Bayesian leak localization in water distribution networks from a day of sparse pressure and demand readings

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

September 2026

Abstract

Leaks in water distribution networks waste treated water and the energy spent on it and open paths for contamination, and locating one from the few pressure and flow sensors a utility has is a long-standing problem. This study asks what one set of conservation and Hazen–Williams rows does with a single addition, a leak variable shared across T snapshots of a day while every snapshot keeps its own demand unknowns, so that a constant outflow is told apart from demand that varies. The data are the two public benchmarks of Irofti, Romero-Ben, Stoican and Puig (2025), who localize leaks with two chained factor graphs over a time window (FGLL) and publish their numbers. On the nine-junction T-Example (3 pressure gauges of 5 mm, 3 demand meters of 1 %, one supply meter, demands known to 0.5 %), 81 cases (9 leaks of 0.5 L/s × 9 noise seeds) at $T = 24$ put the true node top in 63 and top or in an exact tie with its dead-end partner in 81, exclude the leak-free model by at least 4.0 nats in every case (5 nats in 79), and size the leak to 0.029 kg/s RMSE with 80 of 81 intervals covering. The posterior mass on the true node rises from 0.42 at $T = 1$ to 0.75 at $T = 24$; a leak stops being identifiable when the per-snapshot demand prior widens past a few percent (265 nats at 0.5 %, 1.4 at 10 %); a 10 % roughness error biases the leak by −5.478 kg/s when fixed and by −0.001 kg/s after a leak-free-day calibration. On Modena (268 junctions, 20 pressure and 40 demand sensors, 32 scenarios of about 5 L/s), after a roughness calibration on the leak-free day, a score test from the leak-free posterior screens every junction in one solve and the exact evidence ranks a shortlist: the best candidate is 0.48 km from the leak on average (13 exact hits), the five best 0.55 km, and the sensors-only head RMSE is 0.383 m. This study's best-candidate distance (0.48 km, mean over 32 scenarios) is better than the 0.72 km Irofti et al. report for FGLL. Its average-of-five distance (0.55 km) is better than the 0.78 km they report. Its head RMSE (0.383 m, sensors only) is better than the 3.29 m they report. The comparison is on their metric on their data; their code itself could not run here (no GTSAM wheel for this platform), and Modena's posterior probabilities are overconfident because the generating model's parameter noise is systematic. Every number costs 3274 engine solves.

1 Introduction

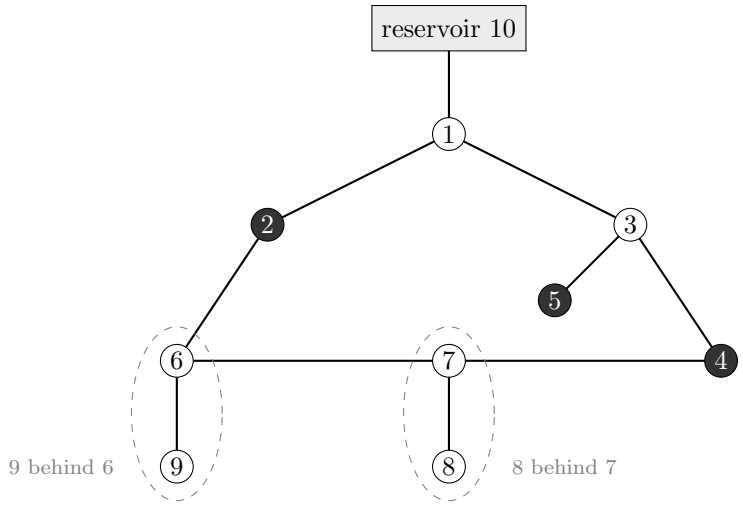
Water distribution networks carry treated water through buried pipes that, in most cities, are older than the instruments watching them. Leaks lose a substantial share of the water that has been pumped and treated, and with it the energy and chemicals spent on it; a leaking pipe under low pressure is a path for contamination, and an unlocated leak weakens supply exactly when droughts or disasters make water scarce. Finding a leak quickly from the few pressure gauges and flow meters a utility can afford is therefore a long-standing problem with environmental, economic and public-health stakes. Puust, Kapelan, Savic and Koppel (2010) review the methods of leakage

management from the hydraulic side; the Battle of the Leakage Detection and Isolation Methods (Vrachimis et al., 2022) put the research community's methods on one shared data set and showed how far apart their answers can be. The difficulty is structural: a network has hundreds of junctions and a handful of sensors, its demands are known only from patterns, and a leak of a few liters per second changes the gauged heads by centimeters.

Methods fall into a few families. Model-based methods compare measured heads with a calibrated hydraulic model and locate the leak where the residuals point (the sensitivity- and residual-based approaches that the GHR-S baseline below represents). Data-driven methods interpolate the hydraulic state at unmeasured nodes from the sensors, as graph-signal interpolation (GSI) does, and state estimators such as the unscented Kalman filter refine that interpolation with the network's physics (UKF-AW-GSI; Romero-Ben, Irofti, Stoican and Puig, 2026). Factor-graph methods write the physics itself as the estimator: Han, Eguchi, Mehrotra and Venkatasubramanian (2018) showed that a water network is a factor graph on which state estimation and fault identification run at scale under disaster damage, and Irofti, Romero-Ben, Stoican and Puig (2025) chained two factor graphs for leak localization — a leak-free state estimation over a time window, then a localization graph on the pressure residuals with a persistence factor — and compared it with GSI, UKF-AW-GSI and GHR-S on a synthetic nine-junction network, on Modena and on L-Town, releasing the data. Both factor-graph methods rest on the sparse nonlinear least-squares machinery that Dellaert and Kaess (2017) describe for robotics.

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 water network's mass balances and Hazen–Williams pipe laws, joins several instants of a day into one model that shares a single leak variable, ranks one leak hypothesis per junction by how well it explains the readings, screens the hundreds of candidates of a large network before the exact comparison, and asks which added sensor would help most and whether its own uncertainty estimates can be trusted. Section 3 states the framework and each of these constructs precisely.

This study takes Irofti et al.'s data and asks a question that is narrower than theirs and older than both papers: given a hydraulic model, readings at a few nodes over a day, and demands known only approximately, which node leaks, how much, and how sure can one be? The construction is one joint model: T copies of the network, one per snapshot, each with its own heads, flows and demand unknowns, sharing one leak variable. The leak is then a hypothesis — one per junction, plus none — compared by the normalized Laplace evidence, so that the output is a posterior probability per node rather than a score. Nothing here is new as a construction; Irofti et al.'s persistence factor plays the role the shared variable plays, and Bayesian model comparison of leak hypotheses predates both. What the study adds is the accounting: calibrated intervals on the leak size, the exact loss of identifiability as demand uncertainty grows, the effect of a roughness error and its repair, the next sensor by expected information gain, two simultaneous leaks, and, on Modena, a screening step that makes the exact comparison affordable on 268 candidates. Section 2 states the data and the model, Section 3 the methods, Section 4 the results, Section 5 their limits, and the appendices the reproduction recipe, the data and the references.



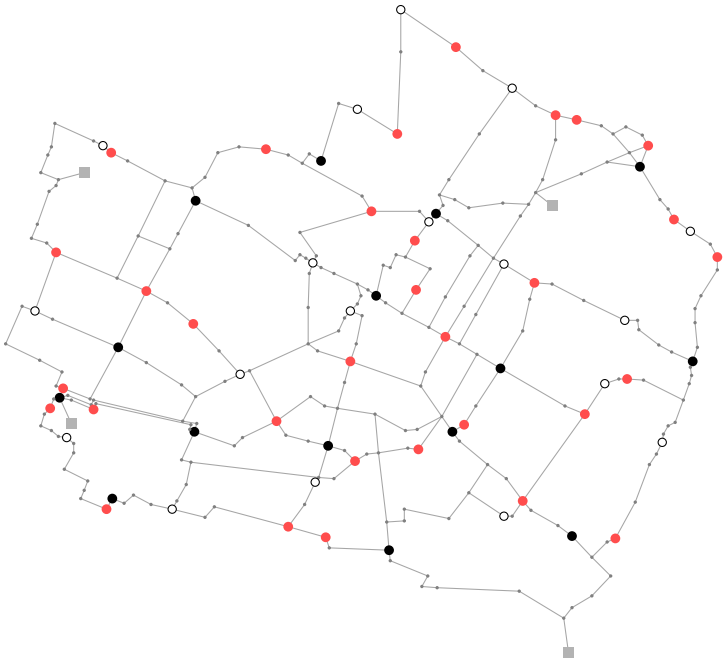
filled: pressure and demand sensor; dashed: the two pairs the sensors cannot separate

Figure 1: Irofti et al.'s T-Example: 9 junctions, one reservoir at 100 m, 10 pipes (C = 140), sensors at junctions 2, 4, 5 and at the reservoir, 288 five-minute steps per day, one 0.5 L/s leak per scenario. Junctions 8 and 9 are dead ends behind unmetered pipes, so a leak at 8 is the same hypothesis as a leak at 7, and 9 the same as 6.

2 System and data

The data. Figure 1 draws the T-Example. The authors ship, for each of nine scenarios, a leak-free day and a leaky day of heads and demands at every node in 5-minute steps, with a constant 0.5 L/s leak at one junction (about 0.6 % of the mean inflow); the pipe list, lengths, diameters and roughness; and the sensor sets: pressure and demand at junctions 2, 4, 5 and at the reservoir, whose demand meter is the supply. A head is the water's energy per unit weight at a node, in meters; a demand is the outflow a junction delivers; the leak is an extra outflow. The network is rebuilt from those arrays (no INP file ships) and its leak-free day is reproduced by the plugin to 0.2 mm. The data carry no noise; this study draws it: 5 mm on heads, 1 % on meters (floor 0.01 kg/s), a fresh draw per seed. Modena (Figure 2) is the same file format at 268 junctions with hourly samples, 20 pressure and 40 demand sensors, and 32 scenarios of 4.5–6.5 L/s; the authors put 1 % noise on roughness and diameters and 0.5 % on the demand patterns of its generating model. Everything is simulated data released under the ISC license.

The model. Figure 3 shows the equations around one node of one snapshot. For a window of T snapshots (spread evenly over the day), the joint model holds T copies of the network: heads at the junctions, signed mass flows on the pipes, and one demand variable per junction with a Gaussian prior centered on that snapshot's nominal demand and 0.5 % of it as standard deviation (the demand-pattern noise the authors state for their data). A hypothesis "leak at j" adds one variable, the leak outflow, to junction j's conservation row in every copy, with a zero-mean prior of 2 kg/s; where j carries a demand meter the meter reads demand plus leak, an exact auxiliary row. Physics rows are weighted at 1 g/s and 1 mm. The leak-free model is the hypothesis with no leak variable.



squares: the 4 reservoirs (the other 4 pressure and demand sensors sit there); filled circles: the 16 pressure sensors at junctions (all also demand sensors); open circles: the other 20 demand sensors; red: the 32 leak nodes of the scenarios

Figure 2: Modena, drawn from the coordinates in the authors' file: 268 junctions, 4 reservoirs, 317 pipes, hourly samples over 75 steps, sensors at 20 pressure and 40 demand nodes, and the 32 single-leak scenarios of about 5 L/s.

The questions. (1) Which junction leaks, and with what posterior probability, over the nine leaks and nine noise seeds? (2) How does that probability grow with the number of snapshots? (3) How wide may the demand prior be before the leak dissolves into it? (4) What does a roughness error do, and does a leak-free day repair it? (5) Which pressure sensor to add next; can two simultaneous leaks be told apart from one; is the Laplace posterior trustworthy? (6) On Modena, how far is the best candidate from the leak, against the published methods?

3 Methods

3.1 The hydraulic factor graph

For a water network the nodes are junctions and reservoirs and the edges are pipes. Each junction carries a head unknown h_i (the water's energy per unit weight, in meters), each pipe a mass-flow unknown m_e (kg/s), and each junction a demand unknown d_i (kg/s), the outflow it delivers to its consumers. Two physical laws tie them together, each written as a residual, an expression that vanishes when the law holds and carries a weight, the inverse variance of how far it may be off; Section 3.2 makes this the general rule. Conservation at junction i says that what flows in flows out or is consumed, and the Hazen–Williams law on pipe $e = (i, j)$ says that the head drop along

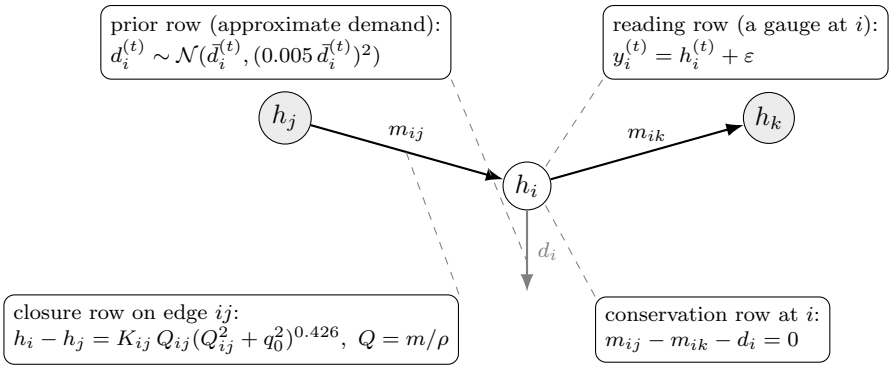


Figure 3: The model around one node of one snapshot t. Unknowns are the head, the mass flows on the links and the demand of that snapshot; the equations are conservation at the node, the Hazen–Williams closure on each link, a reading row per sensor and a prior row per demand. A leak hypothesis adds one more unknown, the leak, to the conservation row of its node in every snapshot of the window; it is the only unknown the snapshots share, and its constancy is what separates it from demand.

the pipe is a fixed function of its flow,

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

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

$$K_e = \frac{10.667\, L_e}{C_e^{1.852} D_e^{4.871}}, \tag{3}$$

with $Q_e = m_e/\rho$ the volumetric flow in m^3/s, $\rho = 1000$ kg/m^3, K_e from the pipe's length L_e, diameter D_e and roughness C_e (SI units, Q in m^3/s), and $q_0 = v_0 A_e$, $v_0 = 10^{-3}$ m/s, a small regularizing flow that keeps the law smooth through zero flow. Reservoirs are fixed heads. The physics residuals are weighted at $(1\ \mathrm{g/s})^{-2}$ and $(1\ \mathrm{mm})^{-2}$, so the physics is trusted far more than any reading. The sensors and the prior knowledge attach to the same unknowns as residuals of the same form: a pressure gauge at i, a demand meter at i and the prior that a demand stays close to its nominal pattern $\bar{d}_i$ each contribute one,

$$r = y_i - h_i, \tag{4}$$

$$r = y_i - d_i, \tag{5}$$

$$r = d_i - \bar{d}_i, \tag{6}$$

weighted by σ_h^{-2}, σ_d^{-2} and $\sigma_{\bar{d}}^{-2}$ respectively, with $\sigma_h = 5$ mm on the T-Example and 10 mm on Modena, $\sigma_d = 1\,\%$ of the reading with a floor of 0.01 kg/s, and $\sigma_{\bar{d}} = 0.5\,\%$ of the nominal demand with a floor of 0.005 kg/s, the demand-pattern noise the authors state for their data. Figure 3 is this graph around one junction.

3.2 The probabilistic model

The residuals of the previous subsection are all of one kind, and the engine treats them as such; 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,

$$p(z \mid y) \propto \prod_k \exp\!\big(-\tfrac{1}{2}\, w_k\, r_k(z)^2\big). \tag{7}$$

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:

$$C(z) = \tfrac{1}{2} \sum_k w_k\, r_k(z)^2, \tag{8}$$

$$z^* = \arg\min_z C(z), \tag{9}$$

$$z \leftarrow z - (J^\top W J)^{-1} J^\top W\, r(z) \quad \text{until } \|J^\top W\, r(z)\| < \epsilon. \tag{10}$$

Near the minimum the cost (8) 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:

$$p(z \mid y) \approx \mathcal{N}\big(z^*, \Lambda^{-1}\big), \tag{11}$$

$$\Lambda = J^\top W J, \tag{12}$$

$$\sigma_i = \sqrt{(\Lambda^{-1})_{ii}}. \tag{13}$$

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,

$$\log Z = -C(z^*) + \tfrac{d}{2} \log 2\pi - \tfrac{1}{2} \log\det \Lambda + \sum_k \tfrac{1}{2} \log \frac{w_k}{2\pi}. \tag{14}$$

Differences of $\log Z$ between hypotheses, from (14), 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 A window of snapshots

A single instant cannot separate a leak from a demand at the same junction: both are outflows in the same conservation factor. What separates them is time. A leak is constant across the day while

demands vary, so the model holds T copies of the network, one per instant t, each with its own heads, flows and demands and its own physics, reading and prior factors, together with a small set of shared variables θ that appear once and enter every copy's factors,

$$p(z^{(1)}, \dots, z^{(T)}, \theta \mid y) \propto p(\theta) \prod_{t=1}^{T} \prod_{k} \exp\big(-\tfrac{1}{2}\, w_k\, r_k(z^{(t)}, \theta)^2 \big). \tag{15}$$

In the leak hypotheses θ in (15) is the leak; in the roughness calibration it is the set of pipe roughnesses. The shared variable is the only one adjacent to all T copies, and that adjacency is what identifies it. The instants are spread evenly over the day, $T = 24$ in the main runs and 6 on Modena.

3.4 Leak hypotheses

Hypothesis M_j, a leak at junction j, adds one shared variable, the leak outflow ℓ, with prior $\ell \sim \mathcal{N}(0, \tau^2)$ and $\tau = 2$ kg/s on the T-Example, 3 kg/s on Modena, to junction j's conservation factor in every copy; the leak-free model M_0 has no such variable. Where j carries a demand meter, the meter reads what actually leaves the junction, demand plus leak, and its factor says so, so the meter cannot hide the leak,

$$r_j^{(t)} = \sum_{e \to j} m_e^{(t)} - \sum_{j \to e} m_e^{(t)} - d_j^{(t)} - \ell, \tag{16}$$

$$r = y_j^{(t)} - \big(d_j^{(t)} + \ell \big). \tag{17}$$

Every hypothesis is solved to its MAP state, from heads at the reservoir level and flows of 1 kg/s to a gradient norm below 10^{-8} (below 10^{-6} on the weakly constrained Modena estimation, at most 500 iterations), and scored by the normalized evidence (14); the posterior probability of each hypothesis is its evidence normalized under equal prior odds,

$$P(M_j \mid y) = \frac{Z_j}{\sum_k Z_k}. \tag{18}$$

Two hypotheses whose evidence agrees to 10^{-6} nats are indistinguishable to the data — on the T-Example a dead end and the junction it hangs from, since no sensor sits between them — and are ordered by name.

3.5 Screening before the exact comparison

Modena has 268 candidate junctions, and solving that many joint models per scenario would be slow. The leak-free optimum z_0^* already carries most of the answer: adding a leak variable at junction j and taking one Gauss–Newton step of the augmented system, eliminated onto the new variable against the cached factorization of Λ_0, gives the leak the junction would take, its information and

the evidence it would gain, a Rao score test,

$$g_j = -\sum_r w_r J_r^\top, \tag{19}$$

$$S_j = \sum_r w_r + \tau^{-2} - g_j^\top \Lambda_0^{-1} g_j, \tag{20}$$

$$\hat{\theta}_j = \frac{\sum_r w_r F_r}{S_j}, \tag{21}$$

$$z_j = |\hat{\theta}_j| \sqrt{S_j}, \tag{22}$$

$$\Delta \log Z_j \approx \tfrac{1}{2}\, \hat{\theta}_j^2 S_j - \tfrac{1}{2} \log(S_j \tau^2), \tag{23}$$

where r runs over j's conservation factors in the T copies, F_r and w_r are their residuals and weights at z_0^*, J_r their rows of the null Jacobian and $\partial r / \partial \ell = -1$. This costs one sparse solve per junction. It is exact for a linear model and, on the unmetered junctions of the T-Example, within 0.05 nats of the exact evidence; on metered junctions it over-scores, because the exact hypothesis also changes what the meter reads. It therefore shortlists — the 20 best unmetered and 5 best metered candidates — and the exact evidence (14) decides.

3.6 The next sensor and the validity of the posterior

Two further questions use the same machinery. Which pressure gauge to add next is answered by the expected information gain: each of the five leading hypotheses predicts the candidate reading y' as a Gaussian, with mean the head at its MAP and variance the sensor noise plus the head's posterior variance, and the gain is the expected fall of the entropy H of the hypothesis posterior once y' is seen,

$$p(y' \mid M_j, y) = \mathcal{N}\big(y';\ a_j^\top z_j^*,\ \sigma_h^2 + a_j^\top \Lambda_j^{-1} a_j\big), \tag{24}$$

$$P(M_j \mid y, y') \propto P(M_j \mid y)\, p(y' \mid M_j, y), \tag{25}$$

$$\mathrm{EIG}(y') = H\big[P(M \mid y)\big] - \mathbb{E}_{y'}\, H\big[P(M \mid y, y')\big], \tag{26}$$

the expectation taken by Gauss–Hermite quadrature on each component of the predictive mixture. Whether the Gaussian picture of the posterior can be trusted is tested on the two leading hypotheses in two ways: the exact cost is profiled along the leak, the other unknowns re-optimized at each fixed ℓ, at one and two posterior standard deviations from ℓ^* and compared with the quadratic $k^2/2$ the Gaussian predicts for $k = \pm 1, \pm 2$, and 64 samples drawn from the Gaussian are reweighted by the true cost, the weights summarized by their effective sample size,

$$\Delta C(k) = C\big(z^*_{\ell^* + k\sigma_\ell}\big) - C(z^*), \tag{27}$$

$$\omega_s = \frac{\exp(-C(z_s))}{\mathcal{N}(z_s;\ z^*, \Lambda^{-1})}, \tag{28}$$

$$\mathrm{ESS} = \frac{(\sum_s \omega_s)^2}{\sum_s \omega_s^2}, \tag{29}$$

where z^*_ℓ is the optimum over the other unknowns at fixed ℓ and z_s are the samples. Two simultaneous leaks are a hypothesis set of the leak-free model, nine singles and 36 pairs, on a leaky day synthesized by the plugin's own forward solve, since the data hold single leaks only.

3.7 Roughness

A wrong pipe roughness changes the head drop the model expects for a given flow, and a leak search can book the difference to a leak. Two treatments of a Hazen–Williams C set 10 % too low are compared: left fixed, or calibrated once on a leak-free day, every C_e a shared variable across the T copies with a prior about its nominal value $\bar{C}_e$,

$$r = C_e - \bar{C}_e, \tag{30}$$

weighted by σ_C^{-2} with $\sigma_C = 10\,\%$ of $\bar{C}_e$ in the T-Example experiment and 2 % on Modena, and then held at its MAP in the leak hypotheses. Freeing the roughnesses inside every leak hypothesis was also tried and is not reported: on the C-parameterized closure Gauss–Newton overshoots C through zero and stalls; a $\log C$ parameterization is the remedy. On Modena the nominal model misses the leak-free day by 13.6 cm rms — the authors put 1 % noise on roughness and diameters in the generating model — so the calibration is the first step there.

3.8 The Modena protocol

Each Modena scenario is run the same way. The roughnesses are calibrated on 6 leak-free snapshots. Each leaky-day head reading is then corrected by the calibrated model's own error at that sensor and step on the leak-free day (the authors' residual methods use the leak-free day at every node; this uses it at the sensors only), and its noise is widened by 5 % of its change from the leak-free day, the generating model's parameter noise in head-loss terms,

$$\tilde{y}_{s,t} = y^{\text{leak}}_{s,t} + \left(\hat{h}^{\text{cal}}_{s,t} - y^{\text{free}}_{s,t}\right), \tag{31}$$

$$\sigma^2_{s,t} = (1\ \text{cm})^2 + \left(0.05\,|y^{\text{leak}}_{s,t} - y^{\text{free}}_{s,t}|\right)^2, \tag{32}$$

with s a sensor, t a step, $\hat{h}^{\text{cal}}$ the calibrated model's leak-free-day head and y^{free} the leak-free-day reading. The leak-free model is solved on 6 leaky snapshots, every junction is screened, the shortlist's exact hypotheses are solved warm-started from the leak-free optimum, and the shortest-path distance, from the file's own distance matrices, from the true leak to the best candidate and the mean over the five best are recorded. The leak-free head estimation of the authors' Table IV is repeated with the sensors alone, the metered supply minus the metered demands split evenly over the unmetered junctions as a 100 % prior.

3.9 Metrics

The questions of Section 2 are answered with measures chosen so that each can be read against a reference. Localization is judged first by the rank of the true junction when the hypotheses are ordered by evidence (14), the operational quantity since a crew inspects candidates in rank order, with exact ties counted as ties rather than as successes or failures; then by the posterior probability of the true junction (18), whose mean over cases should match the fraction of cases the truth tops if the probabilities are calibrated (on Modena, where the model error is systematic, it is overconfident and the rank is the deliverable); and, on Modena, by the measures Irofti et al. publish, the shortest-path distance along the pipes from the true leak to the best candidate and the mean over the five best, in km and in pipes, averaged over the 32 scenarios. Along-pipe distance is what a repair crew walks, the five-best average asks whether the shortlist as a whole points to the right neighborhood, zero is an exact hit, and the published FGLL means, 0.72 and 0.78 km, are the reference. Detection is judged by the evidence gain of the true hypothesis over the leak-free model, the difference of

Table 1: The 81 T-Example cases at $T = 24$, by leak node: metered node or not; cases with the true node top; cases where it is second in an exact tie with its dead-end partner; mean posterior mass on it; the smallest evidence gain over the leak-free model across the 9 seeds; and the mean MAP leak with its mean posterior standard deviation (true leak 0.5 kg/s). Totals: top in 63, top or exact tie in 81, in the top two in 81; smallest gain 4.0 nats; 80 of 81 leak intervals ($\pm 2\sigma$) cover the truth.

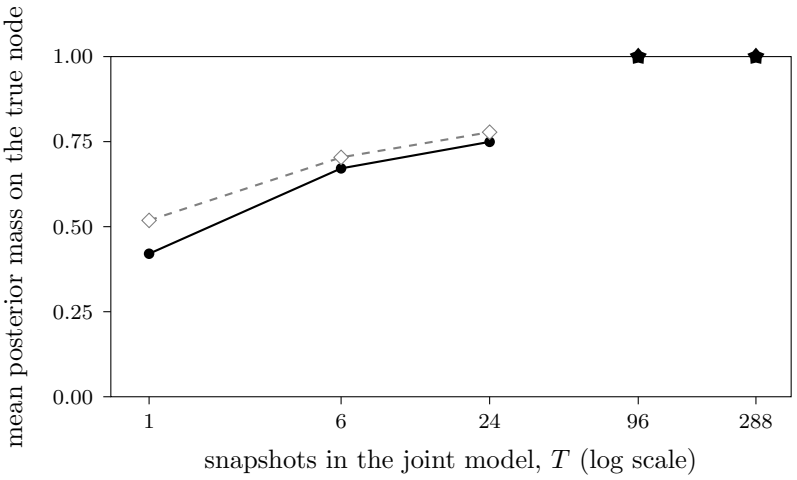
leak	metered	top	tie	$\bar{P}$(true)	min gain [nats]	leak [kg/s]
1	no	9/9	0/9	0.895	4.0	0.508 ± 0.109
2	yes	9/9	0/9	1.000	150.6	0.502 ± 0.027
3	no	9/9	0/9	1.000	85.7	0.508 ± 0.035
4	yes	9/9	0/9	1.000	233.8	0.504 ± 0.022
5	yes	9/9	0/9	1.000	1,250.3	0.504 ± 0.010
6	no	9/9	0/9	0.488	119.4	0.511 ± 0.031
7	no	9/9	0/9	0.490	140.2	0.509 ± 0.029
8	no	0/9	9/9	0.488	140.9	0.510 ± 0.029
9	no	0/9	9/9	0.490	120.7	0.514 ± 0.031

their values of (14) in nats, five nats (odds of about 150 to 1) being the pre-committed floor below which a leak is not called detected. Estimation is judged by the root-mean-square error of the MAP leak in kg/s and by the number of cases whose true leak lies within two posterior standard deviations (13) of the estimate, a calibrated Gaussian covering 95.45 % or about 77 of 81 cases, and, on Modena, by the head RMSE at the unmeasured junctions as in Irofti et al.'s Table IV, against the published 3.29 m. The screening is judged by the rank of the true junction in the score-test ordering (22) of all junctions, which says whether the shortlist of 25 contains the truth. Cost is counted in engine solves — MAP solves, Newton solves and screenings — rather than timed, so that it does not depend on the machine.

4 Results

Leak node on the T-Example. Table 1 summarizes the 81 cases. The true node is top in 63 and top or in an exact tie in 81; it is never below second. The runner-up in the tied cases is the dead-end partner at exactly the same evidence, ordered by name. The leak-free model loses by at least 4.0 nats (seed 6, leak at 1) and by 5 nats or more in 79 of 81 cases, so detection is never in doubt; the smallest gains belong to the junction fed directly by the reservoir, whose leak lowers the gauged heads by millimeters. The leak estimates average a bias of 0.009 kg/s with an RMSE of 0.029 kg/s against a mean posterior standard deviation of 0.036 kg/s; 80 of 81 $\pm 2\sigma$ intervals cover the truth, against about 77 expected of a calibrated Gaussian, so the interval is honest to slightly conservative. The mean posterior mass on the true junction, 0.76, is close to the fraction of cases it tops, 0.78, which is the mark of a calibrated probability. Metered junctions are found with near certainty (their meter reads demand plus leak), the others by the pattern of gauged heads and the supply meter.

Number of snapshots. Figure 4 shows the posterior mass on the true node against T. One snapshot gives 0.42 on average and 14 of 27 tops; 24 give 0.75 and 21. The leak's posterior standard deviation falls from 0.196 to 0.036 kg/s, and the evidence gain over the leak-free model grows from 5.6 to 276.6 nats. For the hardest leak (junction 1, fed directly by the reservoir) the two long



circles: mean P(true) over 27 cases; diamonds: fraction with the true node top (a dead-end tie counts against it); stars: leak at 1, three hypotheses only

Figure 4: Posterior mass on the true node against the number of snapshots. 9 leaks × 3 noise seeds at $T = 1, 6, 24$; the two long windows ($T = 96, 288$) rank only the true node, its best competitor at $T = 24$ and the leak-free model for one leak, because a full ranking costs 11 s per hypothesis at $T = 288$ (8,065 unknowns).

Table 2: Roughness mismatch as a nuisance (nine leaks, seed 0, $T = 24$). The estimator's Hazen–Williams C is set 10 % below the true 140 and either left fixed or calibrated once (every C a shared variable with a ±10 % prior) on the leak-free day (calibrated mean $C = 132.8$, range 123.3–142.8). Columns: true node top and in the top two, mean posterior mass, mean evidence gain over no leak, mean bias of the leak estimate and its mean posterior standard deviation.

roughness	top	top-2	$\bar{P}$(true)	gain [nats]	bias [kg/s]	σ [kg/s]
true C	7/9	9/9	0.775	265.5	0.003	0.036
C 10 % low, fixed	1/9	2/9	0.111	11,522.6	−5.478	0.036
C 10 % low, calibrated on the leak-free day	7/9	9/9	0.768	261.3	−0.001	0.036

windows give 1.000 at $T = 96$ and 1.000 at $T = 288$ against its competitor 3 and the leak-free model.

Demand uncertainty. A leak at a node is, to the physics, a demand increase at that node; only its constancy across snapshots separates the two. Figure 5 widens the per-snapshot demand prior (6): the mean evidence gain of the true hypothesis falls from 265.5 nats at 0.5 % to 1.4 at 10 %, the leak-free model is excluded by 5 nats in 9, 8, 3 and 1 of nine leaks, and the true node is top in 7, 7, 7 and 4. This is the mechanism behind the earlier Net3 result, where 30 % demand priors hid 5 kg/s leaks: not a solver limit but the information the readings carry.

Roughness. Table 2 compares the three treatments. With the true C the estimate is unbiased (0.003 kg/s). Fixed 10 % low, the model needs less flow to produce the observed head drops, and the deficit against the metered supply is booked to the leak as a negative outflow: bias −5.478 kg/s, true node top in 1 of nine. Calibrating once on the leak-free day gives bias −0.001 kg/s and 7 tops with the original width. The three sensors identify the ten roughnesses only in combination

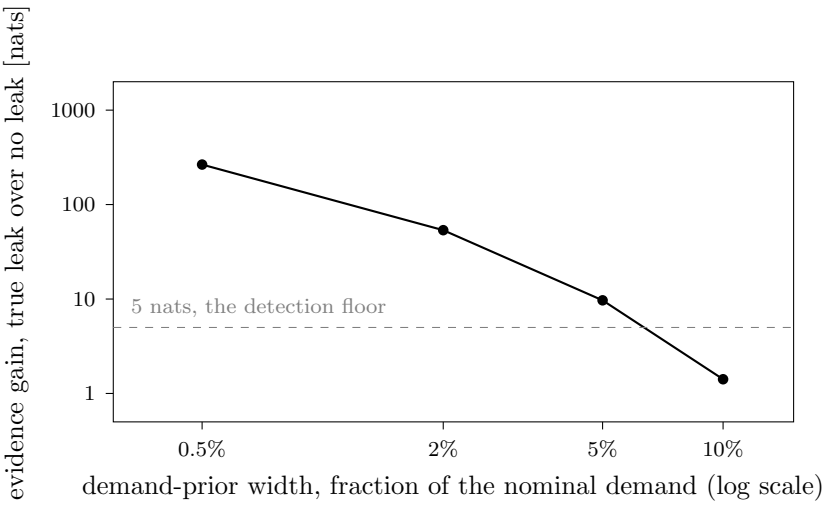


Figure 5: Identifiability against demand uncertainty for a 0.5 L/s leak. The mean evidence gain of the true hypothesis over the leak-free model across the nine leaks (seed 0, $T = 24$) as the per-snapshot demand prior widens from 0.5 % to 10 %. A leak is a demand increase the data cannot distinguish from demand uncertainty at the same node; only its constancy across snapshots identifies it.

(calibrated range 123.3–142.8), which is enough for the leak.

Next sensor, two leaks, validity. For the leak at 1 (seed 0), the pressure reading that would most reduce the entropy of the hypothesis posterior is at junction 1 (0.002 nats of 0.055), then 9 (0.001) and 3 (0.000); the posterior is already nearly certain, so little remains to gain. With leaks of 0.5 kg/s at junctions 3 and 6 at once, among 46 hypotheses the top is leak at 3 and 6 with probability 0.491 (its exact tie partner leak at 3 and 9, 0.491, is the same hypothesis through the dead end); the true pair ranks 1 with 0.491 and beats the best single-leak hypothesis (leak at 6) by 79.9 nats. The Laplace posteriors of the two leading hypotheses are both valid: effective sample ratios 1.000, 1.000 of 1 and curvature discrepancies 1.1e-03, 6.9e-04.

Modena. The nominal model (C = 130) misses the leak-free day by 13.6 cm rms (43.1 cm at worst); calibrated on 6 leak-free snapshots with a ±2 % prior it misses by 3.3 cm at the junctions and by 4.6 mm at the sensors (calibrated C 127.5–132.8, mean 130.1). Table 3 sets this study beside the published means, Figure 6 shows every scenario and Table 4 lists them. The best candidate is 0.48 ± 0.59 km (2.03 pipes) from the leak on average, exact in 13 scenarios and within one pipe in 17; the five best average 0.55 km. This study's best-candidate distance (0.48 km, mean over 32 scenarios) is better than the 0.72 km Irofti et al. report for FGLL. Its average-of-five distance (0.55 km) is better than the 0.78 km they report. Its head RMSE (0.383 m, sensors only) is better than the 3.29 m they report. The screening (22) put the true node in the shortlist in 31 of 32 scenarios (median screening rank 2.5) and the exact evidence put it top in 13; the leak-free model was excluded by 5 nats in 32. The sensors-only head estimation gives 0.383 ± 0.175 m at the 252 unmeasured junctions over 25 samples (0.036 m with the nominal pattern known); the authors' interpolation-based estimators sit at 2.1–4.41 m because they carry no hydraulic model of the unmeasured nodes' demands. Each scenario costs 214 s here: one leak-free solve, one screening, and 25 exact hypotheses.

Table 3: Modena comparison. The first four columns are the means Irofti et al. (2025) publish (their Tables IV and V; standard deviations omitted here) for the 32 scenarios; the last is this study (mean ± standard deviation over the same scenarios). Distances are shortest-path along pipes. The head RMSE row is the leak-free estimation at the 252 unmeasured junctions from the sensors alone; with the nominal demand pattern also known the study's is 0.036 m. Runtimes are the authors' machine against this one and are not comparable beyond their order of magnitude.

metric	GHR-S	GSI	UKF-AW-GSI	FGLL	this study
best candidate [km]	1.47	0.93	0.69	0.72	0.48 ± 0.59
best candidate [pipes]	7.37	4.38	3.28	3.34	2.03 ± 2.24
average of five [km]	1.44	0.93	0.80	0.78	0.55 ± 0.36
average of five [pipes]	7.12	4.37	3.81	3.60	2.39 ± 1.24
head RMSE, unmeasured nodes [m]	4.41	2.10	1.41	3.29	0.383 ± 0.175
seconds per scenario	–	–	936	60.7	214

Cost. 3238 MAP solves, 36 Newton solves and 32 screenings produce every number in this report (3274 engine solves in all).

Table 4: Modena, scenario by scenario: true leak node, whether it carries a demand meter, its rank in the score-test screening of all 268 junctions, the best candidate with its shortest-path distance in km and pipes, the mean distance of the five best, and the true node's posterior mass and rank when it made the shortlist (31 of 32 did; – otherwise).

#	leak	metered	screen rank	best	km	pipes	avg-5 km	P(true)	rank
1	4	yes	3	4	0.00	0	0.28	1.00	1
2	11	no	4	55	1.32	4	0.82	0.00	4
3	16	no	19	251	1.67	6	1.67	0.00	17
4	25	no	4	196	0.20	1	0.23	0.00	2
5	33	no	1	33	0.00	0	0.41	1.00	1
6	37	no	1	37	0.00	0	1.17	1.00	1
7	40	yes	3	40	0.00	0	0.79	1.00	1
8	49	no	2	184	0.19	1	0.54	0.00	2
9	55	no	1	55	0.00	0	0.40	1.00	1
10	60	no	1	60	0.00	0	0.33	1.00	1
11	73	no	5	70	0.29	3	0.34	0.00	4
12	79	no	24	94	1.17	4	1.27	–	–
13	84	no	12	61	0.86	5	0.68	0.00	10
14	88	no	4	58	0.96	3	0.38	0.00	4
15	91	yes	1	91	0.00	0	0.26	1.00	1
16	103	no	2	237	0.79	3	0.54	0.00	2
17	111	no	1	111	0.00	0	0.41	1.00	1
18	117	no	7	183	0.65	4	0.35	0.00	9
19	126	no	4	111	1.61	5	0.99	0.00	4
20	134	no	1	134	0.00	0	0.17	0.99	1
21	143	no	2	145	0.05	1	0.48	0.01	2
22	155	no	2	94	0.78	3	0.55	0.00	2
23	161	no	1	161	0.00	0	0.29	1.00	1
24	172	no	2	171	0.66	3	0.39	0.00	2
25	180	no	5	178	0.12	1	0.23	0.41	2
26	183	no	1	183	0.00	0	0.48	1.00	1
27	189	no	1	189	0.00	0	0.28	1.00	1
28	207	no	7	9	1.06	3	1.36	0.00	8
29	228	no	6	228	0.00	0	0.23	0.99	1
30	243	no	3	220	2.16	9	0.58	0.00	3
31	259	no	8	142	0.42	4	0.27	0.00	6
32	268	no	1	92	0.28	2	0.46	0.00	3

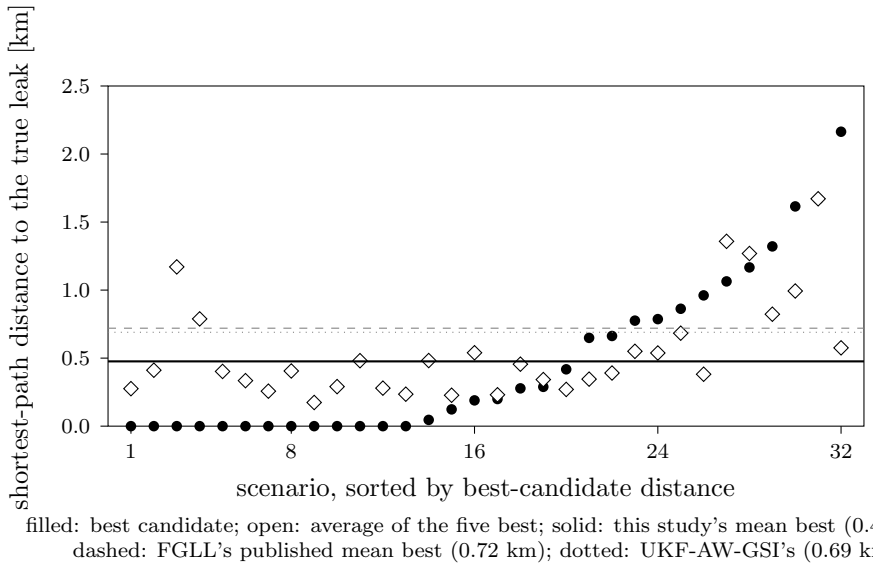


Figure 6: Modena, 32 leak scenarios: the shortest-path distance from the true leak to the best candidate and the mean over the five best, scenario by scenario, against the published means of FGLL and UKF-AW-GSI (their per-scenario values are not published). 13 of 32 best candidates are the leak node itself.

5 Discussion

Three limits bound the T-Example results. The noise is drawn here at stated levels the authors did not use (their data are noise-free), so the absolute probabilities are a property of those levels; the relative picture — which leaks are hard, how the window helps, when demand uncertainty wins — is not. The two dead ends behind unmetered pipes are exactly indistinguishable for any method with these sensors, and the report counts them as ties rather than as successes or failures. The pre-committed floor of 5 nats on excluding the leak-free model was breached once (the junction fed directly by the reservoir, 4.0 nats at its worst seed) and is stated as observed. On Modena the generating model carries 1 % roughness and diameter noise the estimator does not know; calibration on the leak-free day, differential readings (31) and a 5 % response-error term (32) in the head noise make localization work, but they do not make the posterior probabilities honest: a systematic model error is not Gaussian noise, so the top hypothesis carries probability near one whether it is right or wrong. The ranking — best candidate 0.48 km from the leak on average — is the deliverable there, and the probability is reported with that caveat. The comparison with FGLL, GSI and UKF-AW-GSI is on their published means over the same scenarios; their code was not run here, and their per-scenario values are not published, so no paired test is possible. The screening step is exact for unmetered junctions and biased for metered ones, which is why the shortlist keeps the two groups apart; a screening that also changes what the meter reads would remove the bias. Freeing every roughness inside every hypothesis failed for a numerical reason (Gauss–Newton on a C-parameterized law), not a statistical one. All leaks here are single and constant; the data's own leaks vary with pressure by about a percent on the T-Example and by a third on Modena, and a pressure-dependent leak law would fit the latter better.

6 Conclusion

This report asked six questions of one construction, a joint model of T snapshots of a water network sharing a single leak variable, on the T-Example and Modena data of Irofti et al. (2025). Which junction leaks, and how sure can one be: on the T-Example the true junction is top or in an exact structural tie with its dead-end partner in 81 of 81 cases, the leak-free model is excluded by at least 4.0 nats, and the leak is sized to 0.029 kg/s RMSE with 80 of 81 intervals covering the truth. How the answer grows with the window: the posterior mass on the true junction rises from 0.42 at one snapshot to 0.75 at 24. How wide the demand prior may be: identifiability dissolves once the per-snapshot prior widens past a few percent, the evidence for the true leak falling from 265 nats at 0.5 % to 1.4 at 10 %. What a roughness error does: a 10 % error becomes a leak bias of -5.478 kg/s when left fixed and -0.001 kg/s after a calibration on a leak-free day. Which sensor to add, whether two leaks can be told from one, and whether the posterior can be trusted: the most informative added gauge is at junction 1, the true pair of simultaneous leaks beats the best single-leak hypothesis by 79.9 nats, and the Laplace posteriors of the leading hypotheses pass the profile and importance checks. How far the best candidate is from the leak on Modena: a score test screens all 268 junctions and the exact evidence ranks a shortlist, placing the best candidate 0.48 km from the leak on average against FGLL's published 0.72 km and the five best 0.55 km against 0.78 km; the comparison is on their metric and their data, not their code, and the posterior probabilities on Modena are overconfident, so the ranking is the deliverable there. Two dead ends behind unmetered pipes remain one hypothesis to any method with these sensors. A companion study applies the same conservation and closure rows to district-heating networks, where a second conserved quantity rides on the same flows.

A Reproduction

```
cd docs/terra_graph_engine/case_studies/water_networks/leak_localization_modena
python data/fetch_modena.py           # optional; needs HYDRAULICS_DATA_DIR
PYTHONPATH=../../../../../src python run.py
PYTHONPATH=../../../../../src python run.py --add-layout   # the Modena map's coordinates
PYTHONPATH=../../../../../src python make_report.py --pdf
```

`run.py` writes `results.json` (every number at six significant digits) and `timings.json`; `make_report.py` reads only `results.json`. Seeds 0–8 draw the reading noise. The authors' FGLL code needs GTSAM, which has no wheel for this platform, so their numbers are quoted from the paper.

B Data and problems

The data are Irofti et al.'s release (github.com/pirofti/FGLL, commit 51240b8, ISC license): the nine T-Example files are committed with the study, the two Modena files are fetched by pinned commit and verified by sha1. Under `problems/`, four directories: *a dead end is one hypothesis* (the exact tie of leaks at 7 and 8); *how many snapshots does a leak need* (the leak's width against T); *screen every junction, then confirm the few* (the score test against the exact evidence); *the meter at the leaking node* (demand plus leak as an auxiliary row, and why the evidence must count every factor). Each is a statement, a script that solves it with the plugin, and its results file.

References

[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. Code and data: github.com/pirofti/FGLL (ISC license).

[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.

[Romero-Ben et al. 2026] Romero-Ben, L., Irofti, P., Stoican, F., Puig, V. (2026). Dual unscented Kalman filter architecture for sensor fusion in water networks leak localization. IEEE Transactions on Control Systems Technology 34(1), 343–354. doi:10.1109/TCST.2025.3610975.

[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.

[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). doi:10.1061/(ASCE)WR.1943-5452.0001601.

[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.