# Supplementary Information

Pressure-Response-Based Online Cascaded System for Detecting and Localizing Unreported Leaks in Water Distribution Networks

Zijun Zhang[1], Zhao Liu[1,*], and Wenjie Li[1]

[1] School of Automation and Electrical Engineering, University of Jinan, Jinan, Shandong 250022, China.

[*] Correspondence: Zhao Liu; cse_liuz@ujn.edu.cn.

## Supplementary Methods

### Data Splitting and Training Settings

Scenario identifiers were assigned to the training, validation, and test subsets in proportions of 70%, 15%, and 15%, respectively. Normal and leakage samples associated with the same identifier remained in the same subset. The detection test sets contained 30000, 21000, 12000, and 15000 samples for L-town, Modena, Hanoi, and FOWM, respectively, with normal and leakage samples constructed in pairs. Localization evaluation used only the 15000, 10500, 6000, and 7500 leakage samples, respectively.

The detector and localizer were trained independently using cross-entropy loss, the Adam optimizer, and a batch size of 32, with initial learning rates of 0.003 and 0.005, respectively. Binary F1 was used for detection, and support-weighted multiclass F1 was used for localization. Model weights for each candidate window were selected by validation-set F1. Online evaluation used the detection and localization windows in Table 2 of the main text, together with the fixed probability thresholds, 5-min updates, and rule of three consecutive confirmations described in Section 2.2 of the main text.

### Offline Screening Criteria for Detection and Localization Windows

The selection of offline detection and localization windows considered both performance improvement and the amount of historical data required. Composite scores were used to identify candidate knee points where performance gains leveled off as window length increased. Task-specific performance constraints were used to determine acceptable windows. The shortest candidate window satisfying the constraints was then selected.

The candidate detection-window lengths were 0.25 h, 0.5 h, 0.75 h, 1 h, 2 h, and 3 h, and the candidate localization-window lengths were 0.25 h, 0.5 h, 0.75 h, 1 h, 2 h, 3 h, 4 h, 5 h, and 6 h. The offline test set was used to calculate detection F1 and false-positive rate (FPR), and localization Top-3 hit rate and the average shortest-path distance (ASD) defined in Equations (11)–(12) of the main text. Windows were screened using Equations (S1)–(S9). Table 2 of the main text reports the results for the recommended windows on the screening data. Tables 3, S2, and S3 draw on the

same complete set of online scenarios evaluated after the models and windows were fixed. Table 3 reports the online evaluation per 5000 leakage scenarios and 5000 no-leak scenarios for each network. Tables S2 and S3 provide supplementary statistics for the specified leakage scenarios selected from that set, reporting the proportion of correctly detected and localized scenarios and the distribution of delays to the first formal localization output, respectively.

The maximum and minimum of each metric were calculated within each network. The FPR tolerance increment and Top-3 retention proportion were set according to network type.

For the $j$th detection window $W_j$, the composite detection score was defined as

$$S_j^{\mathrm{det}} = \frac{1}{2}\left[\frac{\mathrm{F1}_j - \mathrm{F1}_{\min}}{\mathrm{F1}_{\max} - \mathrm{F1}_{\min}} + \frac{\mathrm{FPR}_{\max} - \mathrm{FPR}_j}{\mathrm{FPR}_{\max} - \mathrm{FPR}_{\min}}\right] \tag{S1}$$

Here, $S_j^{\mathrm{det}}$ is the composite detection score for window $W_j$. The F1 score for that window is $\mathrm{F1}_j$. The maximum and minimum F1 scores across all candidate detection windows in the same network are $\mathrm{F1}_{\max}$ and $\mathrm{F1}_{\min}$, respectively. The FPR for window $W_j$ is $\mathrm{FPR}_j$. The maximum and minimum FPR values across those windows are $\mathrm{FPR}_{\max}$ and $\mathrm{FPR}_{\min}$, respectively.

The F1 term represents overall detection performance. The FPR term is reverse-normalized so that a lower FPR yields a higher score. The two terms are weighted equally. If either metric has the same value across all candidate windows, making its denominator zero, the corresponding normalized term is set to 1. The same rule applies to the localization score below.

For the $j$th localization window $W_j$, the composite localization score was defined as

$$S_j^{\mathrm{loc}} = \frac{1}{2}\left[\frac{\text{Top-3}_j - \text{Top-3}_{\min}}{\text{Top-3}_{\max} - \text{Top-3}_{\min}} + \frac{\mathrm{ASD}_{\max} - \mathrm{ASD}_j}{\mathrm{ASD}_{\max} - \mathrm{ASD}_{\min}}\right] \tag{S2}$$

Here, $S_j^{\mathrm{loc}}$ is the composite localization score for window $W_j$. The proportion of samples with the true location among the three highest-ranked candidate regions or nodes at that window is $\text{Top-3}_j$. The maximum and minimum Top-3 values across all candidate localization windows in the same network are $\text{Top-3}_{\max}$ and $\text{Top-3}_{\min}$, respectively.

$\mathrm{ASD}_j$ is the ASD for window $W_j$, in m. $\mathrm{ASD}_{\max}$ and $\mathrm{ASD}_{\min}$ are the maximum and minimum ASD values, respectively, across all candidate localization windows in the same network.

The Top-3 term directly reflects the proportion of true locations included among the three highest-ranked candidates. The ASD term is reverse-normalized so that a smaller spatial error yields a higher score.

Knee points were calculated separately for detection and localization in each network. For either task, let the $J$ candidate windows be ordered by increasing length as $W_1, \ldots, W_J$, with corresponding composite scores $S_1, \ldots, S_J$. Window length and composite score were normalized as

$$u_j = \frac{W_j - W_1}{W_J - W_1}, \qquad v_j = \frac{S_j - S_1}{S_J - S_1} \tag{S3}$$

Here, $u_j$ is the normalized window length, and $v_j$ is the composite score normalized relative to the first and last windows. When $S_J \neq S_1$, the shortest and longest windows map to (0,0) and (1,1), respectively. The straight line connecting these normalized endpoints is the linear-gain baseline, given by $v = u$.

The perpendicular distance from the $j$th internal candidate window to the linear-gain baseline was

$$d_j = \frac{|v_j - u_j|}{\sqrt{2}}, j = 2, \ldots, J-1 \tag{S4}$$

The candidate knee-point window was defined as

$$W_{\text{knee}} = W_{j^*}, j^* = \arg\max_{2 \le j \le J-1} d_j \tag{S5}$$

Here, $W_{\text{knee}}$ is the candidate knee-point window, $j^*$ is its index, and $d_j$ is the perpendicular distance from the $j$th internal window to the linear-gain baseline. The endpoints were excluded from knee-point selection. If $S_J = S_1$ and $v_j$ was therefore undefined, the distance criterion was skipped and the longest candidate window was selected.

The search began at the candidate detection knee point and proceeded toward longer windows. The shortest window satisfying the FPR constraint was selected:

$$W^*_{\text{det}} = \min_{W_j \ge W^{\text{det}}_{\text{knee}}} \{W_j \mid \text{FPR}_j \le \text{FPR}_{\text{min}} + \delta_{\text{FPR}}\} \tag{S6}$$

Here, $W^*_{\text{det}}$ is the recommended detection window. The candidate knee-point window on the composite detection-score curve is $W^{\text{det}}_{\text{knee}}$. The minimum FPR across all candidate detection windows in the same network is $\text{FPR}_{\text{min}}$. The allowed FPR tolerance increment is $\delta_{\text{FPR}}$.

This criterion selects the shortest detection window that satisfies the FPR constraint. The value of $\delta_{\text{FPR}}$ was set according to network type:

$$\delta_{\text{FPR}} = \begin{cases} 0.003, & \text{distribution-oriented networks} \\ 0.030, & \text{transmission-oriented networks} \end{cases} \tag{S7}$$

FPR is expressed as a proportion in Equation (S6). The prespecified FPR tolerance increments were 0.003 and 0.030 for the two network types, equivalent to 0.3 and 3.0 percentage points, respectively.

The search began at the candidate localization knee point and proceeded toward longer windows. The shortest window satisfying the Top-3 constraint was selected:

$$W^*_{\text{loc}} = \min_{W_j \ge W^{\text{loc}}_{\text{knee}}} \left\{W_j \mid \text{Top-3}_j \ge \eta_{\text{Top-3}} \text{Top-3}_{\text{max}}\right\} \tag{S8}$$

Here, $W^*_{\text{loc}}$ is the recommended localization window. The candidate knee-point window on the composite localization-score curve is $W^{\text{loc}}_{\text{knee}}$. The maximum Top-3 value across all candidate localization windows in the same network is $\text{Top-3}_{\text{max}}$. The Top-3 retention proportion is $\eta_{\text{Top-3}}$.

This criterion selects the shortest localization window that satisfies the Top-3 coverage constraint. The value of $\eta_{\text{Top-3}}$ was set according to network type:

$$\eta_{\text{Top-3}} = \begin{cases} 0.990, & \text{distribution-oriented networks} \\ 0.950, & \text{transmission-oriented networks} \end{cases} \quad \text{(S9)}$$

If neither the candidate knee-point window nor any longer candidate window satisfied the corresponding constraint, the longest candidate window was selected as a fallback. Test-set metrics for the recommended windows are reported in Table 2 of the main text. Detection F1 and FPR correspond to the recommended detection window. Localization Top-3 and ASD correspond to the recommended localization window.

The vertical dashed lines in Figures 4 and 5 of the main text indicate the recommended detection and localization windows, respectively.

The window-screening sensitivity analysis used the offline metrics for each candidate window, keeping the normalization, knee-point determination, selection of the shortest window satisfying the constraints, and fallback rules in Equations (S1)–(S9) unchanged. The weights of the F1 term in the detection score and the Top-3 term in the localization score were each set to 0.25, 0.40, 0.50, 0.60, and 0.75 in turn, with the weight of the other term equal to 1 minus the specified weight. The FPR tolerance increment $\delta_{\text{FPR}}$ was set to 0.50, 0.75, 1, 1.25, and 1.50 times the baseline value for the corresponding network in Equation (S7). The Top-3 retention proportion $\eta_{\text{Top-3}}$ was set to 0.980, 0.985, 0.990, 0.995, and 1.000 for distribution-oriented networks and 0.930, 0.940, 0.950, 0.960, and 0.970 for transmission-oriented networks. There were 25 combinations per task per network and 200 combinations in total; the screening results are reported in Table S10.

**Construction of Leak Regions in Modena**

The construction of leak regions in Modena was informed by the hydraulic-simulation and regional-partitioning approach of Zhang et al. [1]. In this study, an orifice leak was introduced in turn at each candidate node, and the pressure response relative to the normal condition was calculated and used as the input to K-means clustering. The resulting 56-region mapping was used to label the classes of the training and evaluation samples and was kept fixed during subsequent model training and localization evaluation.

The resulting clusters were then adjusted subject to constraints on region size and topological connectivity. Clustering metrics were calculated from the adjusted labels.

The Modena hydraulic model contains 272 hydraulic nodes, including four reservoir nodes. The remaining 268 nodes are candidate leak nodes. Let $m$ denote the number of candidate leak nodes and $n$ the number of hydraulic nodes. For Modena, $m = 268$ and $n = 272$.

First, the hydraulic model was run under normal, no-leak conditions to obtain the baseline pressure head at each hydraulic node. Pressure-dependent orifice leaks were then applied at each of the 268 candidate leak nodes, and the hydraulic simulation was rerun.

For region construction, the signed pressure response at hydraulic node $j$ due to a leak at candidate node $i$ was defined as

$$\Delta P_{ij} = P_{ij} - P_j \tag{S10}$$

Here, $i = 1, \dots, m$ indexes candidate leak nodes, and $j = 1, \dots, n$ indexes hydraulic nodes. $\Delta P_{ij}$ is the signed pressure response at hydraulic node $j$ caused by a leak at candidate node $i$, in m. $P_j$ is the pressure head at hydraulic node $j$ under normal conditions, in m. $P_{ij}$ is the pressure head at hydraulic node $j$ when a leak occurs at candidate node $i$, in m.

This definition preserves the direction of the pressure-head change caused by a leak. The pressure-response row feature vector for candidate leak node $i$ was therefore written as

$$\mathbf{r}_i = [\Delta P_{i1} \quad \Delta P_{i2} \quad \cdots \quad \Delta P_{in}] \tag{S11}$$

Here, $\mathbf{r}_i$ is the pressure-response row feature vector for candidate leak node $i$. Its elements are measured in m, and its length is $n$. The pressure-response matrix was formed by stacking the row feature vectors for all candidate leak nodes:

$$\mathbf{R} = [\Delta P_{ij}]_{m \times n}, i = 1, \dots, m, j = 1, \dots, n \tag{S12}$$

Here, $\mathbf{R}$ is the pressure-response matrix, and its elements are measured in m. For Modena, $\mathbf{R} \in \mathbb{R}^{268 \times 272}$. Rows of $\mathbf{R}$ correspond to candidate leak nodes, and columns correspond to hydraulic nodes. Each element represents the nodal pressure response under the corresponding candidate leak condition.

The signed pressure response was retained for region construction to distinguish pressure-head decreases from increases. In contrast, the absolute values of the pressure responses defined in Equations (2) and (4) were used as classifier inputs. Region construction was based on the similarity of pressure-response patterns caused by different candidate leak nodes. Physical topological connectivity was addressed during the subsequent constraint adjustments.

Leaks for region construction were represented by the same pressure-dependent orifice model used in the controlled simulations, as given in Equation (10):

$$q_{\mathrm{l}} = C_{\mathrm{d}} A \sqrt{2gH} \tag{S13}$$

Here, $q_{\mathrm{l}}$ is the leak flow rate, in $\mathrm{m}^3/\mathrm{s}$. The orifice discharge coefficient is $C_{\mathrm{d}}$. The leak area is $A$, in $\mathrm{m}^2$. Gravitational acceleration is $g$, in $\mathrm{m/s}^2$. The pressure head at the leak node is $H$, in m.

For Modena, $A = 0.00018019\ \mathrm{m}^2$, $C_{\mathrm{d}} = 0.75$, and $g = 9.81\ \mathrm{m/s}^2$. Using the mean normal-condition pressure head at candidate leak nodes, $H_{\mathrm{ref}} \approx 25.13$ m, as the reference head, this leak area corresponds to a reference leak flow of approximately 3.0 L/s.

The leak area specified above was used to generate the pressure-response matrix for Modena partitioning. The online evaluation used the fixed leak area for each network given in Section 2.4 of the main text, corresponding to a reference leak flow of approximately 1.58 L/s at the respective reference head. Under both settings, instantaneous leak flow varied with local pressure head according to Equation (S15).

Each candidate leak node was treated as a clustering object, with its corresponding pressure-response row from matrix $\mathbf{R}$ used as its feature vector. To determine the number of regions in

Modena, values of $k$ from 10 to 100 were evaluated individually. Here, $k$ denotes the number of candidate regions and is dimensionless.

For each value of $k$, an initial K-means clustering was performed. Region-size constraints were then applied, followed by topological-connectivity correction. An undirected topological graph was constructed from the connected nodes and pipes in the Modena network. Pipe lengths were used as edge weights.

If a region consisted of multiple disconnected components in the network topology, its main connected component was retained. The remaining disconnected components were reassigned to adjacent regions with more similar pressure-response features.

After these adjustments, the silhouette coefficient and Inertia (within-cluster sum of squares) were calculated from the adjusted labels. The silhouette coefficient was the primary criterion for selecting the number of regions. Inertia was used to show changes in within-cluster compactness across different region counts.

The silhouette coefficient is calculated by normalizing the difference between the mean distance of each object to other objects in its cluster and the mean distance to objects in the nearest neighboring cluster. The coefficient ranges from $-1$ to 1. Higher values indicate tighter within-cluster grouping and greater separation between clusters.

The number of regions corresponding to the maximum silhouette coefficient was selected as optimal. The coefficient reached its maximum at $k = 56$ (Figure S1). Therefore, the Modena network was divided into 56 leak regions.

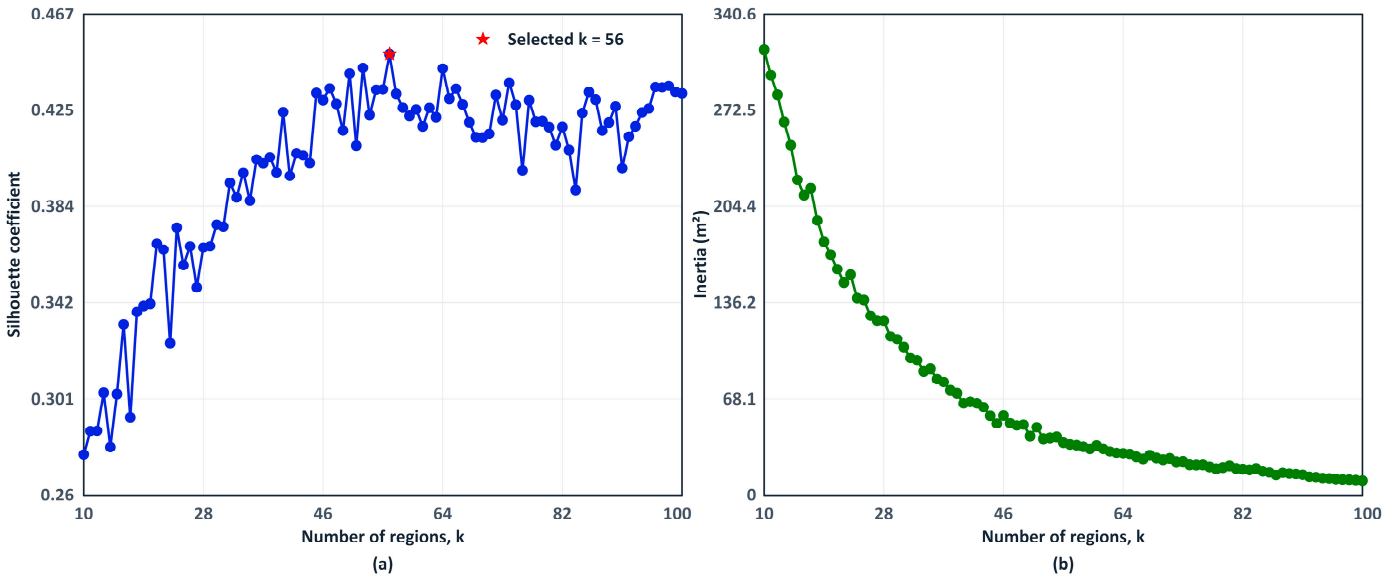


**Figure S1.** Selection of the number of leak regions in Modena. (a) Silhouette coefficients for different numbers of regions; (b) Inertia (within-cluster sum of squares) for different numbers of regions.

The center node of region $z$ was defined using pipe-length-weighted topological distance:

$$c_z = \underset{v \in V_z}{\operatorname{argmin}} \sum_{u \in V_z} d_z(v, u) \tag{S14}$$

Here, $z$ is the region index. The center node of region $z$ is $c_z$. The set of candidate leak nodes in that region is $V_z$. The nodes $v$ and $u$ are arbitrary members of $V_z$. $d_z(v, u)$ is the pipe-length-weighted shortest-path distance between nodes $v$ and $u$ in the subgraph corresponding to region $z$, in m. The operator argmin denotes the node that minimizes the objective function.
Thus, the center node is the node with the smallest sum of pipe distances to all nodes in the region. The topological distribution of the regions and their center nodes is shown in Figure S2.

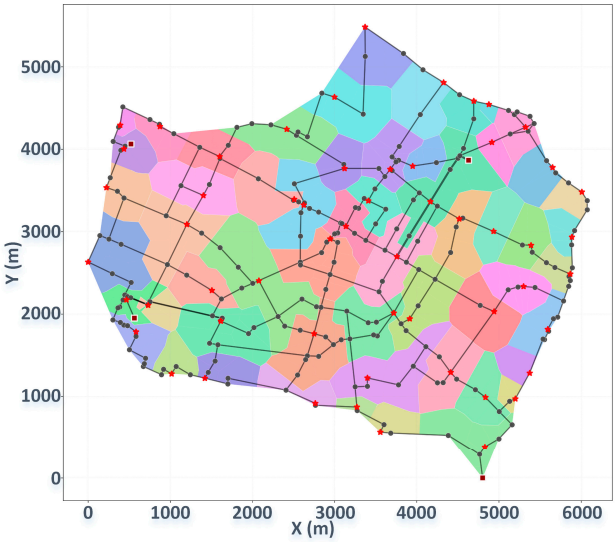


**Figure S2.** The 56 leak regions in Modena. Red stars indicate the center node of each region. Red squares denote reservoirs.

**Leak-Area Conversion for Controlled Scenarios**

The American Water Works Association (AWWA) M36 manual defines unreported leaks as leaks unknown to the public and discovered through active leak detection by the water utility. Such leaks typically do not surface [2] (pp. 174, 195).
For estimating unavoidable annual real losses, AWWA M36 uses a reference flow of 25 gpm, approximately 1.58 L/s, for unreported leaks and breaks in mains or pipelines. The reference pressure is 70 psi [2] (Table 3-22, p. 102).
To account for differences in pressure levels and leak intensities among networks, the target leak-flow range under reference-head conditions was set to 1.0–3.0 L/s. This range includes the reference flow above.
Leak simulations used the pressure-dependent orifice model in the Water Network Tool for Resilience (WNTR):

$$q_l(t) = C_d A \sqrt{2gH(t)} \tag{S15}$$

Here, $q_{\mathrm{l}}(t)$ is the leak flow rate at time $t$, in $\mathrm{m}^3/\mathrm{s}$. The orifice discharge coefficient is $C_{\mathrm{d}} = 0.75$. The leak area is $A$, in $\mathrm{m}^2$. Gravitational acceleration is $g = 9.81\ \mathrm{m/s}^2$. The pressure head at the leak node at time $t$ is $H(t)$, in m.

To standardize reference flows across networks, the target reference flows were converted to corresponding leak-area ranges using $H_{\mathrm{ref}}$, the mean normal-condition pressure head at candidate leak nodes in each network. The conversion used

$$q_{\min} = 0.001\ \mathrm{m}^3/\mathrm{s}, q_{\max} = 0.003\ \mathrm{m}^3/\mathrm{s} \tag{S16}$$

$$A_{\min} = \frac{q_{\min}}{C_{\mathrm{d}}\sqrt{2gH_{\mathrm{ref}}}} \tag{S17}$$

$$A_{\max} = \frac{q_{\max}}{C_{\mathrm{d}}\sqrt{2gH_{\mathrm{ref}}}} \tag{S18}$$

Here, $q_{\min}$ and $q_{\max}$ are the lower and upper bounds of the target reference-flow range, and $A_{\min}$ and $A_{\max}$ are the corresponding leak areas.

During simulation, the instantaneous leak flow at a leak node was dynamically calculated using Equation (S15). Thus, leak flow varied with nodal pressure and operating conditions.

The reference leak-to-demand ratio was defined as the target reference flow under reference-head conditions divided by the sum of nodal base demands of the network:

$$\gamma = \frac{q_{\mathrm{ref}}}{Q_{\mathrm{sys}}} \tag{S19}$$

Here, $q_{\mathrm{ref}}$ is the target reference flow under reference-head conditions. $Q_{\mathrm{sys}}$ is the sum of nodal base demands reported in Table 1 of the main text, calculated without demand-pattern multipliers. The two quantities have the same units, and $\gamma$ is dimensionless and reported as a percentage.

In Table S1, $Q_{\mathrm{sys}}$ was taken from Zhang et al. [3]. $H_{\mathrm{ref}}$ was calculated from normal-condition hydraulic simulations. Leak areas were converted from the target reference flows and $H_{\mathrm{ref}}$, and $\gamma$ was calculated using Equation (S19).

The lower and upper bounds of the ratio of reference leak flow to system base demand correspond to $q_{\mathrm{ref}} = 1.0$ L/s and 3.0 L/s, respectively. Based on the system base demands in Table 1 of the main text and Table S1, the ratios of the 1.58 L/s reference leak flow to base demand are 5.3559%, 0.3883%, 0.0285%, and 0.3578% for L-town, Modena, Hanoi, and FOWM, respectively.

**Table S1. Parameters for leak-area conversion in the four evaluation networks.**

| Network | Reference head $H_{\mathrm{ref}}$ (m) | Sum of nodal base demands $Q_{\mathrm{sys}}$ (L/s) | Target reference flow (L/s) | Leak-area range ($m^2$) | Reference leak-to-demand ratio |
|---|---|---|---|---|---|
| L-town | 46.27 | 29.5 | 1.0–3.0 | 0.00004426–0.00013278 | 3.390%–10.169% |
| Modena | 25.13 | 406.94 | 1.0–3.0 | 0.00006006–0.00018019 | 0.246%–0.737% |
| Hanoi | 41.86 | 5538.90 | 1.0–3.0 | 0.00004653–0.00013959 | 0.018%–0.054% |
| FOWM | 55.81 | 441.63 | 1.0–3.0 | 0.00004030–0.00012091 | 0.226%–0.679% |

## Supplementary Results

### Proportion of Correctly Detected and Localized Scenarios and Its Uncertainty

The leakage scenarios used for the supplementary statistics were selected from the complete set of simulated scenarios used for online evaluation, with leak locations distributed evenly across localization classes. All 33 and 56 regional center nodes were used in L-town and Modena, respectively, and all 31 and 44 candidate nodes were used in Hanoi and FOWM, respectively. For each leak location, 5, 2, 5, and 4 hydraulic conditions were used in L-town, Modena, Hanoi, and FOWM, respectively. Each condition was combined with 24 leak-onset times from 0 to 23 h, yielding 3960, 2688, 3720, and 4224 leakage scenarios, respectively. The proportion of correctly detected and localized scenarios was calculated from these scenarios, and its uncertainty was estimated by resampling scenario groups. Correct localization is determined by a Top-1 hit in the first formal localization output after leak onset. The numerator is the number of correctly detected and localized scenarios, and the denominator is all leakage scenarios in that set, including scenarios in which no leak was detected or no formal localization output was produced. The results are presented in Table S2. Each scenario group comprised scenarios sharing the same leak node and hydraulic condition but corresponding to 24 leak-onset times. Resampling was stratified by leak node, with scenario groups resampled with replacement within each stratum for 5000 iterations. All leak-onset times within a selected group were retained in each iteration, and the proportion of correctly detected and localized scenarios was recalculated. The 95% bootstrap confidence interval was calculated using the percentile method: the proportions obtained from the 5000 resampling iterations were sorted in ascending order, and the 2.5th and 97.5th percentiles were taken as the lower and upper bounds, respectively, retaining approximately the middle 95% of the sorted results.

**Table S2. Proportion of correctly detected and localized scenarios and 95% confidence intervals in the supplementary simulations.**

| Network | Leakage scenarios | Correctly detected and localized scenarios | Proportion of correctly detected and localized scenarios (%) | 95% confidence interval (%) |
|---|---|---|---|---|
| L-town | 3960 | 3622 | 91.46 | 90.58–92.30 |
| Modena | 2688 | 2206 | 82.07 | 80.84–83.30 |
| Hanoi | 3720 | 1873 | 50.35 | 47.77–52.98 |
| FOWM | 4224 | 1781 | 42.16 | 40.67–43.70 |

**Delay to the First Formal Localization Output**

The delay to the first formal localization output is defined as the time difference between the true leak onset and the first formal localization output. The analyses in Tables S2 and S3 used the same set of leakage scenarios selected from the online scenario set in the main text. Table S3 presents the number of these leakage scenarios, the number that produced a first formal localization output, and their delay distribution. Delays were calculated only for scenarios that produced an output and summarized using the median and lower and upper quartiles. The lower and upper quartiles correspond to the 25th and 75th percentiles, respectively, and describe the spread of the delay distribution. Quantiles were estimated by linear interpolation after sorting the samples. When the localization window is longer than the detection window, the configuration wait given by Equations (5) and (6) of the main text ensures that the localization window at the end of the wait has the same starting time as the detection window at the alarm time. Window length therefore affects both the pressure history used for localization and the time at which candidate locations are output.

**Table S3. Distribution of delays to the first formal localization output in the supplementary simulations.**

| Network | Leakage scenarios | Scenarios producing a first formal localization output | Median (min) | Lower quartile (min) | Upper quartile (min) |
|---|---|---|---|---|---|
| L-town | 3960 | 3960 | 100 | 95 | 105 |
| Modena | 2688 | 2466 | 70 | 70 | 70 |
| Hanoi | 3720 | 3437 | 205 | 170 | 265 |
| FOWM | 4224 | 3671 | 255 | 235 | 305 |

**Annual False-Alarm Statistics**

Annual false-alarm statistics are based on the complete alarm sequence for each number of consecutive confirmations in the one-year L-town replay. For each setting, alarms at adjacent 5-min update times are merged into continuous alarm intervals. An interval that does not overlap

with any true leak period is counted as a false-alarm interval. The effective leak-free observation time is calculated as the number of time steps for which a complete detection window can be formed and the system is leak-free, multiplied by 5 min. False-alarm frequency is calculated as the number of false-alarm intervals divided by the effective number of leak-free observation days. Cumulative false-alarm duration is calculated as the number of alarm time steps during leak-free periods multiplied by 5 min, including false-alarm intervals and periods of continued alarms after leak termination. Event alarm delay is the time difference between the true leak onset and the first alarm within that leak period. If an alarm occurs at the update time corresponding to leak onset, the delay is 0 min. The effective leak-free observation time was 269.8056 d. The complete alarm sequence for three consecutive confirmations contained 63 false-alarm intervals, giving a false-alarm frequency of 0.2335 events/d. Another 6 alarm intervals continued beyond leak termination. The false-alarm intervals and periods of continued alarms after leak termination contained a total of 197 alarm time steps during leak-free periods, equivalent to 16.4167 h. Annual alarm statistics and alarm delays for the six leak events RL1–RL6 in Table 4 of the main text are reported in Tables S8 and S9, respectively.

**Probability Thresholds and Number of Consecutive Confirmations**

The probability-threshold analysis used the selected detection window for each network: 1 h for L-town and Modena, 3 h for Hanoi, and 2 h for FOWM. Comparisons at thresholds of 0.60–0.95 for L-town and Modena and 0.50–0.95 for Hanoi and FOWM kept the model and test samples fixed within each network and changed only the decision threshold. The offline detection results for the four networks are presented in Tables S4–S7, respectively. The fixed probability threshold $\tau$ was set to 0.50, 0.60, 0.70, 0.80, 0.90, and 0.95 in turn; a window was classified as leakage when its leakage probability exceeded $\tau$. FPR is the proportion of normal windows classified as leakage, recall is the proportion of leakage windows correctly detected, precision is the proportion of windows classified as leakage that actually contain leakage, and F1 is the harmonic mean of precision and recall. When the threshold increased from 0.50 to 0.90, FPR in Hanoi decreased from 7.6000% to 0.1333% and recall decreased from 83.78% to 59.03%; FPR in FOWM decreased from 4.8000% to 0.0267% and recall decreased from 78.39% to 63.00%.

**Table S4. Offline detection results at different probability thresholds in L-town.**

| Probability threshold | FPR (%) | Recall (%) | Precision (%) | F1 (%) |
|---|---|---|---|---|
| 0.50 | 0.273 | 96.7933 | 99.7184 | 98.23 |
| 0.60 | 0.1733 | 96.5467 | 99.8208 | 98.1564 |
| 0.70 | 0.0867 | 96.1133 | 99.9099 | 97.9749 |
| 0.80 | 0.0533 | 95.6467 | 99.9443 | 97.7483 |
| 0.90 | 0.0067 | 94.3667 | 99.9929 | 97.0984 |
| 0.95 | 0.0000 | 92.3600 | 100.0000 | 96.0283 |

**Table S5. Offline detection results at different probability thresholds in Modena.**

| Probability threshold | FPR (%) | Recall (%) | Precision (%) | F1 (%) |
|---|---|---|---|---|
| 0.50 | 0.133 | 97.1714 | 99.8630 | 98.50 |
| 0.60 | 0.1238 | 97.0857 | 99.8726 | 98.4595 |
| 0.70 | 0.0667 | 97.0190 | 99.9313 | 98.4537 |
| 0.80 | 0.0190 | 96.8476 | 99.9803 | 98.3890 |
| 0.90 | 0.0095 | 96.4952 | 99.9901 | 98.2116 |
| 0.95 | 0.0000 | 95.7238 | 100.0000 | 97.8152 |

**Table S6. Offline detection results at different probability thresholds in Hanoi.**

| Probability threshold | FPR (%) | Recall (%) | Precision (%) | F1 (%) |
|---|---|---|---|---|
| 0.50 | 7.6000 | 83.7833 | 91.6834 | 87.5555 |
| 0.60 | 4.7500 | 80.2000 | 94.4085 | 86.7261 |
| 0.70 | 2.5667 | 76.1000 | 96.7373 | 85.1866 |
| 0.80 | 1.0833 | 70.7167 | 98.4912 | 82.3244 |
| 0.90 | 0.1333 | 59.0333 | 99.7746 | 74.1780 |
| 0.95 | 0.0000 | 45.6333 | 100.0000 | 62.6688 |

**Table S7. Offline detection results at different probability thresholds in FOWM.**

| Probability threshold | FPR (%) | Recall (%) | Precision (%) | F1 (%) |
|---|---|---|---|---|
| 0.50 | 4.8000 | 78.3867 | 94.2298 | 85.5812 |
| 0.60 | 2.9600 | 76.4667 | 96.2733 | 85.2345 |
| 0.70 | 1.4400 | 73.6933 | 98.0834 | 84.1568 |
| 0.80 | 0.5333 | 69.6800 | 99.2404 | 81.8737 |
| 0.90 | 0.0267 | 63.0000 | 99.9577 | 77.2880 |
| 0.95 | 0.0000 | 57.2667 | 100.0000 | 72.8275 |

The consecutive-confirmation analysis used the full-year L-town pressure inputs and a fixed detection model, with a 1 h detection window, $\tau = 0.50$, and rolling-window updates every 5 min. The same sequence of window-level leakage probabilities was evaluated with 1–10 confirmations. One confirmation corresponds to a single positive prediction, which produces an alarm when the leakage probability in the current window exceeds the threshold. Two or more confirmations require this condition to be satisfied at the corresponding number of consecutive update times. Annual alarm statistics and event-specific alarm delays are presented in Tables S8 and S9, respectively. The statistical definitions are provided under "Annual False-Alarm Statistics". All settings detected all six events, with alarm confirmation rates of 100%. Increasing the number of confirmations from 1 to 3 reduced the number of false-alarm intervals from 263 to 63 and false-alarm frequency from 0.9748 events/d to 0.2335 events/d, while increasing the median alarm delay from 30 min to 40 min. Increasing the number of confirmations further to 6 reduced the number of

false-alarm intervals to 9, with a false-alarm frequency of 0.0334 events/d and a median alarm delay of 55 min. Both tables use the complete alarm sequences. The condition for retaining alarm intervals used for temporal coverage in Table 4 and the alarm periods shown in Figure 7 is described in Section 3.3 of the main text.

**Table S8. Annual alarm statistics for different numbers of consecutive confirmations in L-town.**

| Number of consecutive confirmations | Number of false-alarm intervals | False-alarm frequency (events/d) | Number of detected events | Median alarm delay (min) |
|---|---|---|---|---|
| 1 | 263 | 0.9748 | 6 | 30 |
| 2 | 110 | 0.4077 | 6 | 35 |
| 3 | 63 | 0.2335 | 6 | 40 |
| 4 | 35 | 0.1297 | 6 | 45 |
| 5 | 15 | 0.0556 | 6 | 50 |
| 6 | 9 | 0.0334 | 6 | 55 |
| 7 | 6 | 0.0222 | 6 | 60 |
| 8 | 3 | 0.0111 | 6 | 65 |
| 9 | 1 | 0.0037 | 6 | 70 |
| 10 | 1 | 0.0037 | 6 | 75 |

**Table S9. Alarm delays for individual leak events with different numbers of consecutive confirmations in L-town.**

| Number of consecutive confirmations | RL1 (min) | RL2 (min) | RL3 (min) | RL4 (min) | RL5 (min) | RL6 (min) |
|---|---|---|---|---|---|---|
| 1 | 70 | 35 | 0 | 25 | 65 | 0 |
| 2 | 75 | 40 | 5 | 30 | 100 | 5 |
| 3 | 100 | 45 | 10 | 35 | 170 | 10 |
| 4 | 105 | 50 | 15 | 40 | 175 | 15 |
| 5 | 110 | 55 | 20 | 45 | 180 | 20 |
| 6 | 115 | 60 | 25 | 50 | 185 | 25 |
| 7 | 120 | 65 | 30 | 55 | 190 | 30 |
| 8 | 125 | 70 | 35 | 60 | 195 | 35 |
| 9 | 130 | 75 | 40 | 65 | 200 | 40 |
| 10 | 295 | 80 | 45 | 70 | 205 | 45 |

**Window-Screening Sensitivity**

In Table S10, “Number of combinations” is the total number of combinations of score-weight and screening-constraint settings, and “Selections of the original recommended window” is the

number of combinations in which the original recommended window was selected. "Window range" spans the shortest to the longest selected window across the settings. "Change in F1 or Top-3 hit rate" is the range of differences obtained by subtracting the value for the original recommended window from the detection F1 or localization Top-3 hit rate for each selected window, expressed in percentage points. The original recommended detection and localization windows were 1 h and 2 h for L-town, 1 h and 2 h for Modena, 3 h and 4 h for Hanoi, and 2 h and 5 h for FOWM, respectively.

**Table S10. Window-screening sensitivity results.**

| Network | Task | Original recommended window (h) | Number of combinations | Selections of the original recommended window | Window range (h) | Change in F1 or Top-3 hit rate (percentage points) |
|---|---|---|---|---|---|---|
| L-town | Detection | 1 | 25 | 14 | 0.75–3 | -0.28 to 0.65 |
| L-town | Localization | 2 | 25 | 14 | 1–6 | -0.99 to 0.84 |
| Modena | Detection | 1 | 25 | 25 | 1–1 | 0.00 to 0.00 |
| Modena | Localization | 2 | 25 | 20 | 0.5–2 | 0.00 to 0.00 |
| Hanoi | Detection | 3 | 25 | 20 | 2–3 | -2.19 to 0.00 |
| Hanoi | Localization | 4 | 25 | 15 | 3–4 | -2.40 to 0.00 |
| FOWM | Detection | 2 | 25 | 20 | 2–3 | 0.00 to 1.45 |
| FOWM | Localization | 5 | 25 | 15 | 4–5 | -3.20 to 0.00 |

**Computational Cost**

Mean detection-stage processing time is calculated by dividing cumulative detection-stage processing time by the number of detection windows. Mean localization-stage processing time is calculated by dividing cumulative actual localization-stage processing time by the number of actual localization calls. Actual localization calls include calls triggered by false alarms. The results are reported in Table S11. Table S12 presents the localization-call counts included in the evaluation, all-window localization reference counts, and their cumulative processing times, calculated per 5000 leakage scenarios. Cumulative localization processing time and all-window localization reference processing time are obtained by multiplying the corresponding call counts by the same mean localization-stage processing time per operation. The calculations use unrounded counts and mean processing times. The resulting times are converted to hours and reported to three decimal places. The percentage reduction in localization processing time equals the localization-call reduction rate.

Forward-pass inference was tested on an AMD Ryzen 9 7950X central processing unit (CPU) using a single thread and a batch size of 1, timing only the model's forward computation. Inputs from 20

evaluation samples were used for each model. After 30 warm-up runs, the 20 inputs were repeated for 10 rounds, yielding 200 timed runs. The mean forward-pass inference time is the arithmetic mean of these 200 measurements. The peak process working set is the maximum resident physical-memory usage of the process since launch, including the interpreter, runtime libraries, model, and inputs. The number of parameters is the total number of scalar elements in the model parameters. The peak process working set is expressed in MiB, where 1 MiB equals 1048576 bytes. The measurements are listed in Table S13.

**Table S11. Mean processing time per stage operation in the supplementary simulations.**

| Network | Mean detection-stage processing time (ms) | Mean localization-stage processing time (ms) |
|---|---|---|
| L-town | 210.6 | 209.2 |
| Modena | 190.6 | 189.5 |
| Hanoi | 140.7 | 137.6 |
| FOWM | 304.9 | 317.5 |

**Table S12. Localization-call counts and cumulative localization processing time.**
Calculated per 5000 leakage scenarios.

| Network | Leakage scenarios | Localization calls | All-window localization reference count | Cumulative localization processing time (h) | All-window localization reference processing time (h) | Reduction (%) |
|---|---|---|---|---|---|---|
| L-town | 5000 | 982294 | 1690000 | 57.082 | 98.207 | 41.88 |
| Modena | 5000 | 866051 | 1690000 | 45.598 | 88.979 | 48.75 |
| Hanoi | 5000 | 591434 | 1570000 | 22.605 | 60.008 | 62.33 |
| FOWM | 5000 | 599335 | 1510000 | 52.865 | 133.192 | 60.31 |

**Table S13. Number of model parameters, CPU forward-pass inference time, and peak process working set.**

| Network | Task | Input window | Number of parameters | Mean forward-pass inference time (ms) | Peak process working set (MiB) |
|---|---|---|---|---|---|
| L-town | Detection | 1 h | 195330 | 9.019 | 598.0 |
| L-town | Localization | 2 h | 197345 | 9.637 | 585.6 |
| Modena | Detection | 1 h | 129538 | 15.072 | 554.6 |
| Modena | Localization | 2 h | 133048 | 14.665 | 546.0 |
| Hanoi | Detection | 3 h | 99202 | 8.065 | 510.2 |
| Hanoi | Localization | 4 h | 101087 | 7.883 | 506.8 |
| FOWM | Detection | 2 h | 100866 | 7.604 | 526.0 |

| Network | Task | Input window | Number of parameters | Mean forward-pass inference time (ms) | Peak process working set (MiB) |
|---|---|---|---|---|---|
| FOWM | Localization | 5 h | 103596 | 8.122 | 520.7 |

**Spatial Localization Statistics**

The spatial localization statistics use leakage samples from the offline test sets described in Section 2.4 of the main text. The numbers of leakage test samples were 15000, 10500, 6000, and 7500 for L-town, Modena, Hanoi, and FOWM, respectively, with localization windows of 2 h, 2 h, 4 h, and 5 h. Each leakage test window counts as one localization sample. For each regional localization sample, the distance from the true leak node to the center node of the highest-probability region is calculated, including samples with a correctly classified region. For node-level localization, the distance from the true leak node to the predicted node is calculated. All networks use pipe-length-weighted shortest-path distances, in m. When calculating path lengths, pump and valve connections are assigned a length of 0 m. If multiple direct connections exist between the same pair of nodes, the shortest connection is used for the distance calculation. The mean, median, and 90th percentile of the distances are summarized across all leakage test samples and reported in Table S14. The 90th percentile is the distance that approximately 90% of samples do not exceed. For each L-town sample, the nodes in the Top-3 candidate regions are combined, and pipes with both endpoints within this node union are selected. Their lengths are summed after duplicate pipe identifiers are removed, giving the total pipe length for that sample. The mean across all samples is then calculated and reported in Table S15.

**Table S14. Offline localization distance statistics.**

| Network | Localization window (h) | Leakage test samples | Mean (m) | Median (m) | 90th percentile (m) |
|---|---|---|---|---|---|
| L-town | 2 | 15000 | 211.3 | 179.2 | 369.7 |
| Modena | 2 | 10500 | 197.9 | 110.1 | 450.4 |
| Hanoi | 4 | 6000 | 540.6 | 0.0 | 1750.0 |
| FOWM | 5 | 7500 | 692.5 | 0.0 | 2231.1 |

**Table S15. Total pipe length within the Top-3 candidate regions in L-town.**

| Network | Localization window (h) | Leakage test samples | Mean total pipe length (m) |
|---|---|---|---|
| L-town | 2 | 15000 | 3843.9 |

The Top-1 hit rate for each node class is the proportion of its leakage test samples with a Top-1 hit. Hanoi and FOWM have 31 and 44 node classes, respectively, of which 10 and 23 have a Top-1 hit rate below 50%. The Top-3 hit proportion among samples without a Top-1 hit is calculated as the number of samples without a Top-1 hit but with a Top-3 hit divided by the total number of samples without a Top-1 hit; the results are reported in Table S16.

**Table S16. Node-level offline localization statistics.**

| Network | Number of classes with a Top-1 hit rate below 50% | Top-3 hit proportion among samples without a Top-1 hit (%) |
|---|---|---|
| Hanoi | 10 | 72.29 |
| FOWM | 23 | 47.83 |

A regional boundary node is defined as a demand node connected to at least one pipe crossing between regions. The remaining mapped demand nodes are classified as regional interior nodes. Leakage test sample counts, Top-1 hit rates, and pipe-length-weighted shortest-path distances from the true leak node to the center node of the predicted region are calculated separately according to the group containing the true leak node. The denominator of the Top-1 hit rate for each group is the number of leakage test samples in that group. The Top-1 hit rates for regional boundary and interior nodes were 64.38% and 89.27%, respectively, in L-town and 99.08% and 99.25% in Modena. The corresponding mean distances from the true leak node to the center node of the predicted region were 256.7 m and 193.0 m in L-town and 220.9 m and 150.9 m in Modena. The results are reported in Table S17.

**Table S17. Offline localization results for regional boundary and interior nodes.**

| Network | Group | Number of nodes | Leakage test samples | Top-1 hit rate (%) | Mean distance (m) | 90th percentile (m) |
|---|---|---|---|---|---|---|
| L-town | Regional boundary | 210 | 4307 | 64.38 | 256.7 | 404.1 |
| L-town | Regional interior | 572 | 10693 | 89.27 | 193.0 | 350.2 |
| Modena | Regional boundary | 158 | 7048 | 99.08 | 220.9 | 483.6 |
| Modena | Regional interior | 110 | 3452 | 99.25 | 150.9 | 338.9 |

**Processing Statistics for Leakage and No-Leak Scenarios**

The computational cost statistics use both leakage and no-leak scenarios from the online evaluation. Each complete detection window is counted as one detection operation. Actual localization calls include calls triggered by false alarms in no-leak scenarios and before leak onset in leakage scenarios. The detection-window counts, actual localization-call counts, and processing times for detection and actual localization are summed separately for the two scenario classes. The cumulative quantities for each class are calculated per 5000 scenarios and then combined, as reported in Table S18. Cumulative processing time includes input construction, model inference, and the corresponding processing operations. The mean processing times in Table S11 are

calculated using all detection windows and actual localization calls from both leakage and no-leak scenarios in the supplementary simulations.

**Table S18. Detection-window counts, localization calls, and cumulative processing time for leakage and no-leak scenarios.**

Calculated per 5000 leakage scenarios and 5000 no-leak scenarios.

| Network | Leakage scenarios | No-leak scenarios | Detection windows | Actual localization calls | Cumulative detection and localization processing time (h) |
|---|---|---|---|---|---|
| L-town | 5000 | 5000 | 3500000 | 983428 | 262.464 |
| Modena | 5000 | 5000 | 3500000 | 866051 | 230.900 |
| Hanoi | 5000 | 5000 | 3260000 | 611566 | 151.267 |
| FOWM | 5000 | 5000 | 3380000 | 656757 | 344.153 |

## Supplementary References


1. Zhang, Q.; Wu, Z.Y.; Zhao, M.; Qi, J.; Huang, Y.; Zhao, H. Leakage zone identification in large-scale water distribution systems using multiclass support vector machines. *J. Water Resour. Plan. Manag.* **2016**, *142*, 04016042. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000661.
2. Kunkel, G.A., Jr. *Water Audits and Loss Control Programs*, 4th ed.; AWWA Manual M36; American Water Works Association: Denver, CO, USA, 2016; ISBN 978-1-62576-100-2 (print); eISBN 978-1-61300-339-8 (electronic). https://doi.org/10.12999/awwa.m36ed4.
3. Zhang, Z.; Liu, Z.; Li, Y.; Li, W.; Cai, H. Small-scale leakage detection in water distribution networks based on parallel CNN–LSTM and cross-modal attention mechanism. *J. Water Resour. Plan. Manag.* **2026**, *152*, 04026056. https://doi.org/10.1061/jwrmd5.wreng-7324.