Applied Science, Computing and Energy, 2026, 5(2), 117-136

Intelligent Cybersecurity Frameworks for Detecting Data Manipulation, Sensor Spoofing, and Operational Anomalies in Critical Water Infrastructure

Ibidun Isaac*, Henry Obiajuibeh, Samuel A. Adeniji
Received: 21 July 2026/Accepted: 16 October 2026/Published: 24 October 2026

Abstract: *Water utilities depend on telemetry that can be wrong for three very different reasons: an attacker has altered or replaced it, an instrument has failed, or the network itself has changed. We developed an intelligent cybersecurity framework that detects these situations and tells them apart. Four evidence layers are computed from SCADA tags alone: hydraulic consistency (tank mass balances, pump characteristics, main headloss, system water balance, night flow and pressure surrogates), signal integrity (instrument noise fingerprints, replay correlation and exact matches against the historian), learned consistency, and departure from the recent operational routine. A gradient boosted classifier trained on a digital twin attack library attributes each alarm to data manipulation, sensor spoofing or an operational anomaly, and a ranking step localises the affected instrument. The framework was tested on 180 days of EPANET simulation of the Net3 network with 21 tags and 52 test events in 13 scenarios. Calibration from telemetry recovered tank areas within 1.5 % and pump curves within 0.1 m. The framework detected 85 % of events, including all spoofing events, with an F1 score of 0.64 against 0.54 for an autoencoder, attributed 93 % of detected events to the correct family against 82 % to 85 % from raw features, and ranked the affected instrument first in 67 % of cases. Replay remained detectable when attackers added fresh noise, through its physical consequences. Physically consistent falsification, background leaks below 2 % of demand, and attack types absent from the training library were the main failure modes.*

Keywords: *Water distribution; Cyber physical security; Sensor spoofing; False data injection; Anomaly attribution; Hydraulic modelling*

Ibidun Isaac*
Cybersecurity Ed. G. Smith College of Business, University of Missouri Saint Louis, USA
Email: Ibidun07@gmail.com

Henry Obiajuibeh
Department of Computer Science, School of Computing, Data, and Mathematical Sciences, Western Illinois University State: Illinois, USA.
Email:: henryibeh89@gmail.com

Samuel A. Adeniji
School of Computer Science, College of Business and Technology, Western Illinois University, Macomb, Illinois, United States.
Email: **Samueladeniji247@gmail.com**

1. 0 Introduction

An operator looking at a SCADA screen in a water utility sees numbers, not water. Every pressure, level and flow on the display has passed through a transmitter, a remote terminal unit, a radio or fibre link, a programmable logic controller and a historian before it is drawn, and at each of those points the number can be wrong. Most of the time it is wrong for mundane reasons: a transmitter drifts, a level float sticks, a radio link drops and the last value is held. Sometimes the process itself has changed in a way the operator did not expect, because a main has burst or a fire crew has opened a hydrant. And occasionally the number is wrong because someone has made it so. Joint advisories from United States federal agencies have described intrusions into water and wastewater control systems in which attackers gained the ability to alter setpoints and displayed values (CISA, 2021), and reviews of documented incidents show that the sector has been targeted repeatedly over two decades (Hassanzadeh et al., 2020; Tuptuk et al., 2021).

From the operator's chair these situations look alike: a reading that does not match expectation. They demand very different responses. A drifting transmitter calls for a technician with a calibrator. A burst main calls for valve crews and customer notices. A falsified reading calls for isolating a compromised network segment, verifying the true state of the plant by hand and alerting the security team, because an attacker who can rewrite one value may be hiding a physical action elsewhere. An alarm that says only that something is unusual leaves the operator to guess which response is needed, and a wrong guess wastes the most valuable resource in an incident, which is time.

Research on automated detection in water networks has advanced considerably since the BATtle of the Attack Detection ALgorithms (BATADAL) competition gave the community a shared benchmark on the C Town network (Taormina et al., 2017, 2018). Model based detectors that check SCADA data against hydraulic understanding (Housh and Ohar, 2018), deep autoencoders trained on normal operation (Taormina and Galelli, 2018), convolutional and recurrent detectors evaluated on the Secure Water Treatment and Water Distribution testbeds (Mathur and Tippenhauer, 2016; Ahmed et al., 2017; Inoue et al., 2017; Kravchik and Shabtai, 2018) and graph neural networks for multivariate sensor streams (Deng and Hooi, 2021) all detect a substantial share of scripted attacks. In parallel, the hydroinformatics community has built a mature literature on detecting bursts and abnormal flows from district meters and pressure loggers (Mounce et al., 2010; Romano et al., 2014; Puust et al., 2010). The two literatures rarely meet. Attack detectors are usually evaluated on records in which every anomaly is an attack, so they are never asked to recognise a burst or a stuck transmitter as something else, and burst detectors assume that the data they receive are honest. To date, a recent systematic review of AI-based detection algorithms for water distribution networks between 2014 and 2025 categorises the literature based on the detection task, data modality, learning paradigm and model architecture (Habib et al., 2026).

The control and security literature offers two further strands that bear directly on this gap. The first concerns integrity attacks that are designed to be consistent with the physics a defender expects, as in false data injection against power system state estimation (Liu et al., 2011) and stealthy attacks on industrial controllers (Cárdenas et al., 2011; Teixeira et al., 2015; Urbina et al., 2016). These works establish that physical models raise the bar for an attacker without making detection certain. The second concerns the signal itself rather than its physical meaning. Replay attacks, in which recorded data are played back to hide a manipulation, motivated physical watermarking schemes in control (Mo and Sinopoli, 2009), and the observation that every sensor and process has a characteristic noise signature has been used to authenticate measurements in a water treatment testbed (Ahmed et al., 2018). Surveys of physics based detection in cyber physical systems argue that such complementary evidence is needed but rarely combined (Giraldo et al., 2018; Addeen et al., 2021).

Taken together, these strands suggest a different design principle. Instead of a single detector that scores how unusual the data are, a utility needs several independent kinds of evidence, each sensitive to a different way in which a reading can be wrong, and a decision stage that uses the pattern of evidence to say what has happened and where (Idowu et al., 2026). Arithmetic manipulation of a value in transit should violate physical balances that link it to other instruments. Substitution of a recorded or synthetic stream should betray itself through its noise, which belongs to another moment or to no instrument at all. A genuine operational upset should leave the physics and the signals intact but move the system away from its usual daily pattern. If those expectations hold, the evidence pattern carries the attribution.

This study develops and tests such a framework on a hydraulically explicit digital twin of a distribution network with two pumping stations, three storage tanks and 21 telemetry tags. The work had four objectives: to formalise a threat taxonomy that places data manipulation, sensor spoofing and operational anomalies, including instrument faults, in one evaluation; to build evidence layers for hydraulic consistency, signal integrity, learned consistency and operational state whose parameters are calibrated from telemetry alone; to combine them in a detector, an attribution classifier and a localisation step; and to probe the limits of the approach through a calibrated bias sweep and an adaptive replay attack. Because utilities cannot supply labelled attacks on their own networks, the evaluation follows the simulation practice of BATADAL and uses EPANET through the Water Network Tool for Resilience (Rossman, 2000; Klise et al., 2017). The code is written so that the same pipeline runs unchanged on historian exports.

2. 0 Materials and methods

2.1. Benchmark network and digital twin

The study used the EPANET Net3 network, a widely used benchmark of 92 junctions and 117 pipes supplied from a lake through pumping station 10 and from a river through pumping station 335, with three elevated storage tanks (Fig. 1). The original time of day controls on pump 10 were replaced by an equivalent speed pattern that runs the pump from 01:00 to 15:00 each day, which allowed actuator attacks and unplanned trips to be imposed by editing the pattern. Pump 335 kept its original control logic, which starts and stops it on the level of tank 1. Hydraulics were solved by EPANET 2.2 through WNTR 1.5 (Klise et al., 2017) at a 15 min hydraulic and reporting step for 180 days, giving 17 280 time steps.

Demand varied stochastically at three scales. The five demand patterns of Net3 were interpolated to 15 min and multiplied by a daily factor that followed a first order autoregressive process (standard deviation of innovations 3.5 %), a weekend reduction of 5 %, a slow seasonal cycle of ±6 % with a period of 120 days and a 15 min autoregressive jitter of 3 %. Mean system consumption was close to 700 L s^{-1}.

Twenty-one SCADA tags were archived (Fig. 1), covering the levels of the three tanks and the flows in their connecting pipes; flow, run status, and discharge pressure at pump 10; flow, status, suction and discharge pressure at pump 335; flow in the river main; and pressure at seven junctions selected for their centrality in the network graph.

Each continuous instrument was assigned a unique noise fingerprint consisting of first-order autoregressive noise, with a coefficient drawn between 0.25 and 0.80, an amplitude drawn around a class value (0.004 m for levels, 0.25 L s^{-1} for flows, and 0.12 m for pressures), and a quantisation step typical of its transmitter (0.001 m, 0.01 L s^{-1}, and 0.01 m). These fingerprints remained fixed throughout the life of the network, as would be expected for real instruments.

2.2. Threat taxonomy and scenarios

Three families of event were defined (Table 1). Data manipulation (DM) covers arithmetic alteration of a value between the instrument and the historian: a step bias on a junction pressure (DM1), a slow ramp on a tank level (DM2), scaling of the pump 10 flow meter (DM3) and a coordinated falsification of the level and inflow of tank 3 that was constructed to satisfy the tank mass balance exactly (DM4). Sensor spoofing (SS) covers substitution of a whole stream by data that do not come from the instrument at that moment, used to conceal a physical action: pump 10 forced on outside its schedule while the tank and pump telemetry are replayed from the previous day (SS1), pump 10 forced off during its schedule while the pump telemetry is replayed (SS2), and an isolation valve closed while the local pressure is replaced by a synthetic profile built from the seven day mean with white noise of the right amplitude (SS3). Operational anomalies (OA) are events without an adversary: pipe bursts (O1),

growing background leaks (O2), unplanned trips of pump 10 (O3), fire flow (O4), a stuck pressure transmitter (O5) and a drifting pressure transmitter (O6).

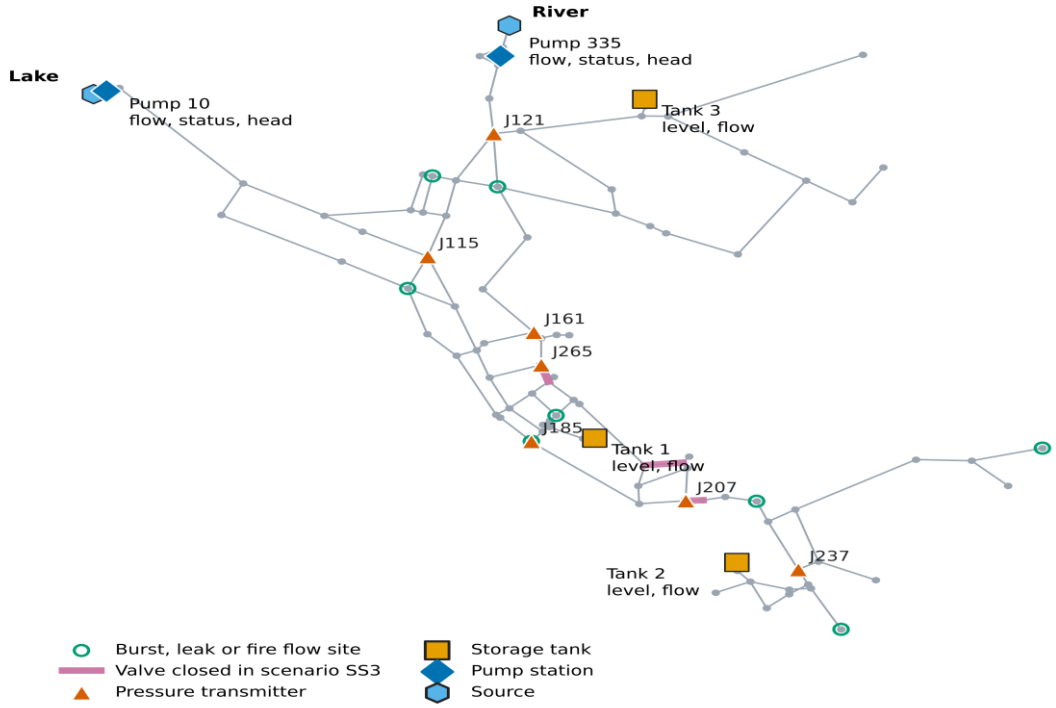


Fig. 1: The Net3 benchmark network with the telemetry used in this study: levels and connecting flows of the three tanks, flow, run status and head at both pumping stations, river main flow and seven junction pressures. Circles mark the junctions where bursts, leaks and fire flow were imposed in the test record, and thick segments the pipes closed in scenario SS3. Days 1 to 56 were used for training and days 57 to 70 for validation; neither contained events. Days 71 to 180 formed the test period

Four instances of each of the 13 scenarios were placed at random, non overlapping times in the test period, giving 16 manipulation events, 12 spoofing events and 24 operational events. Leak, burst and fire flow sites were drawn from 20 junctions spread across the network, and the valves closed in SS3 were selected by a screening simulation as pipes whose closure lowered pressure at a monitored junction by more than 5 m without isolating demand. Events occupied 34.0 % of the test record, of which 23.3 % were operational anomalies, 7.2 % manipulation and 3.5 % spoofing.

A second, independent 180 day record with a different random seed, wider magnitude ranges and seven instances of each scenario (91 events) formed the twin library. It was used only to train the attribution classifier.

Table 1. Scenarios in the evaluation. Durations and magnitudes are the ranges drawn for the test record; replayed data repeat the values archived 24 h earlier

ID	Family	Scenario	Physical action	Telemetry presented to SCADA	Duration (h)	Magnitude
DM1	Manipulation	Pressure bias	None	Junction pressure offset	3 to 12	−5 to −12 m
DM2	Manipulation	Level ramp	None	Tank 1 or 2 level ramped	8 to 24	+0.6 to +1.4 m

ID	Family	Scenario	Physical action	Telemetry presented to SCADA	Duration (h)	Magnitude
DM 3	Manipulation	Flow meter scaling	None	Pump 10 flow multiplied	6 to 14	0.6 to 0.8 times
DM 4	Manipulation	Coordinated falsification	None	Tank 3 level ramp with matching inflow	8 to 24	+0.8 to +1.6 m
SS1	Spoofing	Masked overfilling	Pump 10 forced on outside schedule	Levels, tank flows and pump 10 tags replayed	4 to 10	Full pump output
SS2	Spoofing	Masked supply cut	Pump 10 forced off within schedule	Pump 10 flow, status and pressure replayed	4 to 10	Full pump output
SS3	Spoofing	Masked valve closure	Isolation valve closed	Local pressure replaced by synthetic profile	4 to 12	1 to 5 m local drop
O1	Operational	Pipe burst	Nodal outflow	Honest	3 to 8	30 to 60 L s^{-1}
O2	Operational	Background leak	Growing nodal outflow	Honest	24 to 72	5 to 12 L s^{-1}
O3	Operational	Pump trip	Pump 10 stopped	Honest	2 to 6	Full pump output
O4	Operational	Fire flow	Nodal outflow	Honest	2 to 4	30 to 50 L s^{-1}
O5	Operational	Stuck transmitter	None	Pressure held at last value	6 to 24	Not applicable
O6	Operational	Drifting transmitter	None	Pressure drifts linearly	48 to 96	−3 to −6 m

2.3. Evidence layers

Four evidence layers process the 21 tags (Fig. 2). Every parameter was estimated from the training period by least squares or by robust statistics, without reference to the network model, so that the same procedure can be applied to a utility's historian.

Hydraulic consistency. Six families of balance were written for quantities that the telemetry measures directly. For tank k with connecting flow Q_k, the mass balance residual over one hour is $r_{T,k}(t) = \hat{A}_k\,[L_k(t) - L_k(t-4)] - \Delta t \sum_{i=t-4}^{t-1} Q_k\,(i),$ (1)

where $\hat{A}_k$ is the plan area estimated by regression and Δt is 900 s. For each pump the head gain across the station was compared with a quadratic characteristic fitted to periods in which the pump ran, $r_P = (H_d - H_s) - (a_0 + a_1 Q + a_2 Q^2),$ (2)

and when the pump was reported stopped the residual was the reported flow scaled by 10 L s^{-1}, so that flow through a stopped pump appears as an inconsistency. Flow in the river main was compared with a power law of the head difference between the river and the suction node, $r_R = Q_R - c\,(H_R - H_s)^b$, the form taken by the Hazen and Williams relation. System consumption was inferred from the water balance,

$$D(t) = Q_{P10}(t) + Q_R(t) - \sum_k Q_k\,(t), \quad (3)$$

and compared with its expected value, the median daily profile scaled by the ratio of consumption to profile over the preceding day. The minimum night flow between 02:00 and 04:00, expressed as a fraction of the mean daily consumption, was compared with its training median, following standard water loss practice (Farley and Trow, 2003). Finally, a pressure surrogate for each monitored junction, a ridge regression on tank levels, pump states and flows, the inferred consumption and its square and four harmonics of the time of day, predicted the pressure that the boundary conditions imply. Its residual $r_j = p_j - \hat{p}_j$ is the hydraulic analogue of state estimation. All residuals were centred on their training medians and divided by 1.4826 times their median absolute deviations.

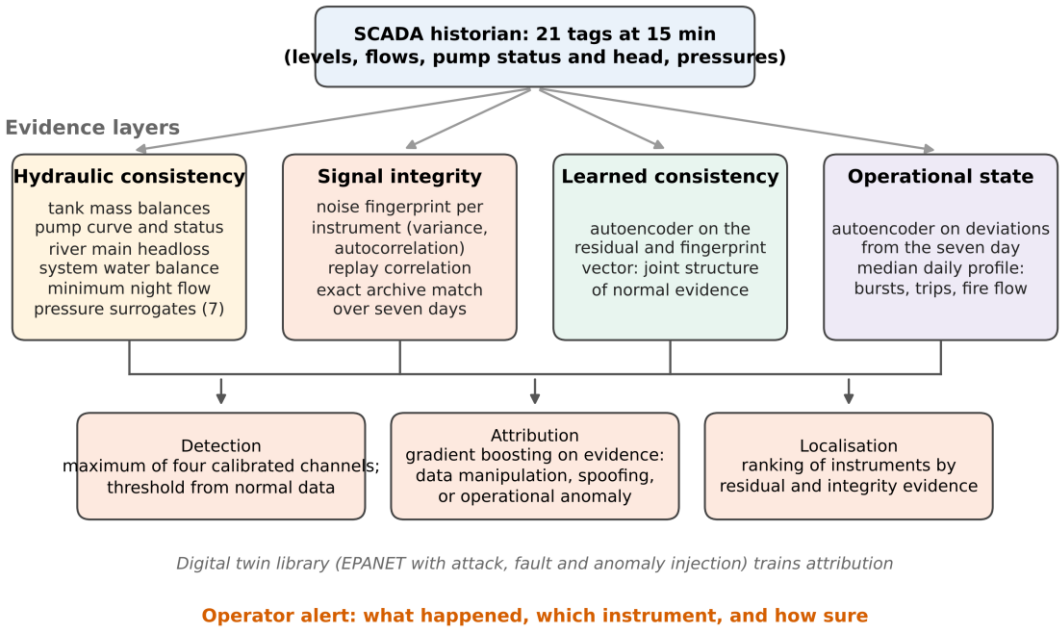


Fig. 2. Architecture of the framework. Four evidence layers computed from historian tags feed detection (maximum of four calibrated channels), attribution to a threat family by a classifier trained on a digital twin library, and localisation of the affected instrument

Signal integrity. For each continuous tag the second difference $e_t = y_t - 2y_{t-1} + y_{t-2}$ removes slowly varying hydraulic behaviour and leaves mostly instrument noise. Over a 6 h window the noise fingerprint combines a robust variance and the lag one autocorrelation of e_t, each standardised against training values,

$$F_s(t) = \sqrt{z_{\mathrm{var}}^2 + z_{\mathrm{ac}}^2}. \quad (4)$$

A stuck transmitter drives the variance to zero, and a synthetic stream with the right amplitude but the wrong colour shifts the autocorrelation. Replay is tested in two ways. The replay correlation is the largest Pearson correlation between the winsorised second differences in the current 6 h window and in every earlier window between 6 h and seven days before,

$$\rho_s(t) = \max_{d \in [24,\,672]} \mathrm{corr}(e_{t-23:t},\ e_{t-d-23:t-d}), \quad (5)$$

and the exact archive match is the length of the current run of samples whose quantised values equal those recorded at a common lag d in the same range. For honest data quantised at 0.01 m with noise of about 0.12 m, the chance that one sample equals the archive value at a given lag is roughly 2 %, so runs longer than a few samples are very unlikely unless the stream has been copied. The run length was divided by its 99.9th training percentile.

Learned consistency. An autoencoder with 32, 8 and 32 hidden units (Hinton and Salakhutdinov, 2006) was trained on the vector of hydraulic residuals, their absolute values and the noise fingerprints. Its reconstruction error responds to combinations of evidence that are individually modest but jointly unusual.

Operational state. Each tag was expressed as its deviation from the median of the same time of day over the previous seven days, which removes daily patterns and slow seasonal drift, and a second autoencoder with 32, 8 and 32 units was trained on these deviations and their changes over one hour. This channel responds to events that are physically consistent and leave honest signals but move the network away from its recent routine, such as bursts, trips and fire flow.

2.4. Detection, attribution and localisation

Each of the four channel scores was normalised by its median and interquartile range on the validation period, smoothed over 1 h and divided by its own 99.75th validation percentile. The detection score is the maximum of the four calibrated channels, and an alarm is raised when the smoothed score exceeds its 99.5th validation percentile.

Attribution assigns each anomaly to one of the three families. A histogram based gradient boosted classifier (Friedman, 2001) with 300 iterations, a learning rate of 0.06 and 31 leaves per tree was trained on all event samples of the twin library. Its inputs summarise the evidence: the absolute hydraulic residuals; the largest and second largest residual, the number above 4 and the variability and trend of the largest over 4 h; the largest fingerprint, replay correlation and archive match, overall and for pressures, levels and flows separately, with the number of instruments exceeding fixed levels; and the largest operational deviation for each class of tag. Class probabilities were smoothed over 2 h. For each detected event the family with the highest mean probability over its alarmed samples was taken as the attribution.

Localisation ranks instruments. The evidence for a tag is the largest of its fingerprint, replay correlation and archive match and of the scaled sum of the hydraulic residuals in which it appears; for example, the pump 10 flow meter enters the pump characteristic and the water balance, whereas the discharge pressure enters only the former. For each detected event the instrument with the highest mean evidence over the alarmed samples was compared with the instruments actually attacked or faulty (Akanni et al., 2026).

2.5. Benchmarks and evaluation protocol

Four detectors served as references: the squared prediction error of principal component analysis retaining 95 % of variance (Qin, 2003), an isolation forest with 300 trees (Liu et al., 2008), an autoencoder with 48, 12 and 48 hidden units trained on the standardised tags and their changes over 1 and 2 h, and a detector that alarms on the largest hydraulic residual alone. All thresholds were set by the same rule. For attribution, an identical classifier was trained on raw tag features and applied both to autoencoder alarms and to the framework's alarms. At sample level we report the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC; Saito and Rehmsmeier, 2015), precision, recall, F1 and the false positive rate. At event level we report the fraction of events with at least one alarm, overall and for each family, the median time to first alarm and false alarm episodes per day. Positives are all three families, since a utility wants to know about bursts and failed transmitters as much as about attacks. Storage tanks take several hours to recover from a major event, and alarms during recovery are neither clearly false nor clearly correct; the 12 h following each event were therefore excluded from scoring for every method. Attribution was scored as accuracy and macro F1 over detected events, and localisation as the fraction of instrument level events whose first-ranked, or one of whose three highest-

ranked, instruments was among those attacked or faulty.

Three further analyses tested generalisation and robustness. In a leave one scenario out test the attribution classifier was retrained thirteen times without one scenario and applied to test events of that scenario. In a bias sweep, steps of 0.5 to 6 m were injected for 6 h into 30 randomly chosen quiet windows of a randomly chosen junction pressure. In an adaptive replay test, the replayed segments of SS1 and SS2 were regenerated with fresh noise of the instrument's own colour added at 0.25 to 2 times its amplitude, defeating exact matching by design. Models were implemented in Python with scikit learn 1.9 (Pedregosa et al., 2011), all seeds were fixed, and every number below is regenerated by the published scripts.

3. 0 Results

3.1. Calibration recovers the physics of the network

The balances identified from 56 days of telemetry agreed closely with the network as built. The plan areas estimated from Eq. (1) were 534, 184 and 1990 m^2 for tanks 1 to 3, within 1.3 %, 0.8 % and 1.4 % of the areas implied by their diameters (527, 182 and 1963 m^2). The quadratic characteristics fitted to running periods reproduced the manufacturer curves in the operating range: at 210 L s^{-1} the fitted head gain of pump 10 was 22.7 m against 22.7 m from the curve, and at 830 L s^{-1} that of pump 335 was 28.5 m against 28.5 m. The exponent of the river main relation was 0.49, close to the value of 0.54 implied by the Hazen and Williams formula. These comparisons matter because they show that the residuals are physical statements an engineer can audit, not arbitrary regressions, and that ordinary operation excites the network enough for identification.

3.2. Evidence signatures of the three families

Each family left a recognisable pattern across the evidence layers (Fig. 3). Under normal operation every score sat near its baseline, between 0.5 and 2.1. Manipulation attacks broke the balance that involves the altered instrument and nothing else: the pressure bias (DM1) produced a median surrogate residual of 32, the level ramp (DM2) a tank balance residual of 8.5 and the flow meter scaling (DM3) a pump characteristic residual of 76, while their noise fingerprints and archive matches stayed at baseline because the instrument's own noise was still present. The coordinated falsification at tank 3 (DM4) did what it was designed to do: no score exceeded 2.7, since the false level and the false inflow satisfy the mass balance and tank 3 has little influence on monitored pressures.

Spoofing produced the opposite signature. The replay scenarios SS1 and SS2 gave median archive matches of 6.5 and 4.3 times their 99.9th training percentile, because the historian already contained the copied values, together with large surrogate residuals, since the replayed tank levels and pump states no longer explained the pressures that honest transmitters reported. In SS2 the water balance residual rose to 6.0 because the replayed pump flow implied a consumption about 200 L s^{-1} higher than the tanks could account for. The synthetic pressure of SS3 had the right amplitude and was not copied, so its archive match stayed at baseline; it was exposed instead by a surrogate residual of 47, because the closed valve depressed pressure at the neighbouring honest transmitter while the spoofed one reported a normal day, and weakly by its fingerprint (3.4), since white noise lacks the colour of the real instrument.

Operational anomalies left the physics largely intact. Bursts, trips and fire flow produced surrogate residuals of at most 3.6 and no integrity evidence; they are physically consistent events, and they were seen mainly through the operational state channel. The stuck transmitter (O5) was the clearest integrity event in the record, with a fingerprint score of 57, and the drifting transmitter (O6) appeared through its surrogate residual (12) with a normal fingerprint. Background leaks of 5 to 12 L s^{-1},

between 0.7 % and 1.7 % of consumption, left no evidence above baseline in any layer.

3.3. A pump shutdown hidden by replayed telemetry

Figure 4 follows one SS2 event. At onset the attacker stopped pump 10 and began to replay the previous day's flow, run status and discharge pressure, so that the SCADA display continued to show about 207 L s^{-1} for 9.75 h while the true flow was zero (Fig. 4a). The tank levels were not falsified, and they fell by 0.41 m in tank 3 and 0.55 m in tank 1 over the event as storage made up the lost supply (Fig. 4b). The consumption implied by the reported telemetry exceeded the actual consumption by a median of 207 L s^{-1} (Fig. 4c), a number that would immediately alarm any engineer who looked at the water balance but that is not displayed on most SCADA screens

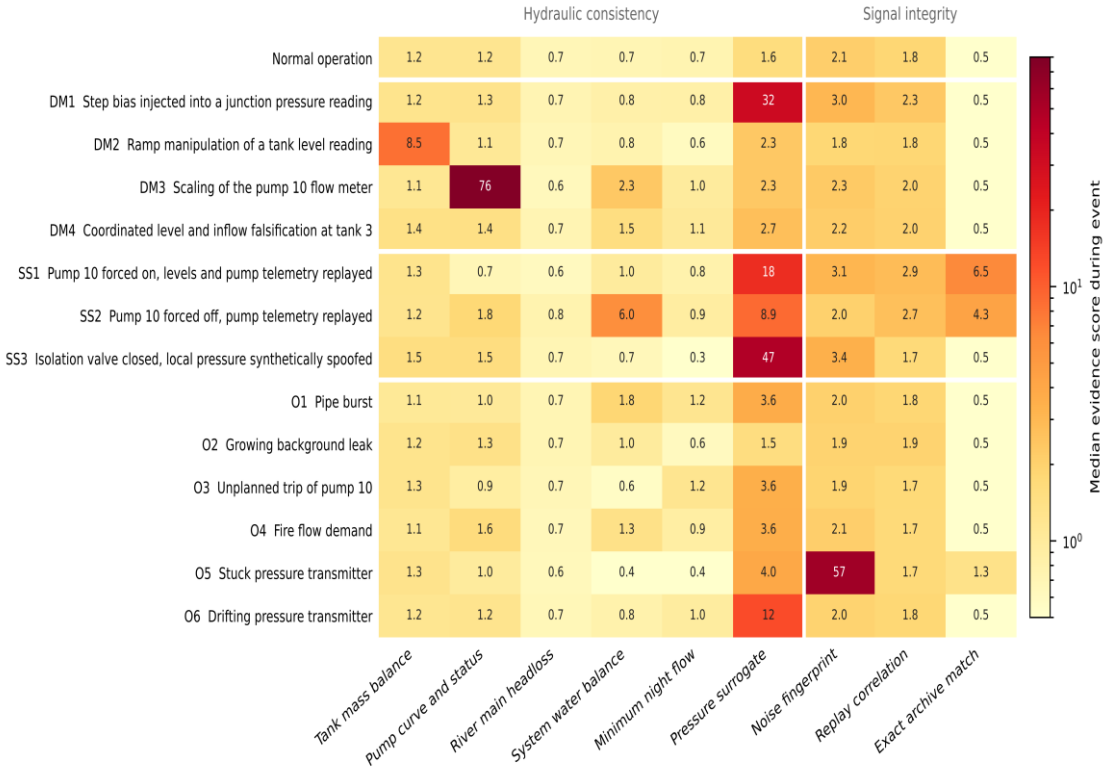


Fig. 3: Evidence signatures. Median score of each evidence family during normal operation and during each scenario in the test period (logarithmic colour scale). Pressure surrogate, tank balance and pump values are the largest over the instruments of that type; the exact archive match is the run length relative to its 99.9th training percentile.

The operational state channel crossed its threshold within the first 15 min because the tanks began to draw down at an hour when they normally fill, and the signal integrity channel followed 45 min later, when the run of samples identical to the previous day exceeded its training maximum (Fig. 4d). The raw tag autoencoder also alarmed early, for the same reason as the operational channel. The hydraulic consistency channel never crossed its threshold, because its water balance residual uses an expected consumption that adapts over a day and the pressure surrogates were only moderately disturbed. What distinguished the framework in this event was therefore not that it raised an alarm but what the alarm said: the attribution classifier labelled the event as spoofing, and the localisation step ranked the pump 10 telemetry first.

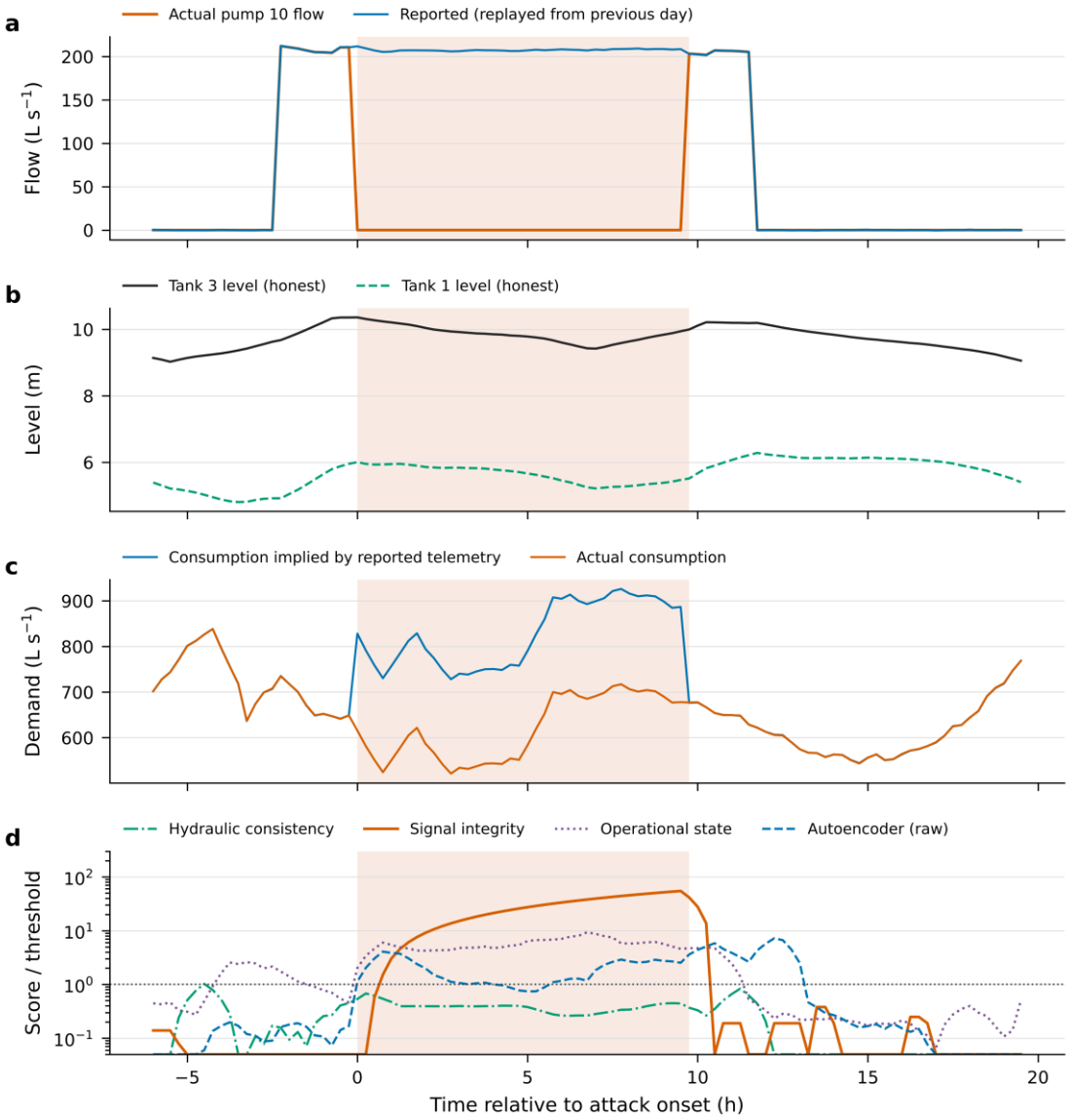


Fig. 4. A supply cut hidden by replayed telemetry (scenario SS2, a test event lasting 9.75 h). (a) Actual and reported flow of pump 10. (b) Honest levels of tanks 1 and 3. (c) Consumption implied by the reported telemetry through Eq. (3) and actual consumption. (d) Smoothed channel scores divided by their thresholds, with the raw tag autoencoder for comparison. Shading marks the attack.

3.4. Detection performance

Over the test period the framework detected 44 of the 52 events (84.6 %), against 40 for the raw tag autoencoder and 22 for the hydraulic residual detector (Table 2). Its F1 score was 0.64 against 0.54 for the autoencoder and its AUPRC 0.795 against 0.775. The hydraulic residual detector had the highest AUROC (0.788) and the highest precision (0.99), but its recall was only 0.19, because it is blind to physically consistent events. The threshold free curves of the three strongest methods were close (Fig. 5): at a false positive rate of 0.1 the framework and the autoencoder both reached a true positive rate near 0.55, and the advantage of the framework in AUPRC came mainly from higher precision at recall between 0.4 and 0.55.

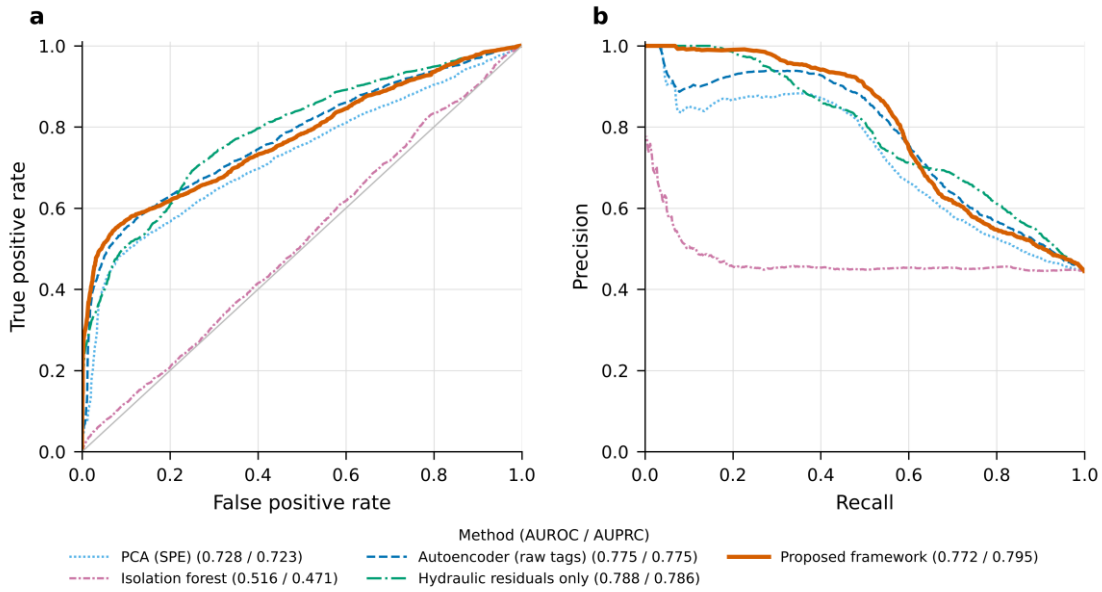


Fig. 5. (a) Receiver operating characteristic and (b) precision recall curves over the test period, with the 12 h after each event excluded. The legend gives AUROC and AUPRC.

The gain in detection came at a cost. The framework raised 0.21 false alarm episodes per day, about one every five days, against 0.07 for the autoencoder and 0.06 for the hydraulic detector, because taking the maximum of four channels adds the false alarms of each. Its false positive rate on normal samples was 3.6 %. The isolation forest performed close to chance (AUROC 0.52), and principal component analysis detected only 14 of the 52 events, a reminder that generic outlier detectors are poorly matched to networks whose normal state changes strongly with the time of day.

Table 2: Detection performance over the 110 day test period (52 events). Event recall is given overall and for manipulation (DM), spoofing (SS) and operational (OA) events. FA, false alarm episodes

Method	AUROC	AUPRC	Prec.	Recall	F1	Event recall	DM	SS	OA	FA per day
PCA (squared prediction error)	0.728	0.723	0.90	0.07	0.13	0.27	0.38	0.58	0.04	0.05
Isolation forest	0.516	0.471	0.70	0.03	0.05	0.35	0.25	0.50	0.33	0.15
Autoencoder, raw tags	0.775	0.775	0.93	0.38	0.54	0.77	0.62	1.00	0.75	0.07
Hydraulic residuals only	0.788	0.786	0.99	0.19	0.31	0.42	0.62	0.67	0.17	0.06
Proposed framework	**0.772**	**0.795**	**0.91**	**0.49**	**0.64**	**0.85**	**0.75**	**1.00**	**0.83**	**0.21**

3.5. Scenario level behaviour

The family level detection rates reveal how the layers complement each other (Fig. 6a). The framework detected all twelve spoofing events, 75 % of manipulation events and 83 % of operational anomalies, against 100 %, 63 % and 75 % for the autoencoder and 67 %, 63 % and 17 % for the hydraulic detector.

Two scenarios illustrate the division of labour. Flow meter scaling (DM3) was caught in all four events by the framework and by the hydraulic detector but in only one by the autoencoder, because a meter that under reads by 20 % to 40 % produces a smaller statistical excursion than the pump curve violation it implies. Conversely, pump trips (O3) and fire flow (O4) were caught by the framework and the autoencoder but almost never by the hydraulic detector, since they are hydraulically consistent.

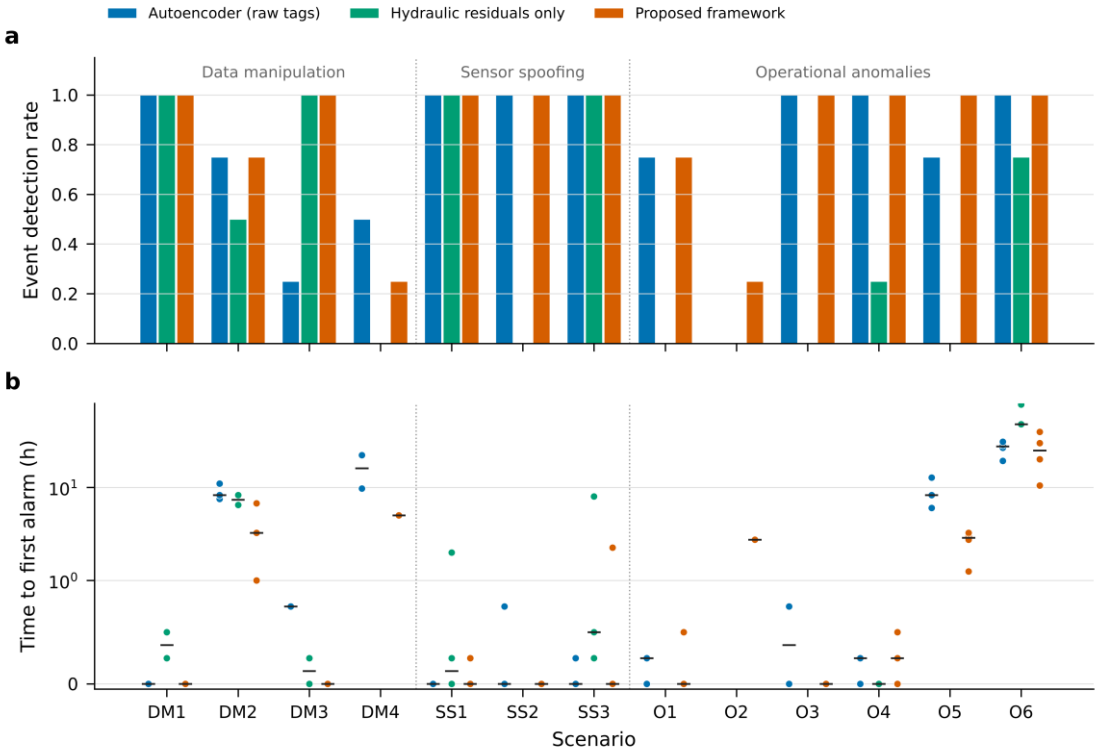


Fig. 6. Scenario level detection. (a) Fraction of the four test events of each scenario that raised at least one alarm. (b) Time from onset to first alarm for each detected event; horizontal bars mark medians, and the axis is linear below 1 h and logarithmic above.

Timeliness followed the same logic (Fig. 6b). The framework flagged the stuck transmitter after a median of 2.9 h against 8.3 h for the autoencoder, since a frozen signal loses its noise long before its value becomes implausible, and the level ramp after 3.3 h against 8.3 h. Drifting transmitters took a day or more to detect with any method. The coordinated falsification DM4 was detected in one of four events by the framework, after 5 h, and in two of four by the autoencoder, through the unusual combination of tank 3 level and network pressures rather than any broken balance. Only one of the four background leaks was flagged, after 2.8 h, and it was the smallest (7.2 L s^{-1}), which suggests that the alarm coincided with an unrelated fluctuation in demand rather than reflecting genuine sensitivity. At about 1 % to 2 % of consumption, in a network without district metering, system level evidence is at or beyond its limit.

3.6. Attribution and localisation

Attribution was the clearest benefit of the evidence layers (Fig. 7, Table 3). Of the 44 events the framework detected, 41 (93.2 %) were assigned to the correct family, with a macro F1 of 0.93. All twelve spoofing events and all twenty operational events were attributed correctly; the three errors were manipulation events labelled as operational. The same classifier trained on raw tag features achieved 85.0 % on the events detected by the autoencoder and 81.8 % on those detected by the framework, and its F1 for the manipulation family fell to 0.57 to

0.59, because a biased transmitter and a genuine change in pressure look alike in raw data but not in the residuals.

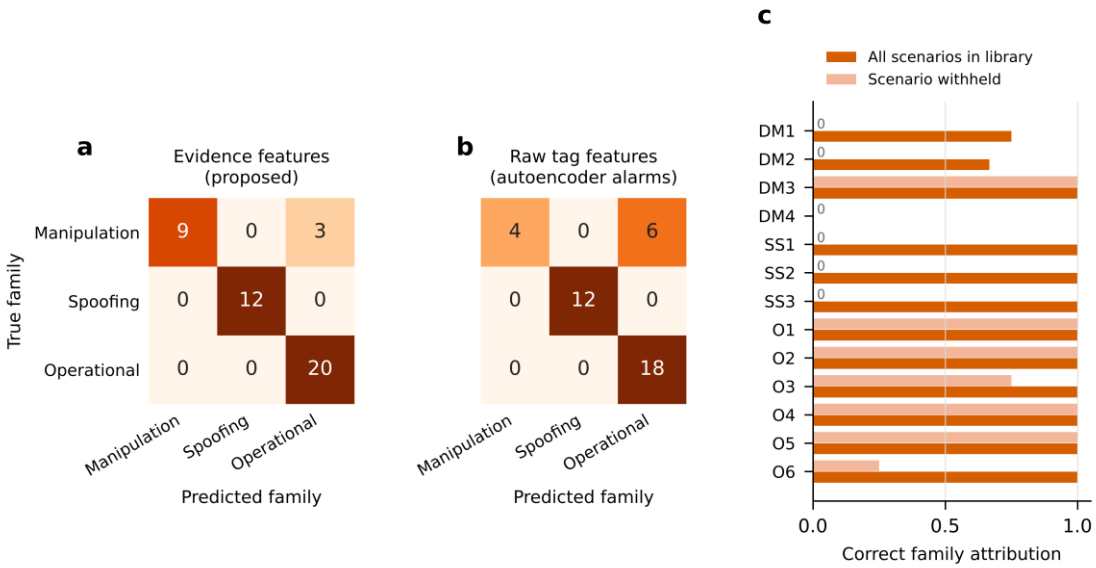


Fig. 7. Attribution of detected events. Event level confusion matrices for the classifier using (a) evidence features on the framework's alarms and (b) raw tag features on the autoencoder's alarms. (c) Correct attribution by scenario when all scenarios are in the twin library and when the scenario in question is withheld.

The leave one scenario out test exposed an asymmetry (Fig. 7c). When an operational scenario was withheld from the library, its events were still attributed correctly in most cases (100 % for bursts, leaks, fire flow and stuck transmitters, 75 % for trips and 25 % for drifts), because the classifier had learned what benign events have in common. When an attack scenario was withheld, its events were rarely recognised: only flow meter scaling (DM3) was still labelled as manipulation, and events of the other withheld attack types were labelled mostly as operational anomalies, with a minority assigned to the wrong attack family. In other words, the classifier generalises the concept of a benign event but not the concept of an attack, and an attack of a new type is more likely to be dismissed as a fault than flagged as an attack.

Localisation placed the attacked or faulty instrument first in 66.7 % of the 36 instrument level events and among the first three in 77.8 %, against 63.9 % and 69.4 % for ranking by the autoencoder's reconstruction error. It was exact for pressure bias, stuck and drifting transmitters and both replay scenarios. It failed for the synthetic pressure of SS3, where honest transmitters that recorded the real effect of the valve closure, including the immediate neighbour, were ranked above the spoofed one in three of four events, and for DM3, where the run status, the discharge pressure and the flow meter all enter the pump characteristic and were ranked first, second and third.

3.7. Sensitivity to bias magnitude and to adaptive replay

The bias sweep located the limits of each detector (Fig. 8a). The framework detected half of the 6 h pressure biases at about 1.3 m and all of them from 2 m. The raw tag autoencoder was more sensitive to small steps, detecting 80 % at 1 m, because an abrupt step in one pressure is a strong statistical outlier even when it is physically small. The hydraulic detector needed about 3.5 m for half detection, reflecting the spread of the surrogate residuals that the time of day and the distribution of demand leave unexplained. A bias of 1 to 2 m, small

compared with the 40 to 50 m of typical service pressure, therefore lies in the range where the framework is less sensitive than a well tuned data driven detector, and its value lies in explaining large manipulations rather than finding the smallest.

Table 3: Event level attribution of detected events to threat families.

Classifier input and alarm source	Events	Accuracy	Macro F1	F1, DM	F1, SS	F1, OA
Evidence framework (proposed)	**features, alarms**	**44**	**0.93**	**0.93**	**0.86**	**1.00**	**0.93**
Raw tag autoencoder alarms	features,	40	0.85	0.81	0.57	1.00	0.86
Raw tag features, framework alarms		44	0.82	0.79	0.59	0.96	0.83

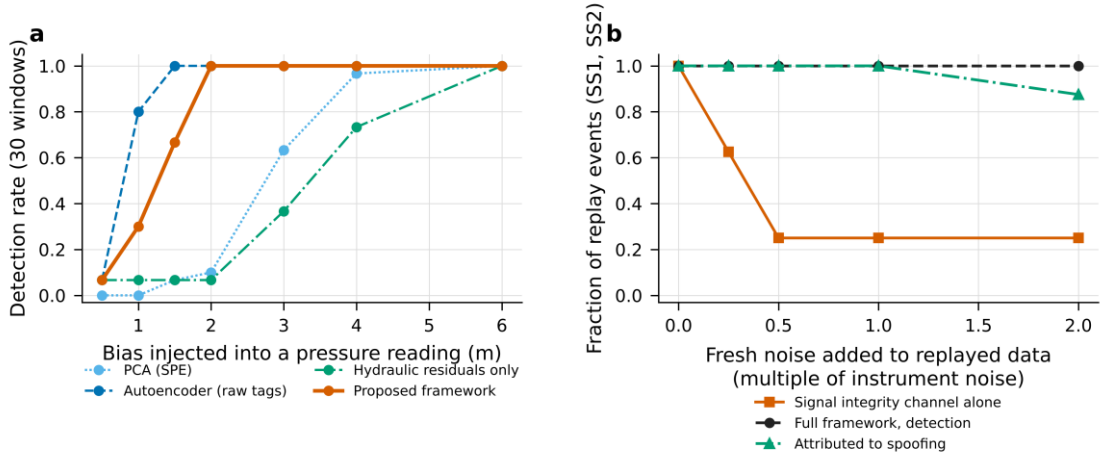


Fig. 8. Robustness. (a) Detection rate of 6 h step biases injected into a junction pressure in 30 quiet windows, against bias magnitude. (b) Detection and attribution of the eight replay events (SS1, SS2) when the attacker adds fresh noise of the instrument's own colour to the replayed data.

The adaptive replay test showed how much each layer contributes when the attacker anticipates exact matching (Fig. 8b). With verbatim replay the integrity channel alone detected all eight SS1 and SS2 events. Adding fresh noise at a quarter of the instrument's amplitude removed every exact match and cut integrity detection to 63 %, and from half the amplitude onward it detected only 25 %, the events in which the replay correlation remained high. The full framework still detected all eight events at every noise level, through the physical consequences of the forced pump state, and attributed all of them to spoofing up to a noise amplitude equal to the instrument's own, and seven of eight at twice that amplitude. The replay correlation of Eq. (5) fell only from a median score of 4.9 to 3.8 across the range, so that partial evidence of copying survives even when exact matching fails.

4. 0 Discussion

4.1. Why several kinds of evidence?

The main finding is that each of the three threat families leaves a different 'footprint'. A manipulation of a value removes a real balance without altering the noise of the instrument, a replacement of a stream

preserves real balance at each level of the stream but carries the wrong noise or a duplicate copy of the past to the instrument, and a real operational event satisfies both noise and real balance tests, but breaks time. This single anomaly score is sufficient to bring these differences together, so that the raw tag auto encoder, a good anomaly detector, correctly attributed the manipulation events less than 50% of the time. Some false alarms are acceptable in keeping evidence separate, as Section 3.4 demonstrated, but it's necessary if the alarm is to carry a diagnosis. This is also the structure of the utilities' response. If a pressure transmitter at junction 207 disagrees with hydraulics, or if that channel is noisy, an alert will be issued, which may indicate a calibration check, or manipulation of that channel. If an alert indicating that the pump 10 telemetry is the same as yesterday's comes up, it is likely a compromised communication path and should be confirmed by site visit. The alert message "The network has left its daily routine" when all instruments are "self-consistent" indicates that the network has gone into a burst or an unscheduled operation. These are all different work orders and the evidence in the framework makes it clear.

4.2. What the framework cannot see

It's not just the successes that are illustrative of the limits. First, it is difficult to detect falsification that is not in conflict with the physics. The mass balance of the tank 3 was the only thing that was built to satisfy for scenario DM4, and it was encountered in only one of the four events. Similarly, the hydraulically equivalent to stealthy false data injection in power systems (Liu et al., 2011, Teixeira et al., 2015): a defender's model also reveals to a capable attacker what to fake. The defence is to introduce instruments whose readings are difficult to forge regardless, such as an independent pressure logger which is close to tank 3 but reports via a different path; the remaining structure of Section 2.3 indicates where such instruments would fill the largest "holes.

Second, small leaks are below the resolution of system level evidence. Background leaks of 1 % to 2 % of consumption left no signature in any layer, which is consistent with the long experience of the water loss community that such leaks are found through district metering and night flow analysis at the scale of a few thousand properties (Farley and Trow, 2003; Puust et al., 2010). The framework does not replace district metered areas; in a network divided into them, the water balance and night flow residuals would be computed for each area and would gain sensitivity in proportion. Third, supervised attribution generalises to benign events better than to attacks. This asymmetry is not an artefact of this study; it follows from the fact that benign events share a physical character, consistency with the hydraulics and with the instrument’s noise, whereas attacks are defined by an adversary’s choices. In deployment the classifier’s output should never be used to suppress an alarm, and a low confidence attribution should be reported as unknown and treated as potentially malicious. Libraries generated systematically from a digital twin, open set classifiers that model the benign classes explicitly, and periodic red team exercises would all reduce the risk that a novel attack is filed as a fault.

Fourth, the evidence comes from simulation. The twin includes stochastic demand at three scales, seasonal drift, instrument specific noise and quantisation, pump controls and storage dynamics, and the calibration recovered the physical parameters from telemetry alone, but real networks add complications: model and topology errors, valves whose status is not telemetered, intermittent communications that cause gaps and held values, and transmitters replaced during the record. Some of these will widen residual distributions; held values after communication failures will resemble stuck transmitters and short replays, and a deployment would need to use the communication status flags that most SCADA systems provide. Validation on

historian records from an operating utility, ideally with staged events, is the necessary next step.

4.3. Implications for practice

Two practical points follow. The exact archive match is cheap to compute and almost free of false alarms on quantised data, yet it is not part of any SCADA product we are aware of; it should be. Its weakness, that a careful attacker adds fresh noise, is real, but the adaptive replay test shows that such an attacker still has to survive the physical consequences of whatever he is hiding. Second, the hydraulic balances in Section 2.3 require no hydraulic model, only telemetry that most utilities already collect. Tank plan areas and pump characteristics estimated from data also serve as continuous checks on instrument calibration, independent of any security purpose.

The framework is equally relevant to utilities with modest resources, including many in Africa and South Asia, where telemetry is often limited to source flows, tank levels and a handful of pressure points, and where security staff are few. The evidence layers degrade gracefully as instruments are removed; each balance needs only the tags it involves. Lastly, the same concepts apply to wastewater collection and treatment, with level, flow, pump and pressure telemetry following similar balances, and actions taken on pumping stations or overflow controls potentially resulting in spills as significant as any drinking water event.

5. 0 Conclusions

There are a variety of reasons why telemetry in water networks may be incorrect, such as manipulation, replacement, or a true change in the network; each of these scenarios leaves a unique signature. Hydraulic balances, which are calibrated from telemetry, reveal manipulated values while noise fingerprints and exact archive matches reveal substituted streams, and departures from the recent routine reveal a true upset. These evidence layers identified 85 % of test events and correctly identified a family for 93 % of the events; this is significantly higher than what could be achieved from the raw signals alone and most of the test events were localised to the instrument affected. This framework requires no hydraulic model, and is based on the telemetry data most utilities already have. The major limitations are the ability to respect the physics in falsification, and the low leakage of system types and attack types not found in the training library. The validation on operating networks, open set attribution, placement of tamper resistant witness instruments are clear priorities for further work.

6.0 References

Addeen, H. H., Xiao, Y., Li, J., & Guizani, M. (2021). A survey of cyber-physical attacks and detection methods in smart water distribution systems. IEEE Access, 9, 99905–99921. https://doi.org/10.1109/ACCESS.2021.3095713

Ahmed, C. M., Ochoa, M., Zhou, J., Mathur, A. P., Qadeer, R., Murguia, C., & Ruths, J. (2018). NoisePrint: Attack detection using sensor and process noise fingerprint in cyber physical systems. In Proceedings of the 2018 Asia Conference on Computer andCbommunications Security (ASIACCS ’18) (pp. 483–497). Association for Computing Machinery. https://doi.org/10.1145/3196494.3196532

Ahmed, C. M., Palleti, V. R., & Mathur, A. P. (2017). WADI: A water distribution testbed for research in the design of secure cyber physical systems. In Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWATER ’17) (pp. 25–28). Association for Computing Machinery. https://doi.org/10.1145/3055366.3055375

Cárdenas, A. A., Amin, S., Lin, Z. S., Huang, Y. L., Huang, C. Y., & Sastry, S. (2011). Attacks against process control systems: Risk assessment, detection, and response. In Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security (ASIACCS ’11) (pp. 355–366).

Association for Computing Machinery. https://doi.org/10.1145/1966913.1966959

Cybersecurity and Infrastructure Security Agency. (2021, October 14). Ongoing cyber threats to U.S. water and wastewater systems (Joint Cybersecurity Advisory AA21-287A). https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-287a

Deng, A., & Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), 4027–4035. https://doi.org/10.1609/aaai.v35i5.16523

Farley, M., & Trow, S. (2003). Losses in water distribution networks: A practitioner's guide to assessment, monitoring and control. IWA Publishing.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Giraldo, J., Urbina, D., Cárdenas, A., Valente, J., Faisal, M., Ruths, J., Tippenhauer, N. O., Sandberg, H., & Candell, R. (2018). A survey of physics-based attack detection in cyber-physical systems. ACM Computing Surveys, 51(4), Article 76. https://doi.org/10.1145/3203245

Habib, M. A., Jurcut, A. D., Ahmed, H., Wei, W., & Salauddin, M. (2026). Cybersecurity in water distribution networks: A systematic review of AI-based detection algorithms. Water, 18(4), Article 519. https://doi.org/10.3390/w18040519

Hassanzadeh, A., Rasekh, A., Galelli, S., Aghashahi, M., Taormina, R., Ostfeld, A., & Banks, M. K. (2020). A review of cybersecurity incidents in the water sector. Journal of Environmental Engineering, 146(5), Article 03120003. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001686

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507. https://doi.org/10.1126/science.1127647

Housh, M., & Ohar, Z. (2018). Model-based approach for cyber-physical attack detection in water distribution systems. Water Research, 139, 132–143. https://doi.org/10.1016/j.watres.2018.03.039

Idowu, K., Leslie, A., Twayana, S., Nitzan, Y., & Waheed, B. O. (2026). *Real-time detection of DDoS and phishing attacks for enhanced banking security*. *International Journal of Scientific Research in Science Engineering and Technology*. https://doi.org/10.32628/IJSRSET2613313

Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017). Anomaly detection for a water treatment system using unsupervised machine learning. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 1058–1065). IEEE. https://doi.org/10.1109/ICDMW.2017.149

Klise, K. A., Bynum, M., Moriarty, D., & Murray, R. (2017). A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study. Environmental Modelling & Software, 95, 420–431. https://doi.org/10.1016/j.envsoft.2017.06.022

Kravchik, M., & Shabtai, A. (2018). Detecting cyber attacks in industrial control systems using convolutional neural networks. In Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC '18) (pp. 72–83). Association for Computing Machinery. https://doi.org/10.1145/3264888.3264896

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

Liu, Y., Ning, P., & Reiter, M. K. (2011). False data injection attacks against state estimation in electric power grids. ACM Transactions on Information and System Security, 14(1), Article 13. https://doi.org/10.1145/1952982.1952995

Mathur, A. P., & Tippenhauer, N. O. (2016). SWaT: A water treatment testbed for research and training on ICS security. In 2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater) (pp. 31–36). IEEE. https://doi.org/10.1109/CySWater.2016.7469060

Mo, Y., & Sinopoli, B. (2009). Secure control against replay attacks. In 2009 47th Annual Allerton Conference on Communication, Control, and Computing (pp. 911–918). IEEE. https://doi.org/10.1109/ALLERTON.2009.5394956

Mounce, S. R., Boxall, J. B., & Machell, J. (2010). Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. Journal of Water Resources Planning and Management, 136(3), 309–318. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000030

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Puust, R., Kapelan, Z., Savić, D. A., & Koppel, T. (2010). A review of methods for leakage management in pipe networks. Urban Water Journal, 7(1), 25–45. https://doi.org/10.1080/15730621003610878

Qin, S. J. (2003). Statistical process monitoring: Basics and beyond. Journal of Chemometrics, 17(8–9), 480–502. https://doi.org/10.1002/cem.800

Romano, M., Kapelan, Z., & Savić, D. A. (2014). Automated detection of pipe bursts and other events in water distribution systems. Journal of Water Resources Planning and Management, 140(4), 457–467. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000339

Rossman, L. A. (2000). EPANET 2 users manual (EPA/600/R-00/057). United States Environmental Protection Agency.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Taormina, R., & Galelli, S. (2018). Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems. Journal of Water Resources Planning and Management, 144(10), Article 04018065. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000983

Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., & Ostfeld, A. (2017). Characterizing cyber-physical attacks on water distribution systems. Journal of Water Resources Planning and Management, 143(5), Article 04017009. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000749

Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., Ostfeld, A., Eliades, D. G., Aghashahi, M., Sundararajan, R., Pourahmadi, M., Banks, M. K., Brentan, B. M., Campbell, E., Lima, G., Manzi, D., Ayala-Cabrera, D., Herrera, M., Montalvo, I., Izquierdo, J., Luvizotto, E., Jr., Chandy, S. E., Rasekh, A., Barker, Z. A., Campbell, B., Shafiee, M. E., Giacomoni, M., Gatsis, N., Taha, A., Abokifa, A. A., Haddad, K., & Lo, C. S. (2018). Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. Journal of Water Resources Planning and

Management, 144(8), Article 04018048. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000969

Teixeira, A., Shames, I., Sandberg, H., & Johansson, K. H. (2015). A secure control framework for resource-limited adversaries. Automatica, 51, 135–148. https://doi.org/10.1016/j.automatica.2014.10.067

Tuptuk, N., Hazell, P., Watson, J., & Hailes, S. (2021). A systematic review of the state of cyber-security in water systems. Water, 13(1), Article 81. https://doi.org/10.3390/w13010081

Urbina, D. I., Giraldo, J. A., Cárdenas, A. A., Tippenhauer, N. O., Valente, J., Faisal, M., Ruths, J., Candell, R., & Sandberg, H. (2016). Limiting the impact of stealthy attacks on industrial control systems. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16) (pp. 1092–1105). Association for Computing Machinery. https://doi.org/10.1145/2976749.2978388

Data and code availability

Data will be made available upon request.

CRediT authorship contribution statement : Conceptualization, Methodology, Software, Formal analysis, Writing (original draft), Writing (review and editing). [Contributions of coauthors to be inserted.]

Declaration of generative AI and AI assisted technologies in the writing process

During the preparation of this work the authors used claude ai in order to developed some figures and improve resolution. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declarations

Consent for publication

Not Applicable

Data Availability Statement:

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflict of Interest

The authors have no competing interests to declare that are relevant to the content of this article.

Ethical Considerations

Not applicable

Funding

The authors declared no external source of funding

Authors' Contributions

All components of the work were carried out by the author