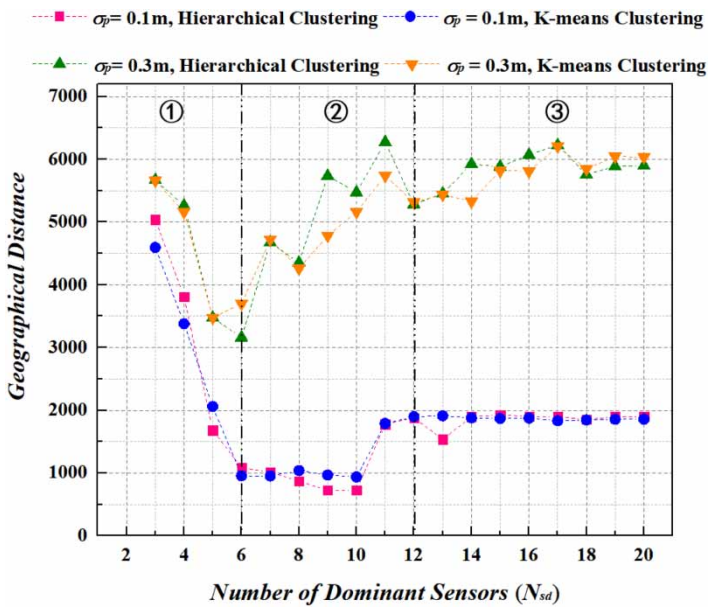

Figure 5 | Localization performance for different dominant sensor sequence lengths.

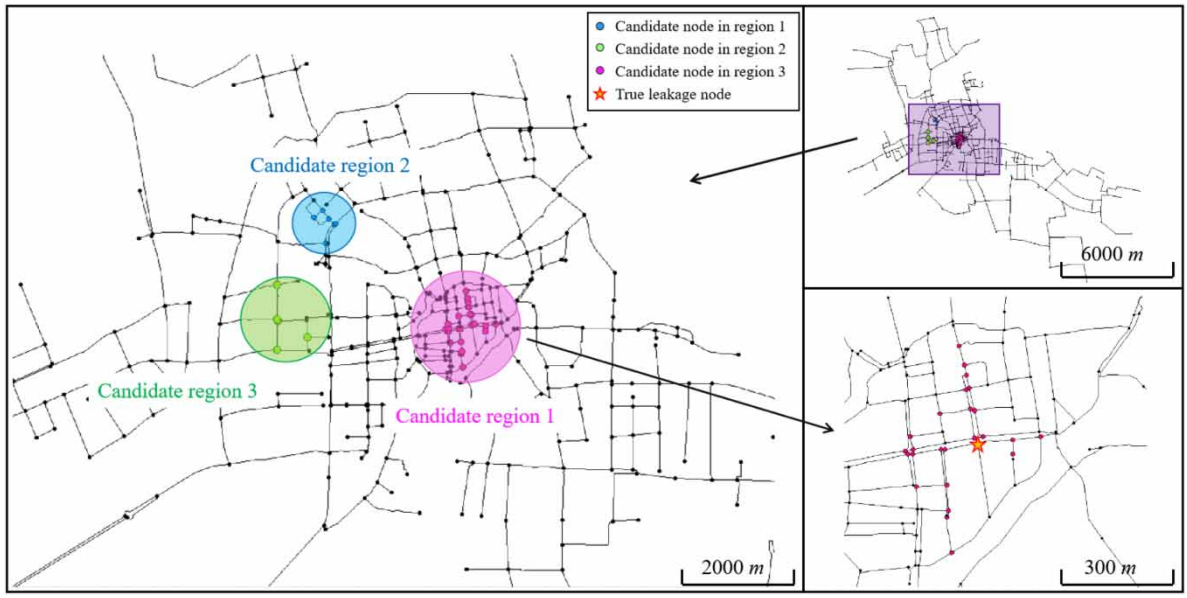

Figure 6 | Candidate leakage region of leakage localization.

leakage node is within the candidate region '1st', indicating that the leakage localization accuracy is good, and the detection priority helps to shorten the leakage localization time.

The summary results of region localization of 14,730 leakage events are illustrated in Table 2. Approximately 97% of leakage events have been located in the first three candidate regions. About 74.5% of leakage events are located in the region '1st', indicating that the leakage can be efficiently located since the operator only needs to inspect the '1st' region. Compared with the smaller leakage events, the events with large leakage have a large probability within the region '1st'. When the leakage rates are within 20–50 m^3/h, the probability within the region '1st' is approximately 70%. It increases from 70% to 79% when the leakage rates increase from 50 m^3/h to 350 m^3/h.

Figure 7 gives the cumulative probability of the geographic distance from the actual leakage position to the candidate nodes in the candidate region '1st' for the 14,730 leakage events. The cumulative probability is approximately 0.35 for 1,000 m, 0.6 for 2,000 m, and 0.7 for 3,000 m, showing the region localization accuracy is acceptable. However, about 30% of leakage events are at a distance greater than 3,000 m. This is because some leakage events are not located in the region '1st' (Table 2).

Leakage localization

The above section gives the detection priority of the candidate regions. The detection priority of the nodes in the region should be determined based on the method presented in the section Candidate Node Analysis.

Table 2 | The ratio of matching results for different flow rate intervals (%)

	Priority			
Flow rate interval	**1st**	**2nd**	**3rd**	**Undetected**
20~30 m^3/h	70.50	19.30	6.30	3.90
30~40 m^3/h	70.10	21.60	5.90	2.40
40~50 m^3/h	71.30	21.00	3.30	4.50
50~100 m^3/h	77.80	15.50	5.10	1.60
100~200 m^3/h	78.10	16.80	3.00	2.10
200~350 m^3/h	79.20	16.50	2.60	1.60
Average	74.50	18.45	4.37	2.68

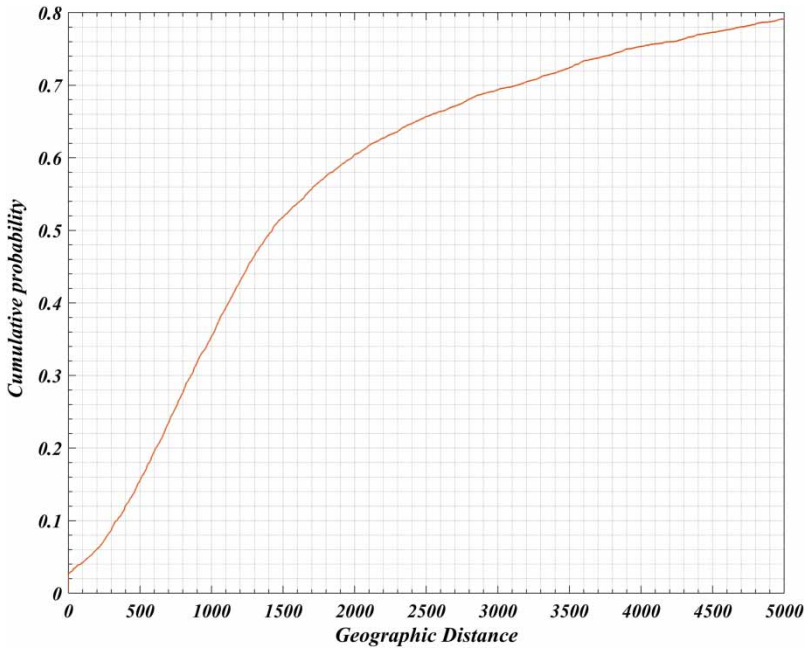

Figure 7 | The cumulative probability distribution of geographic distance.

Figure 8 gives the geographical distance of all nodes in the candidate leakage region '1st' (see Figure 6). Two indicators, namely Sum score and Cumulative ratio, are shown in Figure 8(a) and 8(b) respectively. The node with the highest indicator value is shown with a dashed line. As shown in Figure 8(a), the node with the highest Sum score is close to the actual leakage node and all the nodes are within 2,000 m of the actual leakage node. As shown in Figure 8(b), the node with the highest Cumulative ratio is also close to the actual leakage node.

Figure 9 gives the cumulative probability of the geographic distance from the actual leakage position to the candidate nodes. $Node_{1st}$, $Node_{2nd}$, and $Node_{3rd}$ are the top three priority candidate nodes for every leakage event, respectively. '$Node_{1st+2nd+3rd}$' represents the optimal nodes among the top three priority candidate nodes, which are closest to the actual leakage node. The cumulative probabilities of the distance within 5,000 m are about 0.716, 0.703, 0.685, 0.724 for $Node_{1st}$, $Node_{2nd}$, $Node_{3rd}$, and $Node_{1st+2nd+3rd}$, respectively. The cumulative probability is approximately 0.568, 0.539, 0.518, 0.591 when the distance is less than 2,000 m. The large distance values are mainly due to the uncertainties, including the unknown leakage magnitude, the differences between the real and the estimated nodal water demands, and the measurement noises. However, even in the absence of uncertainty, the leakages in some nodes cannot be located, in the case of the

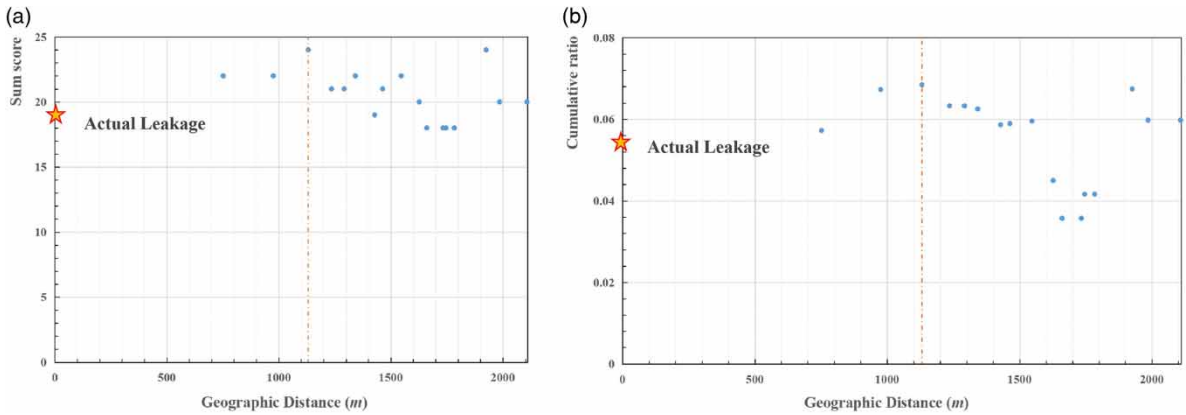

Figure 8 | Indicators versus geographic distance: (a) Sum score and (b) Cumulative ratio.

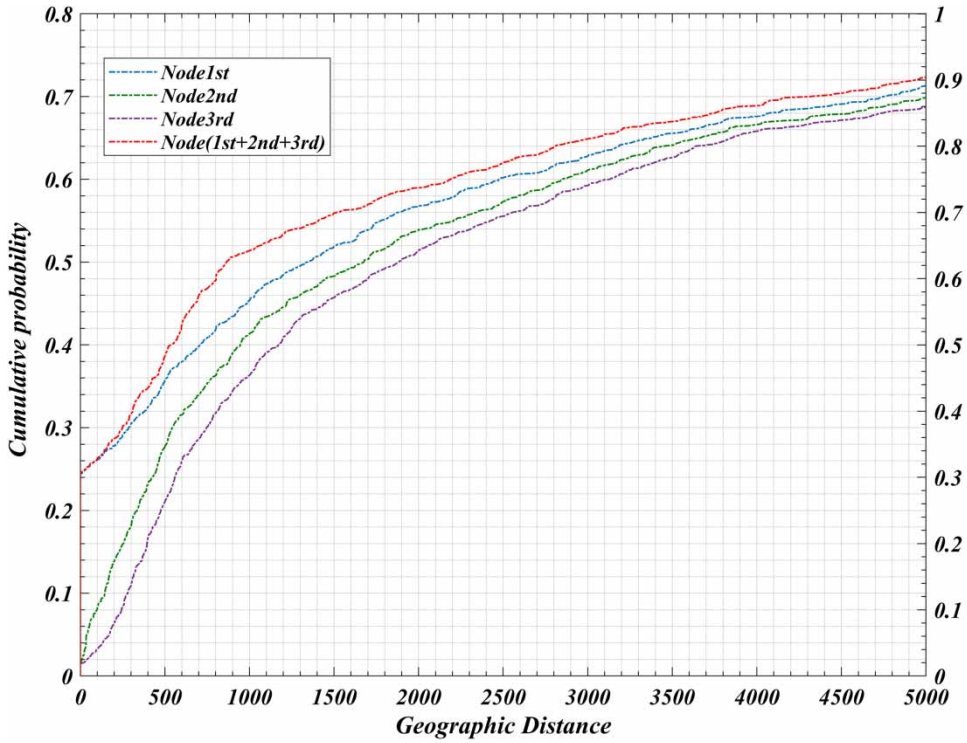

Figure 9 | The cumulative probability distribution of geographic distance versus top-ranked nodes.

nodes being located in the branch of the WDS whereas none of these nodes is equipped with a pressure sensor. Besides, pipe distances are also adopted to evaluate the leakage detection performance of the method and the corresponding results are shown in Appendix A.

CONCLUSIONS

This paper presents a novel model-based method for leakage localization in WDSs. It is characterized by (1) defining specific dominant sensor sequences for each candidate leakage node; (2) utilizing multiple time steps of the measurements which are temporal varying correlated; (3) ranking leakage regions and nodes by their possibility to contain the true leakage. Application to the WDS network highlights the effectiveness and robustness of the method, showing that the method can accurately and efficiently localize the leakage.

The adoption of the dominant sensor sequence can enhance leakage localization performance. There is an optimal number of dominant sensors that can maximize leakage localization accuracy. The optimal sensor number is 6–10 for different conditions in the case study. Using the dominant sensor sequences can improve localization accuracy by about 50% compared with the method that does not use the dominant sensor sequence. Region detection priority helps to shorten the leakage detection time. Approximately 97% of leakage events have been located in the candidate regions, indicating that the candidate leakage region is well defined within the considered leakage intensity. The cumulative probability of the distance between the actual leakage and the node selected is approximately 35% within 1,000 m, 60% within 2,000 m, and 70% within 3,000 m, showing good localization accuracy.

Several research tasks remain open. The proposed approach has been developed assuming only a single leakage occurs. The extension to multiple leakages is possible but it would require numerous leakage scenarios to be simulated. This could be very time-consuming. Considering that the sensor fault usually exists in the sensor networks, it is also of interest to develop a sensor-fault detection method to improve the robustness of the developed method.

ACKNOWLEDGEMENTS

The present research is funded by the National Key R&D Program of China (No. 2016YFC0400600), the National Natural Science Foundation of China (No. 52070165), the National Science and Technology Major Projects for Water Pollution

Control and Treatment (2017ZX07502003-05), and the Fundamental Research Funds for the Central Universities of China (No. 2020QNA4030).

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Abokifa, A. A., Haddad, K., Lo, C. & Biswas, P. 2019 Real-time identification of cyber-physical attacks on water distribution systems via machine learning-based anomaly detection techniques. *Journal of Water Resources Planning and Management* **145** (1), 13.

Arandia, E., Ba, A., Eck, B. & McKenna, S. 2016 Tailoring seasonal time series models to forecast short-term water demand. *Journal of Water Resources Planning and Management* **142** (3), 10.

Blesa, J., Puig, V. & Saludes, J. 2012 Robust identification and fault diagnosis based on uncertain multiple input-multiple output linear parameter varying parity equations and zonotopes. *Journal of Process Control* **22** (10), 1890–1912.

Blesa, J., Nejjari, F. & Sarrate, R. 2014 Robustness analysis of sensor placement for leak detection and location under uncertain operating conditions. *Procedia Engineering* **89**, 1553–1560.

Blesa, J., Nejjari, F. & Sarrate, R. 2015 Robust sensor placement for leak location: analysis and design. *Journal of Hydroinformatics* **18** (1), 136–148.

Duan, H., Ghidaoui, M. S. & Tung, Y. 2010 Energy analysis of viscoelasticity effect in pipe fluid transients. *Journal of Applied Mechanics-T. ASME* **77** (4), 044503. doi:10.1115/1.4000915,(ASCE) 77(0445034).

Duan, H., Lee, P. J., Ghidaoui, M. S. & Tung, Y. 2012 System response function-based leak detection in viscoelastic pipelines. *Journal of Hydraulic Engineering* **138** (2), 143–153. doi:10.1061/(ASCE)HY.1943-7900.0000495,(ASCE).

Duan, H., Pan, B., Wang, M., Chen, L., Zheng, F. & Zhang, Y. 2020 State-of-the-art review on the transient flow modeling and utilization for urban water supply system (UWSS) management. *Journal of Water Supply: Research and Technology-Aqua* **69** (8), 858–893. doi:10.2166/aqua.2020.048,(ASCE).

Farley, B., Mounce, S. R. & Boxall, J. B. 2013 Development and field validation of a burst localization methodology. *Journal of Water Resources Planning and Management* **139** (6), 604–613.

Huang, Y., Zheng, F., Kapelan, Z., Savic, D., Duan, H. & 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. doi:10.1029/2020WR028285,(ASCE) 56(e2020WR02828510).

Kang, D. & Lansey, K. 2009 Real-time demand estimation and confidence limit analysis for water distribution systems. *Journal of Hydraulic Engineering* **135** (10), 825–837.

Kang, J., Park, Y. J., Lee, J., Wang, S. H. & Eom, D. S. 2018 Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. *IEEE Transactions on Industrial Electronics* **65** (5), 4279–4289.

Kleijnen, J. P. C. 2017 Regression and kriging metamodels with their experimental designs in simulation: a review. *European Journal of Operational Research* **256** (1), 1–16.

MacKay, D. J. C. 2004 Information theory, inference, and learning algorithms. *IEEE Transactions on Information Theory* **50** (10), 2544–2545.

MacQueen, J. 1967 Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 1967*, Berkeley, Calif. University of California Press. pp. 281–297.

Manzi, D., Brentan, B., Meirelles, G., Izquierdo, J. & Luvizotto, E. 2019 Pattern recognition and clustering of transient pressure signals for burst location. *Water* **11** (11), 2279.

Meniconi, S., Capponi, C., Frisinghelli, M. & Brunone, B. 2021 Leak detection in a real transmission main through transient tests: deeds and misdeeds. *Water Resources Research, AGU* **57** (3), e2020WR027838. doi: 10.1029/2020WR027838).

Moser, G., Paal, S. G. & Smith, I. F. C. 2018 Leak detection of water supply networks using error-domain model falsification. *Journal of Computing in Civil Engineering* **32** (2), 18.

Perez, R., Puig, V., Pascual, J., Quevedo, J., Landeros, E. & Peralta, A. 2011 Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks. *Control Engineering Practice* **19** (10), 1157–1167.

Pérez, R., Cugueró, M. A., Cugueró, J. & Sanz, G. 2014a Accuracy assessment of leak localisation method depending on available measurements. *Procedia Engineering* **70**, 1304–1313.

Perez, R., Sanz, G., Puig, V., Quevedo, J., Angel, M., Escofet, C., Nejjari, F., Meseguer, J., Cembrano, G., Tur, J. M. M. & Sarrate, R. 2014b Leak localization in water networks a model-based methodology using pressure sensors applied to a real network in barcelona. *IEEE Control Systems Magazine* **34** (4), 24–36.

Ponce, M. V. C., Castanon, L. E. G. & Cayuela, V. P. 2014 Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. *Journal of Hydroinformatics* **16** (3), 649–670.

Puust, R., Kapelan, Z., Savic, D. A. & Koppel, T. 2010 A review of methods for leakage management in pipe networks. *Urban Water Journal* **7** (1), 25–45.

Qi, Z. X., Zheng, F. F., Guo, D. L., Maier, H. R., Zhang, T. Q., Yu, T. C. & 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), 11.

Rossman, L. A. 2000 *EPANET 2: Users Manual*. US EPA, Washington, DC.

Salguero, F. J., Cobacho, R. & Pardo, M. A. 2018 Unreported leaks location using pressure and flow sensitivity in water distribution networks. *Water Supply* **19** (1), 11–18.

Sanz, G., Perez, R., Kapelan, Z. & Savic, D. 2016 Leak detection and localization through demand components calibration. *Journal of Water Resources Planning and Management* **142** (2), 13.

Sarrate, R., Blesa, J. & Nejjari, F. 2014 Clustering techniques applied to sensor placement for leak detection and location in water distribution networks. In *2014 22nd Mediterranean Conference on Control and Automation*. IEEE, New York, pp. 109–114.

Shao, Y., Li, X., Zhang, T., Chu, S. & Liu, X. 2019 Time-series-based leakage detection using multiple pressure sensors in water distribution systems. *Sensors (Basel)* **19** (14), 3070.

Soldevila, A., Blesa, J., Tornil-Sin, S., Duviella, E., Fernandez-Canti, R. M. & Puig, V. 2016 Leak localization in water distribution networks using a mixed model-based/data-driven approach. *Control Engineering Practice* **55**, 162–173.

Soldevila, A., Jensen, T. N., Blesa, J., Tornil-Sin, S., Fernandez-Canti, R. M. & Puig, V. 2018 *Leak Localization in Water Distribution Networks Using A Kriging Data-Based Approach*. IEEE, New York, NY.

Steffelbauer, D. B. & Fuchs-Hanusch, D. 2016 Efficient sensor placement for leak localization considering uncertainties. *Water Resources Management* **30** (14), 5517–5533.

Sun, J. L., Wang, R. & Duan, H. F. 2016 Multiple-fault detection in water pipelines using transient time-frequency analysis. *Journal of Hydroinformatics – IWA* **18** (6), 975–989. doi: 10.2166/hydro.2016.232.

Sun, C., Parellada, B., Puig, V. & Cembrano, G. 2019 Leak localization in water distribution networks using pressure and data-driven classifier approach. *Water* **12** (54), 1–14.

Wu, Y. P. & Liu, S. M. 2017 A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal* **14** (9), 972–983.

Wu, Y., Liu, S., Wu, X., Liu, Y. & Guan, Y. 2016 Burst detection in district metering areas using a data driven clustering algorithm. *Water Research* **100**, 28–37.

Wu, Y. P., Liu, S. M., Smith, K. & Wang, X. T. 2018a Using correlation between data from multiple monitoring sensors to detect bursts in water distribution systems. *Journal of Water Resources Planning and Management* **144** (2), 10.

Wu, Y. P., Liu, S. M. & Wang, X. T. 2018b Distance-based burst detection using multiple pressure sensors in district metering areas. *Journal of Water Resources Planning and Management* **144** (11), 6.

Xie, X., Zhou, Q., Hou, D. & Zhang, H. 2017 Compressed sensing based optimal sensor placement for leak localization in water distribution networks. *Journal of Hydroinformatics* **20** (6), 1286–1295.

Zhang, Q. Z., Wu, Z. Y., Zhao, M., Qi, J. Y., Huang, Y. & Zhao, H. B. 2016 Leakage zone identification in large-scale water distribution systems using multiclass support vector machines. *Journal of Water Resources Planning and Management* **142** (11), 15.

Zhou, X., Tang, Z. H., Xu, W. R., Meng, F. L., Chu, X. W., Xin, K. L. & Fu, G. T. 2019 Deep learning identifies accurate burst locations in water distribution networks. *Water Research* **166**, 12.

First received 8 February 2021; accepted in revised form 28 June 2021. Available online 9 July 2021