Leak detection and localization in a water distribution network from a year of utility SCADA readings

Ricardo Marquez, Ph.D.
rmarquez123@gmail.com

September 2026

Abstract

Leaks lose treated water and the energy spent on it, and a utility must find them from a handful of pressure gauges and flow meters while demand changes by the hour. This inference study takes the Battle of the Leakage Detection and Isolation Methods (BattLeDIM) dataset of the town L-Town (782 junctions, 905 pipes, three pressure-reducing valves, a pump and a tank; 33 pressure sensors, three flow sensors, a tank level and 82 automated meter readings at five-minute resolution) and asks how much of the benchmark's economic score a single hydraulic model, calibrated on 2018 and run forward in time through 2019 without looking ahead, can earn. Every night the inflow balances of two supply zones are compared with the same nights of earlier weeks, a CUSUM of nightly log Bayes factors raises alarms, and each alarm is localized by comparing, with the Laplace evidence, one leak hypothesis per junction in a joint model of hourly snapshots before and after the estimated start that shares a per-sensor offset; the reported pipe and time maximize the expected score. Calibration lowers the pressure misfit from 2.70 to 1.15 cm rms. On 2019 the method raised 16 new-leak alarms, submitted 13 detections, found 7 of 23 leaks with 4 false positives, and scored EUR 128,805 of the EUR 523,154 a perfect submission earns (EUR 113,061 without the submissions for leaks standing since 2018). The official economic score of this method (EUR 128,805) is lower than the best published score (Tongji, EUR 264,873). Its true positive rate (30.4 %) is lower than the best published rate (Under Pressure, 65.2 %). Its false positive count (4) is more than the fewest published (IRI, 1). A second evaluation adds a test for the disappearance of a known leak, with thresholds chosen on 2018; it is not blind to 2019. It found 7 of 23 leaks with 5 false positives and scored EUR 111,561. With the closure test, the official economic score (EUR 111,561) is lower than the best published score (Tongji, EUR 264,873). With the closure test, the true positive rate (30.4 %) is lower than the best published rate (Under Pressure, 65.2 %). With the closure test, the false positive count (5) is more than the fewest published (IRI, 1). The stated probabilities are overconfident: the 90 % credible set of pipes contained the running leak's pipe for 4 of 19 alarms, and the probability that the submitted pipe lies within 300 m of a leak has a Brier score of 0.293 over 19 localized alarms, worse than a constant one half. The main failure of the frozen evaluation is that leaks repaired during the year stay subtracted from later nights. The closure test lowers the flow the detector carries above the true leak flow in Areas A and B from 48 to 23 m^3/h on average between June and October, but the remainder still hides most new leaks there, and it closes the leak that ended in 1 of 6 closures near a true end. A localization takes 2.1 minutes on one thread.

1 Introduction

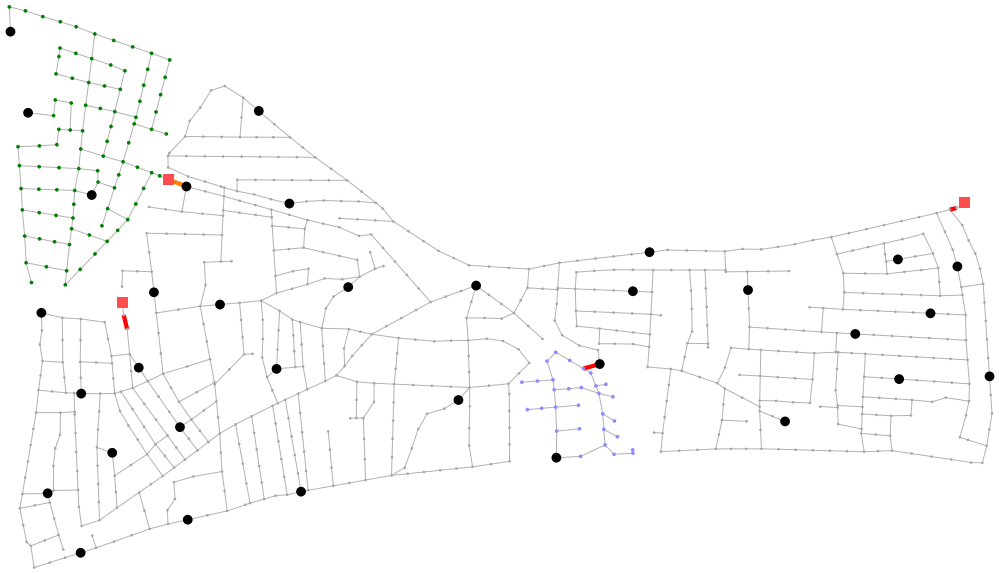
Water distribution networks lose a large share of the water they treat and pump through leaks in buried pipes, and with it the energy and chemicals spent on that water; a leaking pipe at low

pressure is also a path for contamination. Finding a leak soon after it starts and sending a crew to the right street is the operational problem, and it is hard for structural reasons: a town has hundreds of pipes and tens of sensors, demands are known only as patterns, and a leak of a few cubic meters per hour moves the gauged pressures by centimeters. Puust, Kapelan, Savic and Koppel (2010) review the hydraulic methods of leakage management. The Battle of the Leakage Detection and Isolation Methods (Vrachimis et al., 2022) put the community's methods on one realistic dataset: a year of historical SCADA readings of L-Town with a report of repaired leaks, a second year in which 23 leaks run, and an economic score that rewards water saved and charges for false alarms and for the distance a crew must search.

The teams of that competition combined three kinds of steps (Vrachimis et al., 2022). Detection in time used residuals against expected flows or pressures: statistical process control, CUSUM and likelihood-ratio change detection, decomposition of seasonal signals, or clustering. Localization used a hydraulic model: pressure sensitivities and correlation with residuals, simulation of candidate leaks compared with the readings, genetic or particle-swarm optimization over leak positions, or interpolation of the network state from the sensors. Most calibrated the nominal model's demands and roughness first. Factor graphs are a further family: Han, Eguchi, Mehrotra and Venkatasubramanian (2018) wrote a water network as a factor graph for state estimation and fault identification, and Irofti, Romero-Ben, Stoican and Puig (2025) localized leaks with factor graphs over a time window; factor graphs themselves are standard in estimation (Dellaert and Kaess, 2017). The hydraulic laws here are those of EPANET 2.2 (Rossman et al., 2020), including its pressure-reducing valve.

TERRA is a general-purpose engine for probabilistic inference on physical networks. It treats a network as a graph on which the physics is written locally: a conservation law at every node, a constitutive law on every edge, and, attached to the same graph, whatever the sensors report and whatever is known in advance. Each of these is a relation that holds only to within a stated tolerance, and the engine weighs them together to find the most probable state of the system, the uncertainty of every quantity in it, and the relative plausibility of competing explanations of the same readings. Domains enter through plugins that supply their laws, units and file formats; the inference and the bookkeeping are shared. This study uses its hydraulics plugin to write a town's pipes, pressure-reducing valves, pump and tank as conservation and constitutive relations, joins hourly snapshots before and after a suspected leak into one model with shared sensor offsets, ranks one leak hypothesis per junction by how well it explains the change in the readings, and turns the resulting posterior over pipes into the pipe whose report maximizes the expected benchmark score. Section 3 states the framework and each of these constructs precisely.

This report asks how much of the benchmark's score a method earns when it is held to the knowledge a utility has at each moment: the nominal model, the historical year with its repair report, and the second year's readings only up to the decision time, with every setting fixed on the historical year before the second year is read. The construction is a nightly flow-balance detector followed by an evidence ranking of leak hypotheses in a joint model of snapshots before and after the leak. What it adds is a causal, fully specified baseline on the benchmark with stated probabilities that are checked against the ground truth. Section 2 describes the network and the data, Section 3 the model and the method, Section 4 the calibration, the development year and the evaluation year, Section 5 the limits and Section 6 the answers.



black dots: the 33 pressure sensors; blue: Area B (31 junctions, behind PRV-3); green: Area C (92 junctions, fed by the tank); gray: Area A; red squares: the two reservoirs and the tank; red links: the three pressure-reducing valves; orange: the pump

Figure 1: L-Town, drawn from the coordinates of the nominal model: 782 junctions, 2 reservoirs, 1 tank, 905 pipes, 1 pump and 3 pressure-reducing valves. Inflow from the two reservoirs (p227, p235) and the pump flow are metered; the 82 automated meter readings are all in Area C.

2 System and data

Figure 1 shows L-Town. Two reservoirs at a head of 100 m feed Area A through pipes p227 and p235, whose flows are measured; pressure-reducing valves PRV-1 and PRV-2 hold the downstream pressure of Area A's inlets, and PRV-3 feeds Area B (31 junctions). A pump fills the tank T1 from Area A, and the tank alone feeds Area C (92 junctions), where 82 of the demand junctions carry automated meter readings. The head at a node is the water's energy per unit weight in meters (elevation plus pressure); the demand is the flow delivered to consumers there. The data (Vrachimis et al., 2020) are the nominal EPANET model, the SCADA readings of 2018 and of 2019 at five-minute steps — 33 pressures (m), the two inflows and the pump flow (m^3/h), the tank level (m) and the meters (L/h) — and a report of the links and repair times of ten leaks fixed in 2018. The readings were generated by the organizers with a model whose pipe diameters and roughnesses differ from the nominal ones by up to 10 %, with demand patterns the nominal model does not know, and with 23 leaks running in 2019, of which 19 start in 2019 and 4 run on from 2018. A second, "real" model and the leak list were published with the data; this method reads neither.

Figure 2 shows the model around one junction. Its assumptions are these. Flows are mass flows in kg/s at a density of 1000 kg/m^3; heads are in meters. Pipes follow the Hazen–Williams law with a small regularizing flow; roughness and diameter enter only through $CD^{2.63}$, so Area A's pipes are calibrated as three groups by diameter with one roughness factor each, and Areas B and C keep their nominal pipes. A snapshot is an hour of readings reduced to hourly means and solved as a steady state: the tank is a fixed head at its measured level, the pump a link at its measured flow, and the valves follow EPANET 2.2's pressure-reducing valve law. Area C's consumption is the pump flow less the tank's rate of fill; its unmetered junctions and every junction of Areas A and B share one demand scale per area and snapshot. Leaks are constant outflows at junctions

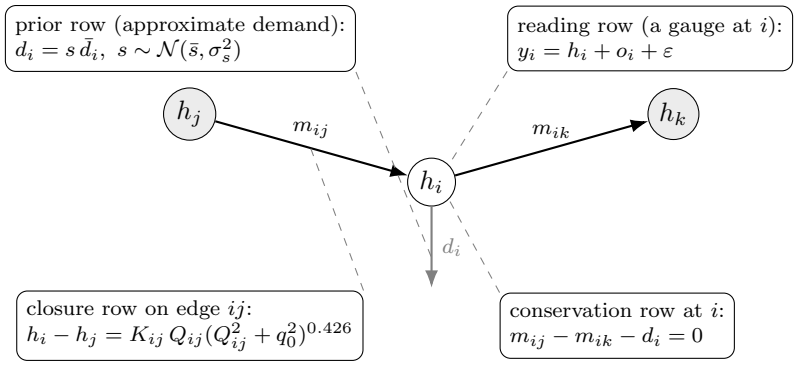


Figure 2: The model around one junction of one snapshot. Unknowns are the head, the mass flows on the incident links and the demand; the rows are conservation at the junction, the Hazen–Williams law on each pipe (on a pressure-reducing valve, the valve's active, open or closed law), a pressure reading with a sensor offset o_i shared by all snapshots, and the demand tied to its area's demand scale s, whose prior comes from the same hours of earlier weeks. A leak hypothesis adds one outflow to one junction's conservation row in every snapshot after the estimated start.

within the hours of one localization; leaks on pipes are placed at their end junctions.

3 Methods

3.1 The hydraulic factor graph of a snapshot

In one snapshot every junction i carries a head h_i and a demand d_i, and every link e a mass flow m_e. Conservation at a junction, the Hazen–Williams law on a pipe from i to j, and a pressure-reducing valve with setting head H (downstream elevation plus setting) are residuals

$$r_i = \sum_{e \to i} m_e - \sum_{i \to e} m_e - d_i - \ell_i, \tag{1}$$

$$r_e = h_i - h_j - K_e\,Q_e\,(Q_e^2 + q_0^2)^{0.426}, \tag{2}$$

$$r_{\text{active}} = h_j - H, \tag{3}$$

$$r_{\text{open}} = h_i - h_j - K_v\,m_e\sqrt{m_e^2 + m_0^2}, \tag{4}$$

$$r_{\text{closed}} = m_e, \tag{5}$$

with $Q_e = m_e/\rho$, $K_e = 10.667 L_e/(C_e^{1.852} D_e^{4.871})$, ℓ_i a leak outflow present only under a leak hypothesis, and K_v the open valve's minor loss. The valve is active (3) while the flow is forward and the upstream head exceeds H; open (4) while the flow is forward and the downstream head is at or below H; and closed (5) on reverse flow or while the downstream head is at or above H. A branch is chosen by a regime loop and re-checked at the fit. Physics rows are weighted at $(1\ \text{g/s})^{-2}$ and $(1\ \text{mm})^{-2}$. The readings and priors attach to the same unknowns: a pressure gauge at i reads the head plus a sensor offset o_i shared by all snapshots, the measured inflows read their pipes' flows, the tank balance reads the flow into Area C, a meter reads its junction's demand, and every

other demand is its nominal pattern value $\bar{d}_i$ times its area's scale s,

$$r = y_i - h_i - o_i, \tag{6}$$

$$r = y_e - m_e, \tag{7}$$

$$r = y_i - d_i, \tag{8}$$

$$r = d_i - s\,\bar{d}_i, \tag{9}$$

$$r = s - \bar{s}, \tag{10}$$

$$r = o_i, \tag{11}$$

weighted by the inverse squares of: 3 cm for a head reading, 1 % (at least 0.1 kg/s) for a measured inflow, 5 % (at least 0.3 kg/s) for Area C's balance, 1 % for a meter; the scale row (9) is exact; the scale prior (10) has standard deviation 0.1 for Areas A and B and 0.3 for Area C in snapshots after a leak's start and 0.5 and 1 before it; the offset prior (11) has 0.5 m. A leak's prior is $\ell \sim \mathcal{N}(0, 15^2)$ kg/s.

3.2 The probabilistic model

The rows of the previous subsection are all of one kind; this subsection states the general model they belong to. A physical network is a graph, and the engine writes its physics on that graph. The state is a vector z of unknowns: one potential per node, one flow per edge, and whatever else is uncertain, such as demands or parameters. Every relation the state must satisfy is a factor, written as a residual $r_k(z)$ that vanishes when the relation holds and carrying a weight w_k, the inverse variance of how far it may be off. A conservation law at a node, a constitutive law on an edge, a sensor reading and a prior on an approximately known quantity are all factors of this one kind; they differ only in their residual and their weight, a physics relation being trusted to a small tolerance, a reading to its sensor's noise, a prior to its stated spread. Treating each factor as an independent Gaussian belief about its residual gives the posterior density of the state given the readings y,

$$p(z \mid y) \propto \prod_k \exp\!\big(-\tfrac{1}{2}\, w_k\, r_k(z)^2\big). \tag{12}$$

The state that maximizes it, the *maximum-a-posteriori* (MAP) state z^*, minimizes the weighted sum of squared residuals $C(z)$ and is found by Gauss–Newton iteration on the sparse Jacobian J of the residuals, $J_{ki} = \partial r_k / \partial z_i$, with W the diagonal matrix of weights, from a flat start until the gradient norm falls below a tolerance ϵ:

$$C(z) = \tfrac{1}{2} \sum_k w_k\, r_k(z)^2, \tag{13}$$

$$z^* = \arg\min_z C(z), \tag{14}$$

$$z \leftarrow z - (J^\top W J)^{-1} J^\top W\, r(z) \quad \text{until } \|J^\top W\, r(z)\| < \epsilon. \tag{15}$$

Near the minimum the cost (13) is approximately quadratic, so the posterior is approximately Gaussian about z^* with information matrix Λ. This is the *Laplace approximation*, and the *posterior standard deviation* σ_i of any unknown is read from the diagonal of Λ^{-1}, from the same sparse factorization the last Gauss–Newton step used:

$$p(z \mid y) \approx \mathcal{N}\big(z^*, \Lambda^{-1}\big), \tag{16}$$

$$\Lambda = J^\top W J, \tag{17}$$

$$\sigma_i = \sqrt{(\Lambda^{-1})_{ii}}. \tag{18}$$

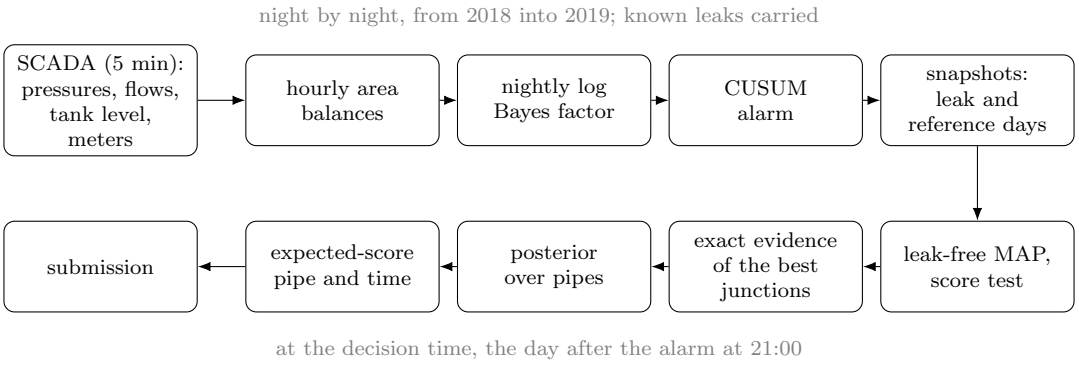


Figure 3: The causal pipeline. Every night the hourly inflow balances of two supply zones are compared with the same nights of earlier weeks; a CUSUM of the nightly log Bayes factors raises an alarm. At its decision time the alarm's area is localized with a joint hydraulic model of snapshots before and after the estimated start, and the pipe and the detection time with the largest expected score are submitted.

When several factor graphs compete to explain the same readings — a leak here, a leak there, no leak — each is compared by its *evidence* $Z = \int \prod_k \exp(-\frac{1}{2} w_k r_k(z)^2)\, dz$, the probability of the readings under the model once the extra freedom of its unknowns has been paid for. In the Laplace approximation the evidence has a closed form, and normalizing every factor by its own weight lets graphs with different numbers of unknowns d and factors compare as densities,

$$\log Z = -C(z^*) + \tfrac{d}{2} \log 2\pi - \tfrac{1}{2} \log\det \Lambda + \sum_k \tfrac{1}{2} \log \frac{w_k}{2\pi}. \tag{19}$$

Differences of $\log Z$ between hypotheses, from (19), are in *nats*, natural-log units of odds; five nats is odds of about 150 to 1. The posterior probability of a hypothesis is its evidence normalized over all hypotheses under equal prior odds. Everything that follows is a choice of variables and factors on this graph, followed by these three operations.

3.3 The pipeline

Figure 3 shows the order of operations: a detector that runs every night on the area inflow balances, and a localization that runs once per alarm on a joint hydraulic model. The next two subsections state each; the third states the submission.

3.4 Nightly detection

Between 01:00 and 05:00 demand is lowest and a leak is the largest part of the inflow. For each night n the five hourly inflows $y_{n,h}$ of a balance — Areas A and B together (the reservoir inflows less the pump flow), or Area C's unmetered part (the pump flow less the tank's fill, less the meters) — are compared with the reference profile D_h, the median of the same hours on the last 3 nights with the same day of the weekly demand pattern, each less the flow of the leaks the detector already carries. The residual $r_h = y_{n,h} - D_h$ is modeled with a demand scale for the night and a possible

new leak,

$$r_h = (s_n - 1)D_h + L_n + e_h, \tag{20}$$

$$s_n \sim \mathcal{N}(1, \sigma_s^2), \tag{21}$$

$$L_n \sim \text{HalfNormal}(\tau), \tag{22}$$

$$e_h \sim \mathcal{N}(0, \sigma_e^2), \tag{23}$$

and integrating the scale and the leak out of (20) gives the night's log Bayes factor of a new leak against none in closed form,

$$\Sigma = \sigma_e^2 I + \sigma_s^2 D D^\top, \tag{24}$$

$$P = \mathbf{1}^\top \Sigma^{-1} \mathbf{1} + \tau^{-2}, \tag{25}$$

$$c = \mathbf{1}^\top \Sigma^{-1} r, \tag{26}$$

$$\log B_n = \log 2 - \log \tau - \tfrac{1}{2} \log P + \frac{c^2}{2P} + \log \Phi\big(c/\sqrt{P}\big), \tag{27}$$

with Φ the standard normal distribution function. The noise levels were estimated on 2018 by median absolute deviations: $\sigma_e = 1.58$ m^3/h and $\sigma_s = 0.097$ for Areas A and B, 0.91 m^3/h and 0.061 for Area C; the night-mean residuals are correlated from night to night with coefficient ρ (0.90 and 0.11), and ρ times the previous night's mean residual is subtracted from r first. The leak scale is $\tau = 20$ m^3/h. The statistic (27) accumulates in a CUSUM,

$$S_n = \max\big(0,\ S_{n-1} + \log B_n - k\big), \tag{28}$$

which raises an alarm and restarts at zero when $S_n \geq h$. The values $k = 0.5$ and $h = 2$ maximize the 2018 expected score over a grid ($k \in \{0.5, 1, 2, 4\}$, $h \in \{2, 4, 8, 16, 32\}$), with detection power measured on leaks added to the 2018 night flows (twelve per balance, steps and three-week ramps, sizes drawn from the flow drops at the ten reported repairs) and false alarms counted as alarms without a reported repair in their balance within 120 nights. An alarm within 45 nights of the start of a known leak that is still growing is that leak's growth. Otherwise the alarm starts a known leak $a + b\,t$ ($b \geq 0$, t in days since its start), its start the most probable night under a step-plus-ramp posterior over the nights since the CUSUM last stood at zero. The known leak is refitted every night on the nights up to 7 before, so that a new leak in the last week stays in the residual, and in 2018 it is closed at a reported repair whose flow drop matches it.

3.5 The closure test

In 2019 no repairs are reported, so a leak the detector carries stays subtracted from later nights after it is fixed unless the readings themselves show that it ended. The closure test asks, every night and for each known leak k with flow $L_k > 0$, whether the night sits L_k below the reference. With the residual r of (20) against the reference less all known leaks, Σ of (24) and no night-to-night correction, the log likelihood ratio of that drop against an unchanged night and its CUSUM are

$$a = \mathbf{1}^\top \Sigma^{-1} \mathbf{1}, \tag{29}$$

$$\log D_{n,k} = -L_k\, c - \tfrac{1}{2} L_k^2\, a, \tag{30}$$

$$T_{n,k} = \max\big(0,\ T_{n-1,k} + \log D_{n,k} - k_c\big), \tag{31}$$

with c of (26). Consecutive nights are correlated, so T is a detection statistic with an empirical threshold rather than log-odds. When some $T_{n,k} \geq h_c$ the drop is estimated as $\hat{L} = -c/a$ with

standard deviation $1/\sqrt{a}$. The candidates are each leak whose statistic reached h_c and each pair of them; the candidate with the largest $\log D$ is closed if its flow lies within $3/\sqrt{a}$ of $\hat{L}$, and it is closed as of the night its statistic last stood at zero. A closure restarts every T of the balance, the CUSUM (28) and the night-to-night correction; a firing that matches no candidate closes nothing and restarts every T only. In 2018, where repairs are reported, a repair closes one open leak whose flow matches its drop: among those, the leak whose localization at its alarm put at least 0.05 of the posterior within 300 m of the repaired pipe, else the one nearest in flow; a repair within three nights after a closure by the test in its balance closes nothing more. Leaks carried into 2019 are submitted from the localization made at their alarm, with the no-leak probability recomputed on 2019-01-01. The thresholds k_c and h_c were chosen on 2018 with the repair report withheld from the detector (Section 4); the other settings are those of the frozen evaluation.

3.6 Localization at the decision time

An alarm on night n is localized at 21:00 on the next day, once the day's readings are in. The joint model holds hourly snapshots at 10:00, 13:00, 16:00, 19:00 on the alarm day and the next day, where the leak runs, and at the same hours one or more whole weeks earlier, before the estimated start, where it does not; 16 snapshots in all. The reference snapshots give each pressure sensor an offset o_i that absorbs whatever part of the model's error is constant in time, so that the leak is placed by the change in the readings rather than by their absolute level. The leak-free model is solved to its MAP and its regimes re-checked; each junction of the alarm's balance is then scored from that fit without re-solving, by the one-step Gauss–Newton estimate of a leak added to its conservation rows in the leak snapshots (a Rao score test),

$$g_j = -\sum_r w_r J_r^\top, \tag{32}$$

$$S_j = \sum_r w_r + \tau_\ell^{-2} - g_j^\top \Lambda_0^{-1} g_j, \tag{33}$$

$$\hat{\ell}_j = \frac{\sum_r w_r F_r}{S_j}, \tag{34}$$

$$\Delta \log Z_j \approx \tfrac{1}{2}\hat{\ell}_j^2 S_j - \tfrac{1}{2}\log(S_j \tau_\ell^2), \tag{35}$$

where r runs over junction j's conservation rows, F_r, w_r and J_r are their residuals, weights and Jacobian rows at the leak-free optimum, and Λ_0 is its information matrix. The 10 junctions with the largest gain (35) are solved exactly, one at a time and started from the leak-free optimum, and their evidence differences (19) replace the one-step values. A pipe's likelihood is the mean of its two end junctions' Bayes factors, the prior over pipes is uniform, and the posterior over pipes p_ℓ is scaled by $1 - p_0$, where p_0, the probability that the alarm is not a leak, is the probability that the balance's leak estimate is not positive given its standard error over the nights since the start (at least 0.05).

3.7 The submission

The official score of a detection at time t on pipe λ is, for a leak within 300 m along the network that runs at t, EUR 0.80 per m^3 of the leak's water from t to the end of the year, less EUR 500 times the distance over 300 m; any other detection costs EUR 500. The submitted pipe maximizes

the expected score under the posterior,

$$E[\lambda] = \sum_{\ell} p_{\ell}(1-p_0)\big[\mathbf{1}(d_{\lambda\ell} \le R)(V - C\,d_{\lambda\ell}/R) - \mathbf{1}(d_{\lambda\ell} > R)\,C\big] - p_0 C, \tag{36}$$

with $R = 300$ m, $C =$ EUR 500, d the benchmark's pipe-to-pipe distance and V the balance's leak estimate times the hours left in the year times EUR 0.80. The submitted time t_m (01:00 of a candidate start night m) maximizes $P(t_m \ge t_{\text{start}})\,V(t_m) - (1-P)\,C$ under the start posterior, and the detection is submitted when (36) is positive. Leaks the detector carries at the end of 2018 are localized on the last two days of 2018 and submitted at 2019-01-01 00:00.

3.8 Protocol

The method reads the nominal model, the 2018 readings and repair report, and the 2019 readings through a reader that refuses any row later than the current decision time; the ground truth is read only by the scoring. All settings — the thresholds, the noise model, the snapshot hours, the priors, the calibrated factors — were fixed on 2018 and recorded before 2019 was read. On 2018 the tank's area was fitted from the pump cycles with a prior on its nominal 16 m diameter, the pattern alignment from the meters, the noise model from the nights, the thresholds from injected leaks, and the conductance factors from 24 snapshots of 12 quiet days (no alarm within seven days and no known leak above 2 m^3/h); the whole pipeline was then run on 2018 with localization, and each localized alarm compared with the reported repairs.

2019 is evaluated twice. The first evaluation uses the settings fixed at the end of 2018 and is the blind result. The second adds a test for the disappearance of a known leak (Section 3.5). Its thresholds are chosen on 2018 alone, but its form responds to a failure seen in the first evaluation, so the second evaluation is not blind. Prototypes of the test were run on 2019 three times before the thresholds were fixed, and the ground-truth leak flows and end dates were read to diagnose the failure, including the rule that every closure restarts all disappearance statistics.

3.9 Metrics

The question of the Introduction is answered with the benchmark's own measures and with checks of the stated probabilities. The economic score, in euros, is the benchmark's summary of what a utility gains; its reference values are the perfect submission (EUR 523,154) and the six best published teams. The true positive rate, found leaks over the 23 leaks, and the false positive count separate sensitivity from cost. The distance from a submitted pipe to its leak, in meters along the network, is what a crew searches. The time from a leak's start to the submitted detection time, in days, is the delay the score charges through lost water. The Brier score of the stated probability that the submitted pipe is within 300 m of a running leak, the mean squared difference between that probability and the 0 or 1 outcome over all localized alarms, measures whether the probabilities can be trusted: 0 is perfect, and always stating 0.5 gives 0.25. Link coverage counts the alarms raised while a leak ran in their balance whose smallest set of pipes holding 90 % of the posterior contains that leak's pipe, reported with a Clopper–Pearson interval; a calibrated posterior covers in about 90 % of such alarms.

4 Results

Calibration on 2018. The pump cycles give the tank an area of 190.3 ± 0.4 m^2, an equivalent diameter of 15.57 m against the nominal 16 m, and the meters lead the nominal pattern by 25

Table 1: Calibrated conductance of Area A's pipe groups on 24 leak-free snapshots of 2018 (posterior mean ± posterior standard deviation, prior $\mathcal{N}(1, 0.1^2)$). The pressure misfit falls from 2.70 cm rms with the nominal pipes to 1.15 cm.

Pipe group	Pipes	Roughness factor f_g	Conductance $f_g^{1.852}$
Area A, $D \leq 100$ mm	580	0.862 ± 0.013	0.760
Area A, $150 \leq D \leq 160$ mm	113	1.066 ± 0.013	1.125
Area A, $D \geq 200$ mm	69	0.923 ± 0.004	0.863

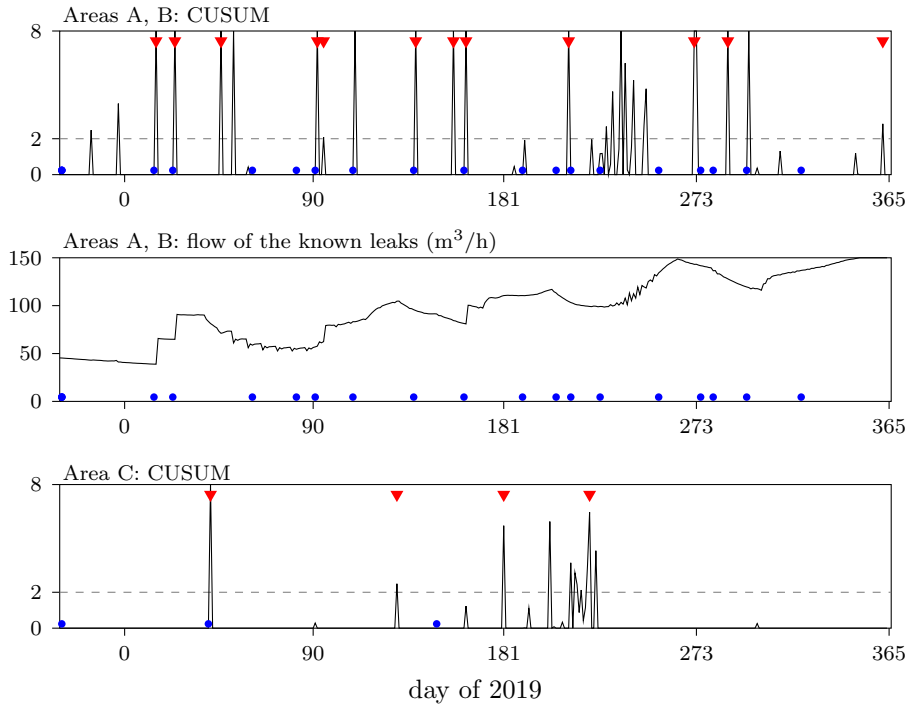
Table 2: The 2018 development run: every new-leak alarm, the balance's leak estimate at the alarm, the no-leak probability, the pipe the localization would submit, and the nearest pipe in the same balance whose repair the 2018 report lists after the alarm (*not connected*: the two pipes lie on different sides of a valve or pump).

Alarm night	Balance	Leak (m^3/h)	P(no leak)	Localized pipe	Next repair	Distance (m)
2018-01-22	C	4.2	0.05	p250	p31	520
2018-02-02	AB	5.9	0.13	p639	p232	346
2018-02-22	AB	3.8	0.27	p108	p461	145
2018-02-25	AB	4.7	0.24	p469	p461	205
2018-03-06	AB	20.1	0.05	p160	p673	227
2018-04-03	AB	8.0	0.05	p152	p158	not connected
2018-05-19	AB	44.4	0.05	p125	p538	290
2018-06-26	AB	0.0	0.50	p182	p183	748
2018-07-09	C	1.5	0.05	p292	p31	293
2018-08-12	AB	13.0	0.05	p217	p183	101
2018-10-06	AB	12.3	0.15	p714	p158	56
2018-10-24	AB	11.7	0.05	p706	p369	917
2018-12-16	AB	0.1	0.49	p849	–	not connected

minutes. Table 1 lists the calibrated groups: the small pipes carry 24 % less conductance than nominal and the large ones 14 % less, which is the direction of the diameter reduction the dataset's description allows, while the middle group carries more. The pressure misfit falls by a factor of 2.4, and the largest remaining misfit, 9.7 cm, is the error the sensor offsets absorb in localization. The calibration days all fall in January, the only period of 2018 quiet enough by the stated criterion, so the factors describe winter flows.

The 2018 development run. Table 2 lists the 13 new-leak alarms of 2018. The report of repaired leaks gives a check the ground truth does not: for 7 of the 12 alarms followed by a repair in their balance, the localized pipe lies within 300 m of such a repair. The alarms whose balance estimate is near zero receive a no-leak probability near one half, which keeps their expected score near or below zero.

The closure test on 2018. The thresholds of the closure test were chosen by running the detector on 2018 with the repair report withheld and matching its closures one to one with the 10 reported repairs: a closure matches in time when it fires between one night before and three nights after the repair, and in size when the closed leak's flow lies within three standard deviations $3/\sqrt{a}$ of the flow drop at the repair; a closure that matches no repair is false. Over $k_c \in \{0, 1, 2\}$ and $h_c \in \{2, 5, 10, 20, 40\}$, the cells within one timing match of the best were ranked by fewest false closures, then smallest median delay, then larger h_c, then smaller k_c. The selected $k_c = 2$, $h_c = 40$ matched 3 of 10 repairs in size (Wilson 95 % interval 11 % to 60 %) and 6 in time (31 % to 83 %), with 0 false closures; ten repairs, several of them merged with concurrent leaks, bound the



black: the CUSUM of the nightly log Bayes factor (clipped at 8; dashed: the threshold $h = 2$), and the flow of the leaks the detector carries; red triangles: alarms that started a new leak; blue dots: starts of the true leaks (post hoc; leaks begun in 2018 at the left edge)

Figure 4: The detection statistic through 2019. The top and bottom panels show the CUSUM of the two balances; the middle panel shows the flow of the leaks the detector carries in Areas A and B, which it subtracts from later nights. An alarm resets the CUSUM to zero.

thresholds only loosely. Widening Σ by each leak's intercept variance left the matches at 3 in size and 6 in time. With the report and the closure test, the 2018 run closed 3 leaks by the test and 6 at reported repairs, 1 of them by the localization's mass near the repaired pipe; 4 reported repairs matched no open leak.

Detection in 2019. Figure 4 shows the CUSUM (28) of both balances from December 2018 through 2019 with the alarms and, post hoc, the true leaks' starts. The detector raised 16 alarms that started a new leak (12 in Areas A and B, 4 in Area C) and carried 6 leaks from 2018 into 2019.

The official score. Table 3 sets the evaluation of the 13 submitted detections beside the published teams. The method found 7 of 23 leaks (30.4 %) with 4 false positives and scored EUR 128,805, 24.6 % of the perfect submission. Without the submissions for leaks standing since 2018 the score is EUR 113,061. The official economic score of this method (EUR 128,805) is lower than the best published score (Tongji, EUR 264,873). Its true positive rate (30.4 %) is lower than the best published rate (Under Pressure, 65.2 %). Its false positive count (4) is more than the fewest published (IRI, 1). The published teams read the whole evaluation year before submitting; this method did not, so the comparison favors them in timing and settings. With the closure test the method submitted 12 detections, found 7 leaks (30.4 %) with 5 false positives and scored

Table 3: The official BattLeDIM evaluation of 2019. The six published teams are the top six of the competition (Vrachimis et al., 2022, Table 4), who submitted during the competition with the 2019 SCADA in hand; this method processed 2019 causally with settings fixed on 2018. The row with the closure test is the second evaluation (Section 3.5), which is not blind. The perfect submission reports every leak at its own pipe at its start.

Submission	True positive rate	False positives	Score
This method	30.4 %	4	EUR 128,805
This method, without the standing-leak submissions	30.4 %	2	EUR 113,061
With the closure test	30.4 %	5	EUR 111,561
Tongji	56.5 %	3	EUR 264,873
Under Pressure	65.2 %	4	EUR 260,562
IRI	43.5 %	1	EUR 210,772
Leakbusters	47.8 %	7	EUR 195,490
Tsinghua	47.8 %	5	EUR 167,981
UNIFE	43.5 %	4	EUR 127,626
Perfect submission	100.0 %	0	EUR 523,154

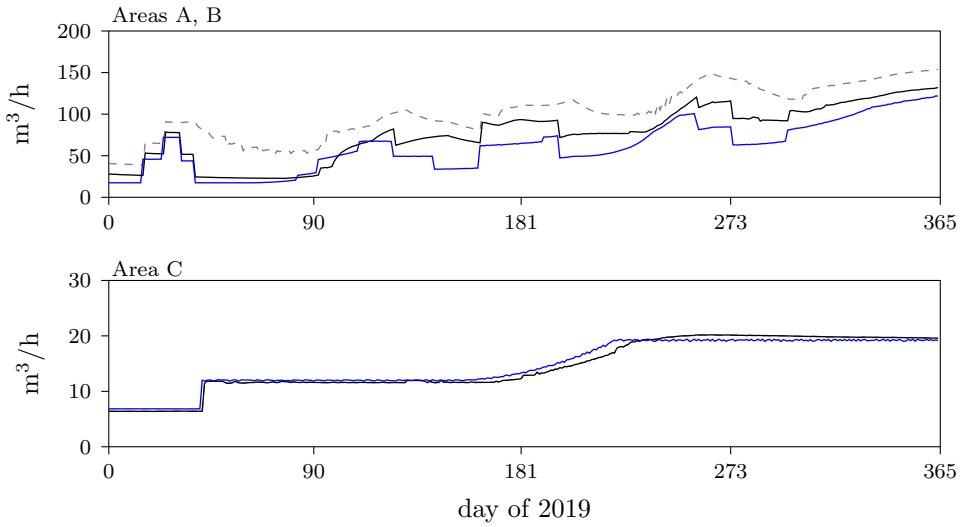
EUR 111,561. With the closure test, the official economic score (EUR 111,561) is lower than the best published score (Tongji, EUR 264,873). With the closure test, the true positive rate (30.4 %) is lower than the best published rate (Under Pressure, 65.2 %). With the closure test, the false positive count (5) is more than the fewest published (IRI, 1).

The flow the detector carries. Figure 5 sets the flow of the known leaks in both evaluations against the true leak flow at night. In December the frozen detector carried 149 m^3/h in Areas A and B and the detector with the closure test 127 m^3/h, against a true 115 m^3/h. The test closed 6 leaks in 2019; 6 of them fired within three nights of the end of a true leak in their balance (7 true leaks ended during the year), and 1 of those closed the known leak whose alarm that ending leak had raised. Another 7 firings matched no known leak or pair and closed nothing. Between June and October the detector carried on average 48.2 m^3/h more than the true leak flow in the frozen evaluation and 23.4 m^3/h more with the closure test; the true flow includes leaks that were never alarmed on, so neither figure is the flow of repaired leaks alone. With the closure test the method found the same leaks as the frozen evaluation. The submissions for leaks standing since 2018 added EUR 15,743 to the frozen score, where they were localized on the last two days of 2018, and EUR -1,000 with the closure test, where each reuses the localization made at its leak's alarm.

Leaks found and missed. Table 4 lists every leak. Found leaks were submitted a median of 0.2 days after their start (or after 2019-01-01 for standing leaks), at a median distance of 145 m. 4 submissions matched no running leak within 300 m.

A posterior over pipes. Figure 6 draws the posterior of the highest-scoring detection on the map: the mass concentrates on a few pipes around the submitted one, and the true leak's pipe lies within the radius the score allows.

Stated probabilities. Over the 19 localized alarms raised while a leak ran in their balance, the 90 % credible set of pipes contained that leak's pipe in 4 (21.1 %, Clopper–Pearson 90 % interval 7.5 % to 41.9 %), against about 90 % for a calibrated posterior. The Brier score of the stated



dashed gray: the flow of the known leaks in the frozen evaluation; black: with the closure test; blue: the true leak flow between 01:00 and 05:00 (post hoc)

Figure 5: The flow of the leaks the detector carries through 2019, as it stood on each night, against the true leak flow. A closure dated to earlier nights lowers the flow only from the night it fires, so the black line drops later than the leak ended.

probability of lying within 300 m is 0.293 over 19 localized alarms, against 0.25 for a constant one half.

Cost. Detection over the two years takes seconds; each localization solves one leak-free joint model of about 39,601 unknowns, one score test over the balance's junctions and 10 exact hypotheses, in 1.7 to 3.0 minutes on one thread.

5 Discussion

The largest failure is structural and visible in Figure 4. In 2018 the detector closed a known leak at a reported repair; in 2019 no repairs are reported, and 7 of the 23 leaks ended during the year, so every leak the detector alarmed on in Areas A and B stayed subtracted from later nights after the utility had fixed it. By December the detector carried 149 m^3/h of such flow. After each unreported repair the residual turned negative (in February the mean nightly residual was -27.8 m^3/h), the CUSUM stayed at zero, and a new leak had to exceed the missing flow before it could raise an alarm; 13 of the 16 leaks missed in Areas A and B started after February. The same negative residuals interact with the night-to-night correction of (28)'s input: subtracting ρ times a large negative previous residual turns the next night's residual positive, so 8 of the 23 alarms of 2019 in Areas A and B were raised on nights whose own residual was negative, among them the tallest spikes of the CUSUM and the alarm of 2019-02-16 whose submission was a false positive. The closure test addresses both effects only in part. With it, 5 alarms of 2019 in Areas A and B fell on nights with a negative residual, and the flow carried above the true flow between June and October fell from 48 to 23 m^3/h, yet the new leaks of Areas A and B were still missed. The remaining excess comes from known leaks whose flow the refit overestimates and that the test never closes, and from firings whose drop matched no single leak or pair; together with the leaks merged by the growth rule, it is consistent with keeping most new leaks below the threshold. The test does not

Table 4: Every leak running in 2019 with the outcome of this method: whether a submission within 300 m during its lifetime found it, the submitted pipe's distance, the time from the leak's start (or from 2019-01-01 for leaks begun in 2018) to the submitted detection time, and the detection's score; the last column is the score of the leak's detection with the closure test, or *no*.

Pipe	Balance	Start	End	Found	Distance (m)	Days to submitted time	Score	With the closure test
p257	C	2018-01-08	2019-12-31	yes	160	0.0	EUR 46,805	EUR 30,061
p427	AB	2018-02-13	2019-12-31	no	–	–	–	no
p654	AB	2018-07-05	2019-12-31	no	–	–	–	no
p810	AB	2018-07-28	2019-12-31	no	–	–	–	no
p523	AB	2019-01-15	2019-02-01	yes	49	0.1	EUR 8,778	EUR 8,778
p827	AB	2019-01-24	2019-02-07	yes	0	0.3	EUR 6,714	EUR 6,714
p280	C	2019-02-10	2019-12-31	yes	226	181.5	EUR 13,846	EUR 13,846
p653	AB	2019-03-03	2019-05-05	no	–	–	–	no
p710	AB	2019-03-24	2019-12-31	no	–	–	–	no
p514	AB	2019-04-02	2019-05-23	yes	145	0.2	EUR 14,751	EUR 14,751
p331	AB	2019-04-20	2019-12-31	no	–	–	–	no
p193	AB	2019-05-19	2019-12-31	no	–	–	–	no
p277	C	2019-05-30	2019-12-31	yes	158	31.1	EUR 22,574	EUR 22,574
p142	AB	2019-06-12	2019-07-17	yes	105	0.2	EUR 17,337	EUR 17,337
p680	AB	2019-07-10	2019-12-31	no	–	–	–	no
p586	AB	2019-07-26	2019-09-16	no	–	–	–	no
p721	AB	2019-08-02	2019-12-31	no	–	–	–	no
p800	AB	2019-08-16	2019-10-01	no	–	–	–	no
p123	AB	2019-09-13	2019-12-31	no	–	–	–	no
p455	AB	2019-10-03	2019-12-31	no	–	–	–	no
p762	AB	2019-10-09	2019-12-31	no	–	–	–	no
p426	AB	2019-10-25	2019-12-31	no	–	–	–	no
p879	AB	2019-11-20	2019-12-31	no	–	–	–	no

identify which leak ended: a drop is matched by size, so where two open leaks have similar flows the test can close the wrong one, which then keeps the other's location; 1 of 6 closures near a true end closed the right leak. Submissions are made at the alarm, so a wrong identity changes later nights' residuals but not an earlier submission. A leak carried into 2019 is submitted from the localization made at its alarm, made on the nights just after its start rather than on the last two days of 2018. The nights between a leak's end and its closure also bias the other leaks' refits. Leaks missed in both evaluations (p123, p193, p331, p426, p427, p455, p586, p653, p654, p680, p710, p721, p762, p800, p810, p879) include leaks that grow over months, which the reference nights absorb before the CUSUM accumulates, and leaks merged by the growth rule with a concurrent leak in the same area. The detector sees a leak only through its area's inflow at night. Area B has one pressure sensor, and PRV-3 fixes the head there, so a leak in Area B shows mainly as more flow through PRV-3 and a lower head at its inlet; Area C's pressures are held by the tank, so its balance, sharpened by the meters, is the signal; in Area A a leak of a few cubic meters per hour moves the gauged pressures by centimeters by day and millimeters at night, which is why localization uses daytime snapshots. Two leaks that start close in time in the same area are not separated by the balance: an alarm during another leak's growth is booked as that growth. The reference nights absorb slow seasonal change but also any leak that began more than three weeks before and was never alarmed, so leaks standing since 2018 are found only if the detector alarmed on them in 2018. The thresholds are at the most sensitive corner of the grid ($h = 2$ is its smallest value), because the 2018 false alarms were counted only as alarms without a later reported repair, a count that unreported leaks make too low. The posterior over pipes treats the model's error as the stated reading noise; where the error has structure the offsets do not absorb, the posterior is too sure of itself, and the coverage and Brier results of Section 4 are the check of how far. The published teams submitted after seeing the whole evaluation year; their scores are an upper reference for a causal method, not a like-for-like



red: posterior probability of each pipe (darker and wider is more probable; pipes below 10^{-4} gray); dashed: the submitted pipe p11; blue: the true leak's pipe p277 (post hoc)

Figure 6: The posterior over pipes at the decision time (2019-07-02 21:00) of the alarm of 2019-07-01 in Area C. The submitted pipe is 158 m from the true leak and the detection scored EUR 22,574.

comparison.

6 Conclusion

This report asked how much of the BattLeDIM benchmark's economic score a method earns when it calibrates on the historical year and then processes the evaluation year forward in time. A nightly flow-balance detector with a CUSUM of log Bayes factors and an evidence ranking of leak hypotheses in a joint hydraulic model of snapshots before and after each alarm found 7 of 23 leaks with 4 false positives and scored EUR 128,805, 24.6 % of a perfect submission. The official economic score of this method (EUR 128,805) is lower than the best published score (Tongji, EUR 264,873). Its true positive rate (30.4 %) is lower than the best published rate (Under Pressure, 65.2 %). Its false positive count (4) is more than the fewest published (IRI, 1). A second, not blind evaluation adds a test for the disappearance of a known leak, with thresholds chosen on 2018; it found 7 leaks with 5 false positives and scored EUR 111,561, lower than the frozen evaluation. With the closure test, the official economic score (EUR 111,561) is lower than the best published score (Tongji, EUR 264,873). With the closure test, the true positive rate (30.4 %) is lower than the best published rate (Under Pressure, 65.2 %). With the closure test, the false positive count (5) is more than the fewest published (IRI, 1). Calibration on 2018 lowered the pressure misfit from 2.70 to 1.15 cm, and on 2018 the localized pipe lay within 300 m of a later repair for 7 of 12 alarms. The stated probability of lying within 300 m scored 0.293 on the Brier scale over 19 alarms. The stated probabilities were overconfident. Without the closure test the detector lost sensitivity after leaks were repaired without a report, because it kept subtracting them; with it the excess carried flow is smaller but still hides most new leaks, and the test matches drops by size without telling which of two similar leaks ended. The next steps are a detector for leaks that grow over months, one that separates concurrent leaks in one area, and a localization whose noise model includes the

structured model error.

A Reproduction

```
cd docs/terra_graph_engine/case_studies/water_networks/battledim_l_town
python data/fetch_battledim.py --dest $HYDRAULICS_DATA_DIR
PYTHONPATH=../../../../../src python run.py --stage calibrate   # 2018: balances, tuning, pipes
PYTHONPATH=../../../../../src python run.py --stage develop     # 2018 with localization
PYTHONPATH=../../../../../src python run.py --stage detect      # 2019, causal, then the score
PYTHONPATH=../../../../../src python run.py --stage subset      # the committed excerpt
PYTHONPATH=../../../../../src python run.py --stage closure          # closure thresholds on 2018
PYTHONPATH=../../../../../src python run.py --stage develop_closure  # 2018 with the closure test
PYTHONPATH=../../../../../src python run.py --stage detect_closure   # 2019 with the closure test
PYTHONPATH=../../../../../src python run.py --stage truth_flow       # the true night leak flow (figure)
PYTHONPATH=../../../../../src python make_report.py --pdf
```

Each stage writes its section of `results.json`; `make_report.py` reads only that file. The injected leaks of the tuning use seed 20190101. The engine's linear algebra runs on one thread. The committed excerpt (2018-01-02 to 01-08 and 2019-01-07 to 01-20) reproduces the calibration of one week, the night statistics of a week of 2019 and a localization on the gate.

B Data and problems

The data are the BattLeDIM dataset (Vrachimis et al., 2020; Zenodo record 4017659, CC BY 4.0), fetched by file and checksum; an excerpt is committed with the study. Under `problems/`, five directories: *the three states of a pressure-reducing valve*; *the official score of a small leak submission*; *a leak alarm from nightly Bayes factors*; *which pipe to report for a leak*; *when a known leak has been repaired*. Each is a statement, a script that solves it and its results file.

References

[Vrachimis et al. 2022] Vrachimis, S. G., Eliades, D. G., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., Polycarpou, M. M. (2022). Battle of the Leakage Detection and Isolation Methods. Journal of Water Resources Planning and Management 148(12), 04022068. doi:10.1061/(ASCE)WR.1943-5452.0001601.

[Vrachimis et al. 2020] Vrachimis, S. G., Eliades, D. G., Taormina, R., Ostfeld, A., Kapelan, Z., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., Polycarpou, M. M. (2020). BattLeDIM: Battle of the Leakage Detection and Isolation Methods (dataset). Zenodo. doi:10.5281/zenodo.4017659. License CC BY 4.0.

[Puust et al. 2010] 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. doi:10.1080/15730621003610878.

[Rossman et al. 2020] Rossman, L. A., Woo, H., Tryby, M., Shang, F., Janke, R., Haxton, T. (2020). EPANET 2.2 User Manual. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-20/133.

[Irofti et al. 2025] Irofti, P., Romero-Ben, L., Stoican, F., Puig, V. (2025). Factor graph optimization for leak localization in water distribution networks. arXiv:2509.10982. doi:10.48550/arXiv.2509.10982.

[Han et al. 2018] Han, Q., Eguchi, R. T., Mehrotra, S., Venkatasubramanian, N. (2018). Enabling state estimation for fault identification in water distribution systems under large disasters. In: 2018 IEEE 37th Symposium on Reliable Distributed Systems (SRDS), pp. 161–170. doi:10.1109/SRDS.2018.00027.

[Dellaert and Kaess 2017] Dellaert, F., Kaess, M. (2017). Factor graphs for robot perception. Foundations and Trends in Robotics 6(1–2), 1–139. doi:10.1561/2300000043.