Applied Science, Computing and Energy, 2025, 3(3), 658-675

Development of Chemistry-Informed Machine Learning Models for Cyber Threat Detection and Protection of Water and Wastewater Infrastructure

Ibidun Isaac

21 July 2025/Accepted: 20 October 2025 /Published: 25 October 2025

Abstract: *Water treatment plants increasingly depend on networked control, and an intruder who manipulates dosing while replaying normal analyser readings can defeat monitors that judge each signal alone. We developed a chemistry informed machine learning (CIML) framework that turns the chemistry of treatment into consistency checks on supervisory control and data acquisition (SCADA) data. Carbonate equilibrium, Kohlrausch conductivity, Nernstian redox potential, first order chlorine kinetics and reservoir mass balance were expressed as eleven residuals computed from historian tags, with coefficients calibrated by least squares on normal operation. A dual channel detector combined a joint autoencoder on raw signals and residuals with a residual energy score, and a gradient boosted triage stage trained on a digital twin attack library separated attacks from benign anomalies. We tested it on a chemically explicit twin of a coagulation, filtration and chlorination plant over 110 test days containing seven attack scenarios, natural storms and four types of instrument fault or process upset. Calibration recovered dosing stoichiometry within 1 % and electrode and decay parameters within 5 %. Storms left every balance residual at baseline, whereas each attack left a distinct fingerprint. CIML with triage reached an area under the precision recall curve of 0.67 and an F1 score of 0.79, against 0.48 and 0.41 for the strongest data driven reference, with 0.19 false alarm episodes per day. A masked hypochlorite underdose missed by a raw signal autoencoder was detected 2.25 h after onset. Triage generalised poorly to unseen attack types, and attacks that leave the chemistry consistent remain hard to detect.*

Keywords: *Water treatment; Cyber physical security; Physics informed-ML; Anomaly detection*

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

1.0 Introduction

Drinking water and wastewater utilities are networkedly control. Coagulant and disinfectant pumps are driven by programmable logic controllers; supervisory control and data acquisition (SCADA) servers record every reading from each analyser; and once a convenience, remote access for small plants is now commonplace to access the treatment system when no one is in the control room. This connectivity that helps reduce operating cost has created greater attack surface (Odozor et al., 2025). In the water sector, incidents in the reviewed literature include opportunistic ransomware attacks on business networks, as well as manipulation of process setpoints (Hassanzadeh et al., 2020; Tuptuk et al., 2021). The stakes were real in February 2021 when an unauthorised user accessed an unauthorised system via remote desktop software, raised the sodium hydroxide setpoint by over two orders of magnitude, and only an operator who saw the cursor moving and reversing the change noticed it (CISA, 2021). On that occasion, the plant was guarded by someone who was watching a screen, and not by any automatic safety device.

That was an obvious occurrence. A more patient opponent would not put his finger on a setpoint which the operator can see, but would control a dosing pump and type yesterday's analyser readings into the

historian, making all the trends on the control room display look normal. Such deception attacks are well known in the control literature (Cárdenas et al., 2011; Urbina et al., 2016) and are able to defeat a monitor (that relies on the reported value of the variable being attacked). However, the physical impacts in a water plant are harsh. If either an underdose or an overdose of hypochlorite occurs, it can be masked and allow pathogens to pass the clearwell with a residual in the clearwell that would appear compliant, or a caustic overdose could result in water at a higher pH than the WHO recommends for both consumers and distribution pipe integrity (WHO, 2022).

Research on automated detection has grown quickly since the release of public testbeds and benchmark datasets. The Secure Water Treatment and Water Distribution testbeds at the Singapore University of Technology and Design (Mathur and Tippenhauer, 2016; Ahmed et al., 2017; Goh et al., 2017) and the Battle of the Attack Detection Algorithms (BATADAL) competition built on the C Town network (Taormina et al., 2017, 2018) gave the community common ground for comparison. On these data, unsupervised learners trained only on normal operation, including deep neural networks, one class support vector machines, convolutional networks and autoencoders, detect a large share of scripted attacks (Inoue et al., 2017; Kravchik and Shabtai, 2018; Taormina and Galelli, 2018), and graph based models have pushed accuracy further on multivariate sensor streams (Deng and Hooi, 2021). Model based approaches that encode hydraulic understanding of the network have performed at least as well on BATADAL, and they offer a clearer rationale for each alarm (Housh and Ohar, 2018). Surveys of the field converge on a familiar pair of weaknesses: purely data driven detectors raise too many false alarms when the process itself changes, and they rarely explain why an alarm was raised (Giraldo et al., 2018; Addeen et al., 2021). The application of machine-learning for water infrastructure security against cyber-physical attacks has been further pushed forward by recent studies. Chandran and Sunil (2025) created a rectilinear hybrid belief classifier network for cyber-physical attacks in water distribution networks, emphasizing the importance of learning-based approaches in infrastructure protection. In the same manner, Sayghe (2025) introduced a digital-twin based intrusion detection system based on process simulation and physical residuals and machine learning for detecting attacks in industrial SCADA systems. The progress made in this area highlights the benefits of combining process behaviour with the data-driven detection, but the adoption of treatment-specific chemical relationships to separate malicious manipulation from natural water-quality variation is an important research topic.

Both weaknesses are acute in water treatment, because the process is driven by a source that is never stationary. A storm can raise raw water turbidity tenfold within hours, depress alkalinity and push the coagulant dose, the coagulated pH and the chlorine demand to values never seen in the training record (Adeyemi et al., 2025). To a detector that learns only the statistical envelope of historian tags, the plant during a storm looks very much like a plant under attack. Instrument faults compound the problem. Fouled redox electrodes, drifting pH probes and chlorine analysers that flatline after a reagent failure are everyday events at any utility, and an alarm system that cannot distinguish them from intrusions will soon be ignored.

Hydraulic models have been the main source of physical knowledge in this literature, which is natural given that most benchmarks concern distribution networks. Treatment plants, however, carry a far richer and largely unexploited body of knowledge: their chemistry. The pH of coagulated water is fixed, given the raw water carbonate system, by the amount of alkalinity that the coagulant consumes (Adeyemi et al., 2025). The

finished water pH follows from the caustic and hypochlorite doses through the same equilibria. Conductivity rises by amounts that the limiting molar conductivities of the added ions predict. The oxidation reduction potential (ORP) of chlorinated water is a Nernstian function of the hypochlorous acid activity, which in turn depends on free chlorine, pH and temperature (Morris, 1966; Copeland and Lytle, 2014). Free chlorine decays in the clearwell and the network according to first order kinetics whose rate constants scale with organic matter and temperature (Rossman et al., 1994; Vasconcelos et al., 1997). Each of these relations links several independently measured signals. An attacker who falsifies one signal therefore creates a contradiction somewhere else, unless he understands the chemistry well enough to falsify every linked signal consistently.

The broader movement toward physics informed and theory guided machine learning provides a natural vehicle for this knowledge (Karpatne et al., 2017; Raissi et al., 2019; Karniadakis et al., 2021; Willard et al., 2022). Its value in environmental systems has been shown for problems such as lake temperature prediction, where process knowledge improves generalisation beyond the conditions seen in training (Read et al., 2019). In cyber physical security, however, physical knowledge has mostly been injected as hydraulic constraints or generic control invariants, and the aqueous chemistry of treatment has received little attention (Idowu et al., 2025). We are not aware of a detector that uses carbonate equilibrium, redox potential, ionic conductivity and disinfectant kinetics together as a consistency check on SCADA data, nor of one that separates attacks from benign chemical and instrumental upsets on that basis.

This study develops and evaluates such a framework, which we term chemistry informed machine learning (CIML). The work had three aims. The first was to translate standard treatment chemistry into a set of residuals that can be computed from historian tags alone, with free parameters calibrated on normal operating data rather than assumed. The second was to combine those residuals with a learned representation of the raw signals in a detector whose alarms are traceable to specific chemical contradictions, and to add a triage stage that decides whether an anomaly looks like an attack or like a storm, a fouled electrode or a failed analyser. The third was to establish, with a deliberately adversarial evaluation, where the approach stops working: how it degrades with sensor noise, how well it transfers to attack types it has not seen, and what an attacker must falsify to defeat it. Because no utility can supply labelled attacks on its own plant, the framework was developed and tested on a chemically explicit digital twin of a conventional surface water treatment plant and its chlorinated network, following the simulation practice established by BATADAL. All code is written so that the same pipeline runs unchanged on plant historian exports.

2. 0 Materials and methods

2.1. Plant configuration and simulated historian

The reference plant treats a turbid tropical river through alum coagulation, sedimentation and rapid gravity filtration, followed by pH correction with sodium hydroxide, chlorination with sodium hypochlorite and a 5000 m^3 clearwell (Fig. 1). Treated water passes to a service reservoir with a plan area of 1500 m^2 and then to a network monitored at three nodes whose nominal travel times from the clearwell are 2, 6 and 14 h at a plant flow of 1200 $m^3\ h^{-1}$. The configuration and dosing ranges were chosen to match common practice in conventional treatment (Crittenden et al., 2012).

Exogenous inputs were generated stochastically. Water temperature followed an annual cycle about 27 °C with a diurnal component and autoregressive noise. Storm events arrived as a Poisson process at a mean rate of about one every 11 days, each

producing a rapid turbidity rise to a lognormally distributed peak with a median of 180 NTU followed by an exponential recession over 14 to 36 h, together with correlated increases in total organic carbon (TOC) and decreases in alkalinity, pH and conductivity. Consumer demand followed diurnal and semidiurnal cycles with 2 % multiplicative noise.

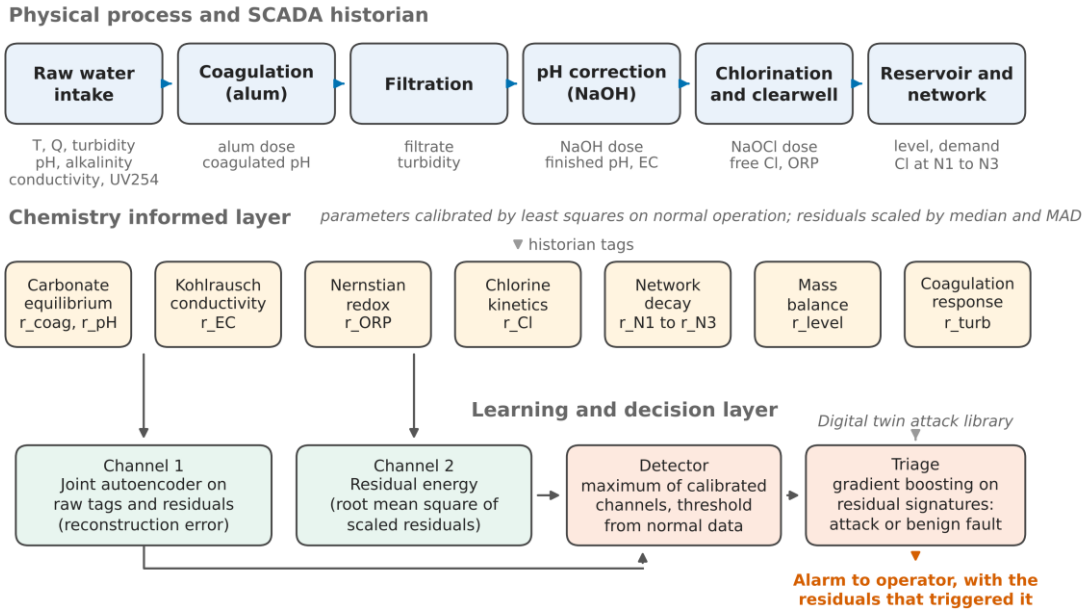


Fig. 1. Architecture of the chemistry informed machine learning (CIML) framework. Historian tags from six process stages feed a residual layer built on carbonate equilibrium, Kohlrausch conductivity, Nernstian redox potential, chlorine kinetics, network decay, reservoir mass balance and coagulation response. A joint autoencoder (Channel 1) and the residual energy (Channel 2) form the detector, and a triage classifier trained on a digital twin attack library decides whether an alarm reflects an attack or a benign event.

The process model was solved at the 5 min historian interval. Coagulated water chemistry was obtained by holding total carbonate constant through coagulation and removing 0.505 mg L^{-1} of alkalinity (as $CaCO_3$) per mg L^{-1} of commercial alum, $Al_2(SO_4)_3 \cdot 14H_2O$, which is the stoichiometric consumption of six equivalents of bicarbonate per mole (Benjamin, 2015). The pH was then found by solving the carbonate charge balance. Sodium hydroxide added 1.25 mg L^{-1} of alkalinity per mg L^{-1}, and the finished water pH was computed in the same way with a small allowance for carbon dioxide loss. Filtrate turbidity responded to any shortfall of the alum dose relative to the dose required by the raw water. Free chlorine in the clearwell followed a continuously stirred tank balance with an instantaneous demand and first order bulk decay, both proportional to filtered TOC and corrected for temperature with an Arrhenius factor of 1.07 per °C (Rossman et al., 1994; Vasconcelos et al., 1997). Network concentrations followed from the clearwell history through travel time delays and combined bulk and wall decay. ORP was generated from a Nernstian dependence on hypochlorous acid, whose fraction was obtained from the temperature dependent dissociation constant of Morris (1966), with an electrode slope of 74 mV per decade, typical of platinum electrodes in chlorinated water (Copeland and Lytle, 2014). Conductivity increments followed from the limiting molar conductivities of the added and consumed ions.

Two proportional integral controllers closed the loops, one trimming caustic to a finished pH setpoint and one trimming hypochlorite to a clearwell residual setpoint, and a level controller set the plant flow from the reservoir level. Operators changed the pH setpoint (7.5 to 7.8) and the chlorine setpoint (1.0 to 1.5 mg L^{-1}) every 3 to 7 days, which gave the normal record realistic excitation. Twenty one tags were archived (Fig. 1), including an online UV absorbance at 254 nm as a surrogate for organic matter, with Gaussian noise representative of process analysers (for example 0.012 pH units, 0.025 mg L^{-1} free chlorine and 4 mV ORP). The main record covered 180 days, or 51 840 samples per tag. Days 1 to 56 were used for training and days 57 to 70 for validation; neither contained attacks or instrument faults, although both contained natural storms. Days 71 to 180 formed the test period.

2.2. Threat model and scenario design

The adversary was assumed to have gained access to the control network, as at Oldsmar, and to be able to change actuator commands and to replace sensor values on their way to the historian and the controllers. Seven attack scenarios were defined (Table 1). They span overt actuator manipulation (A1, A4), deception attacks that combine actuator manipulation with replay of the affected sensor (A2, A3, A6, A7) and an integrity attack that biases a sensor so that the controller itself drives the process to an unsafe state (A5). Replay used the value recorded exactly 24 h earlier, which preserves diurnal patterns and is a stronger deception than a frozen value.

Detection is only half the problem in practice. The test period therefore also contained five classes of benign anomaly (Table 1): storm driven turbidity surges (F1), slow fouling of the ORP electrode (F2), calibration drift of the finished water pH electrode (F3), flatlining of the clearwell chlorine analyser (F4) and an algal bloom that raised chlorine demand (F5). These events are labelled as negatives in every metric reported below, so that a detector earns no credit for alarming on them. Four instances of each attack and of F2 to F5 were placed at random, non overlapping times in the test period, with magnitudes and durations drawn from the ranges in Table 1. Storms occurred naturally, giving 16 storm episodes in the test period. Attack samples made up 9.7 % of the test record and benign anomalies 29.4 %.

A second, independent 180 day record was generated with a different random seed, an earlier start of events, wider magnitude ranges and seven instances of every attack and fault. This record, the twin attack library, was used only to train the triage classifier described in Section 2.4. It never contributed to the training of detectors or to the setting of thresholds.

Table 1: Attack scenarios and benign events used in the evaluation. Magnitudes multiply the controller command unless stated otherwise; replayed signals repeat the value recorded 24 h earlier.

ID	Scenario	Element manipulated	Deception presented to SCADA	Duration (h)	Magnitude
A1	Caustic overdose through HMI setpoint change	NaOH dosing command	None	2 to 6	2.2 to 3.5 times
A2	Caustic overdose with replayed pH	NaOH dosing command	Finished pH replayed; dose readback set to controller command	6 to 18	1.8 to 2.6 times

ID	Scenario	Element manipulated	Deception presented to SCADA	Duration (h)	Magnitude
A3	Hypochlorite underdose with spoofed residual	NaOCl pump	Clearwell free chlorine reported at setpoint; dose readback set to command	6 to 16	0.25 to 0.45 times
A4	Hypochlorite overdose through pump override	NaOCl pump	None	3 to 8	2.0 to 2.8 times
A5	Ramp bias on clearwell chlorine analyser	Free chlorine signal used by the controller	Bias grows linearly over 60 % of the event	10 to 24	+0.35 to +0.60 mg L^{-1}
A6	Coagulant underdose with frozen filtrate turbidity	Alum pump	Filtrate turbidity replayed; dose readback set to command	6 to 14	35 % to 55 % reduction
A7	Reservoir level replay with forced pumping	Plant flow	Reservoir level replayed	3 to 6	+270 to +450 m^3 h^{-1}
F1	Storm driven turbidity surge	Raw water quality	Not applicable	14 to 36	Peak median 180 NTU; duration is the recession time
F2	ORP electrode fouling	ORP signal	Not applicable	24 to 60	Linear drift to −60 to −110 mV
F3	pH electrode calibration drift	Finished pH signal	Not applicable	24 to 60	Linear drift to +0.15 to +0.30
F4	Chlorine analyser flatline	Free chlorine signal	Not applicable	3 to 10	Reading near zero
F5	Algal bloom	Raw water TOC	Not applicable	24 to 48	+1.0 to +2.0 mg L^{-1} TOC

2.3. Chemistry informed residual layer

The residual layer converts historian tags into quantities that should be close to zero whenever the plant obeys its chemistry and its instruments tell the truth. Every residual is computed from measured tags only, and every free parameter is fitted by linear or nonlinear least squares on the training period. No quantity known only to the simulator enters the residuals.

Carbonate balance through coagulation. Total carbonate C_T is obtained from raw water pH and alkalinity A_r (eq L^{-1}) through the charge balance

$$A = C_T\,(\alpha_1 + 2\alpha_2) + \frac{K_w}{[\mathrm{H}^+]} - [\mathrm{H}^+], \qquad (1)$$

where α_1 and α_2 are the bicarbonate and carbonate fractions defined by the first and second acidity constants of carbonic acid (Stumm and Morgan, 1996). The predicted coagulated pH, $\widehat{\mathrm{pH}}_c$, is the root of Eq. (1) with $A = A_r - \theta_1 D_{\mathrm{alum}}$, and the residual is $r_{\mathrm{coag}} = \mathrm{pH}_c - \widehat{\mathrm{pH}}_c - \theta_0$. The coefficient θ_1 is the apparent alkalinity consumption per unit coagulant dose.

Finished water pH balance. With alkalinity $A_f = A_r - \theta_1 D_{\mathrm{alum}} + \theta_2 D_{\mathrm{NaOH}} + \theta_3 D_{\mathrm{Cl}}$ and a carbonate retention factor θ_4, the residual r_{pH} is the difference between the measured

finished pH and the root of Eq. (1) evaluated at $(\theta_4 C_T, A_f)$.

Ionic conductivity balance. Following Kohlrausch's law of independent ionic migration, the increase in electrical conductivity across the plant should be additive in the doses, $r_{\mathrm{EC}} = \kappa_f - \kappa_r - (\beta_1 D_{\mathrm{alum}} + \beta_2 D_{\mathrm{NaOH}} + \beta_3 D_{\mathrm{Cl}} + \beta_0)$. (2)

Redox consistency. The hypochlorous acid concentration is $[\mathrm{HOCl}] = C\,\alpha_{\mathrm{HOCl}}$, with $\alpha_{\mathrm{HOCl}} = \left(1 + 10^{\,\mathrm{pH} - \mathrm{p}K_a(T)}\right)^{-1}$ and $\mathrm{p}K_a(T) = 3000/T - 10.0686 + 0.0253\,T$ for absolute temperature T (Morris, 1966). The ORP residual is $r_{\mathrm{ORP}} = E - \left(\gamma_0 + \gamma_1 \log_{10}[\mathrm{HOCl}] + \gamma_2\,\mathrm{pH}_f + \gamma_3\,T_c\right)$, (3) where E is the measured ORP and T_c the temperature in °C. The coefficient γ_1 is the empirical electrode slope.

Clearwell chlorine kinetics. The expected clearwell residual $\hat{C}$ is propagated forward from the measured hypochlorite dose and flow, $\hat{C}_{t+1} = \hat{C}_t + \Delta t \left[\frac{\max(D_{\mathrm{Cl},t} - d_t,\, 0) - \hat{C}_t}{\tau_t} - k_t \hat{C}_t\right]$, (4) with hydraulic residence time $\tau_t = V/Q_t$, demand $d_t = \delta_0 + \delta_1\,u_t$ and decay constant $k_t = \delta_2\,(u_t/u_0)\,1.07^{\,T_c - 20}$, where u_t is the UV254 absorbance and $u_0 = 0.07$ cm^{-1}. The residual is $r_{\mathrm{Cl}} = C_t - \hat{C}_t$. Because Eq. (4) is driven only by measured inputs and is stable, prediction errors do not accumulate.

Network decay. For node j with travel time $t_j = t_j^0\,Q^0/Q_t$, the expected concentration is the clearwell value t_j earlier multiplied by $\exp\left[-\left(k_{w,j} + k_b\,u_t\,1.07^{\,T_c - 20}\right) t_j\right]$, which separates wall and bulk contributions. The residuals are r_{N1} to r_{N3}.

Reservoir mass balance and coagulation response. The level residual is the hourly sum of the difference between the observed level change and $(Q_{\mathrm{in}} - Q_{\mathrm{dem}})\Delta t/S$, with the plan area S estimated from data. A coagulation response residual compares log filtrate turbidity with a regression on log raw turbidity and alum dose.

The Langelier saturation index (Langelier, 1936) was also computed from measured pH, temperature, conductivity derived dissolved solids and alkalinity. It is a state descriptor rather than a balance and was therefore excluded from the residual energy defined below, although it was supplied to the learning layer.

Each residual was centred on its training median and divided by 1.4826 times its median absolute deviation, giving robust scores z_i that are near unit variance under normal operation.

2.4. Learning and decision layer

The detector has two channels (Fig. 1). Channel 1 is an autoencoder (Hinton and Salakhutdinov, 2006) trained to reconstruct a joint vector of the 21 standardised raw tags, their changes over 30 and 60 min, the eleven residual scores and their changes over 60 min. It uses a symmetric multilayer perceptron with 56, 14 and 56 hidden units, rectified linear activations, Adam optimisation with early stopping and an L2 penalty of 10^{-4}, trained on 14 000 samples drawn at random from the training period. Its anomaly score is the mean squared reconstruction error. Channel 2 is the residual energy, the root mean square of the ten balance residual scores, which responds directly to chemical contradictions whether or not the raw signals look unusual.

Each channel score was normalised by its median and interquartile range on validation samples labelled normal, smoothed with a 30 min moving average and divided by its own 99.75th percentile on those samples. The detector score is the larger of the two calibrated channels, and an alarm is raised when its smoothed value exceeds the 99.5th percentile of the validation distribution. Taking the maximum preserves sensitivity to overt manipulations that disturb the raw signals without breaking any balance, which only Channel 1 sees, and to deceptions that break a balance while leaving each raw signal plausible, which Channel 2 sees more clearly.

The triage stage is a gradient boosted tree classifier (Friedman, 2001), implemented as a histogram based ensemble with 300 iterations, a learning rate of 0.06 and 31 leaves per tree. Its inputs are the residual scores, their rolling 1 h standard deviations and 1 h changes, and the standardised raw tags with their rolling standard deviations. It was trained on all anomalous samples of the twin attack library to output the probability that an anomaly is an attack rather than a benign event. The probability was smoothed over 1 h, and the final CIML alarm requires both a detector alarm and a triage probability above 0.5. Residual signatures carry the information that matters for this decision. A fouled ORP electrode breaks only the redox balance, a drifting pH electrode breaks the pH and redox balances in fixed proportion, and a flatlined analyser produces a residual with almost no variance, whereas the attacks in Table 1 produce coordinated patterns across several balances.

2.5. Benchmarks and evaluation

Four detectors served as references. Three are widely used data driven monitors trained on the same windowed raw tag vectors: the squared prediction error of principal component analysis retaining 95 % of variance (Qin, 2003), an isolation forest with 300 trees (Liu et al., 2008) and an autoencoder on raw tags with 48, 12 and 48 hidden units. The fourth, labelled mechanistic residuals, raises an alarm on the largest absolute residual score and represents a purely model based detector with no learning. To isolate the contribution of chemistry to triage, the raw tag autoencoder was also paired with an identical classifier trained on raw tag features only. All thresholds were set by the same rule, the 99.5th percentile of the smoothed score on validation samples labelled normal, so that every detector was given the same nominal false alarm budget.

Performance was measured on the test period at two levels. At sample level we report the area under the receiver operating characteristic curve (AUROC), the area under the precision recall curve (AUPRC), which is more informative than AUROC when positives are rare (Saito and Rehmsmeier, 2015), and precision, recall and F1 at the operating threshold. False positive rates are reported separately for normal samples and for samples within benign anomalies. At event level we report the fraction of attack events that raised at least one alarm, the median time from attack onset to first alarm and the number of false alarm episodes per day, an episode being a contiguous run of alarms that does not overlap an attack.

Three further experiments probed generalisation and robustness. In a leave one scenario out test, the triage classifier was retrained seven times with one attack type removed from the library and evaluated on test events of the withheld type. In a noise experiment, five additional 110 day records were generated with sensor noise scaled by factors from 0.5 to 3, and all models were retrained on days 1 to 40 and validated on days 41 to 50 of each. In an informed adversary experiment, four 70 day records contained eight instances of scenario A2 in which the attacker falsified progressively more signals: the finished pH alone; pH and ORP; pH, ORP and conductivity; and finally these three together with the chlorine readings at all network nodes. Models were implemented in Python with scikit learn 1.9 (Pedregosa et al., 2011). All random seeds were fixed and every number reported here is regenerated by the scripts in the code repository.

3.0 Results

3.1. Calibration recovers the stoichiometry of treatment

Least squares calibration on 56 days of normal historian data recovered the process parameters encoded in the simulator closely (Table 2). The apparent alkalinity consumption of the coagulant was 0.504 mg L^{-1} as $CaCO_3$ per mg L^{-1} of alum, against a stoichiometric value of 0.505, and the alkalinity contribution of caustic was 1.244 against 1.25. The increments of conductivity

per unit dose were in error range of 6 % with respect to their Kohlrausch values. The slope of the fitted ORP electrode was 73.2 mV/decade of hypochlorous acid (HCl), and the slope of the fitted pH electrode was −24.4 mV/unit, both of which were within 2 % of the values that were generated. The kinetic calibration separated out the instantaneous demand from the bulk decay, and returned the value of 0.087 mg L⁻¹ per mg L⁻¹ of TOC equivalent, which was the value implied by the simulator, and the three wall decay constants matched their generating values within 5 %.

The comparisons are important for two reasons. They demonstrate that the residuals are not random regressions, but rather balances that are physically meaningful and have coefficients which can be compared to literature and stoichiometry by a process engineer. They also demonstrate that the normal operation of the setpoint changes are sufficient to excite the plant for identification and no special plant trials are required for identification. The ORP slope identification was the most difficult case, and, in an earlier configuration, a fixed chlorine setpoint, a poor identification of the slope was obtained and the extrapolation of the redox residual at high pH was bad, hence the need for the operational setpoint changes in the design.

Table 2: Recovery of process parameters by least squares calibration on 56 days of normal historian data. Simulator values derive from stoichiometry, limiting ionic conductivities or literature electrode and decay constants.

Balance	Parameter	Simulator value	Calibrated estimate
Carbonate, coagulation	Alkalinity consumed per unit alum (mg $CaCO_3$ per mg)	0.505	0.504
Carbonate, finished water	Alkalinity added per unit NaOH (mg $CaCO_3$ per mg)	1.250	1.244
	Carbonate retention after treatment	0.970	0.971
Conductivity	Increment per unit alum (µS cm^{-1} per mg L^{-1})	0.330	0.316
	Increment per unit NaOH	2.150	2.187
	Increment per unit NaOCl (as Cl_2)	1.620	1.571
Redox	Electrode slope (mV per decade HOCl)	74.0	73.2
	pH coefficient (mV per pH unit)	−24.0	−24.4
	Temperature coefficient (mV per °C)	0.90	0.92
Chlorine kinetics	Demand per mg L^{-1} TOC equivalent	0.087	0.087
	Bulk decay constant at reference conditions (h^{-1})	0.028	0.029
Network decay	Wall decay, node N1 (h^{-1})	0.010	0.0095
	Wall decay, node N2 (h^{-1})	0.015	0.0145
	Wall decay, node N3 (h^{-1})	0.020	0.0193
Mass balance	Reservoir plan area (m^2)	1500	1437

3.2. Each attack leaves a chemical fingerprint

Under normal operation all residual scores had medians of about 0.7, as expected for robust scores of a half normal magnitude (Fig. 2, top row). Storms, the most disruptive natural events in the record, left every balance residual at that baseline; the largest median score during storms was 0.75. In other words, rises in turbidity of up to 45 times the dry

weather median, falls in alkalinity of up to 42 mg L^{-1} and coagulant dose increases of up to 85 % were all accounted for by the chemistry. This invariance is the main reason the chemistry informed detector suffers few storm alarms.

Every attack broke at least one balance, and the pattern of broken balances was specific to the attack (Fig. 2). The stealthy caustic overdose (A2) broke the redox balance (median score 32) and the conductivity balance (8.4) but left the pH balance almost untouched, because the replayed pH and the spoofed dose readback were mutually consistent; the contradiction appeared only in signals the attacker had not falsified. The underdose (A3) of masked hypochlorite at the first network node during the travel time caused the breakdown of the redox balance and after the travel time the decay balance. The most significant attack on clearwell kinetic balance was the analyser bias attack (A5) where the controller decreased the dose and the biased analyser continued to report the setpoint. The carbonate balances were used to expose the coagulant attack (A6): Underdosing will leave more alkaline and thus a coated PH than indicated by the dose of the coagulant. Only the reservoir mass balance was violated with the level replay (A7). The two manifest acts of attacks were different. The setpoint driven caustic overdose (A1) did not cause any imbalance as all the signals were related to the real process, but it was evident from the Langelier index and from the raw signals. The hypochlorite overdose (A4) was also chemically consistent, and resulted in very little response remaining.

Fingerprints were also found on the instrument due to benign instrument faults, but in a different shape. On the whole, ORP fouling (F2) produced the largest residuals in all the balances, with no apparent trend, pH drift (F3) caused pH and, less strongly, redox imbalance while the analyser flatlined (F4) caused very large but relatively constant residuals on all the balances utilizing clearwell chlorine. The algal bloom (F5), which resulted in an actual demand increase, kept the kinetic and network balances close to baseline levels due to the UV254 surrogate being representative of the increased organic load.

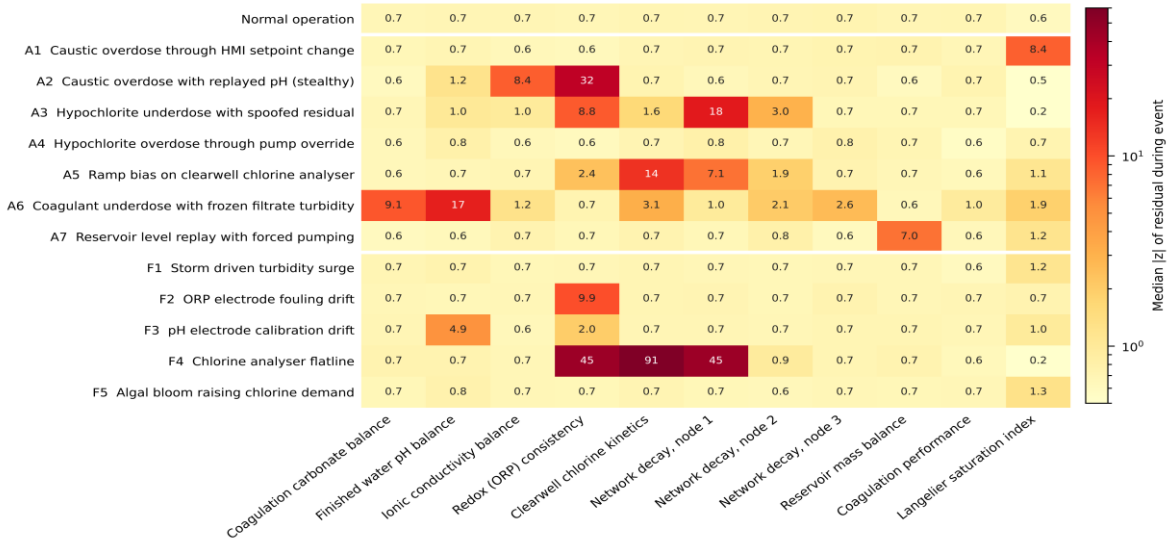


Fig. 2: Residual fingerprints. Median absolute robust score of each residual during normal operation, during each attack scenario (A1 to A7) and during each class of benign event (F1 to F5) in the test period. The colour scale is logarithmic; values near 0.7 correspond to normal operation.

3.3. A masked disinfection failure

The following is an example of scenario A3 (see Figure 3). The attacker initially reduced the hypochlorite feed to 0.78 mg L^{-1} of Cl2 and started to report the setpoint residual and the normal dose. The actual clearwell residual followed the exponential decay curve defined by the residence time and achieved a value of 0.33 mg L^{-1} after approximately 12 h, and was maintained at a fairly stable level around 1.2 mg L^{-1} all the time (Fig. 3a, b).

The measured ORP, which the attacker had not falsified, fell steadily away from the value implied by the reported chlorine, pH and temperature, by 29 mV on average over the event (Fig. 3c). The chemistry informed detector crossed its threshold 2.25 h after onset and remained in alarm until the feed was restored (Fig. 3d).

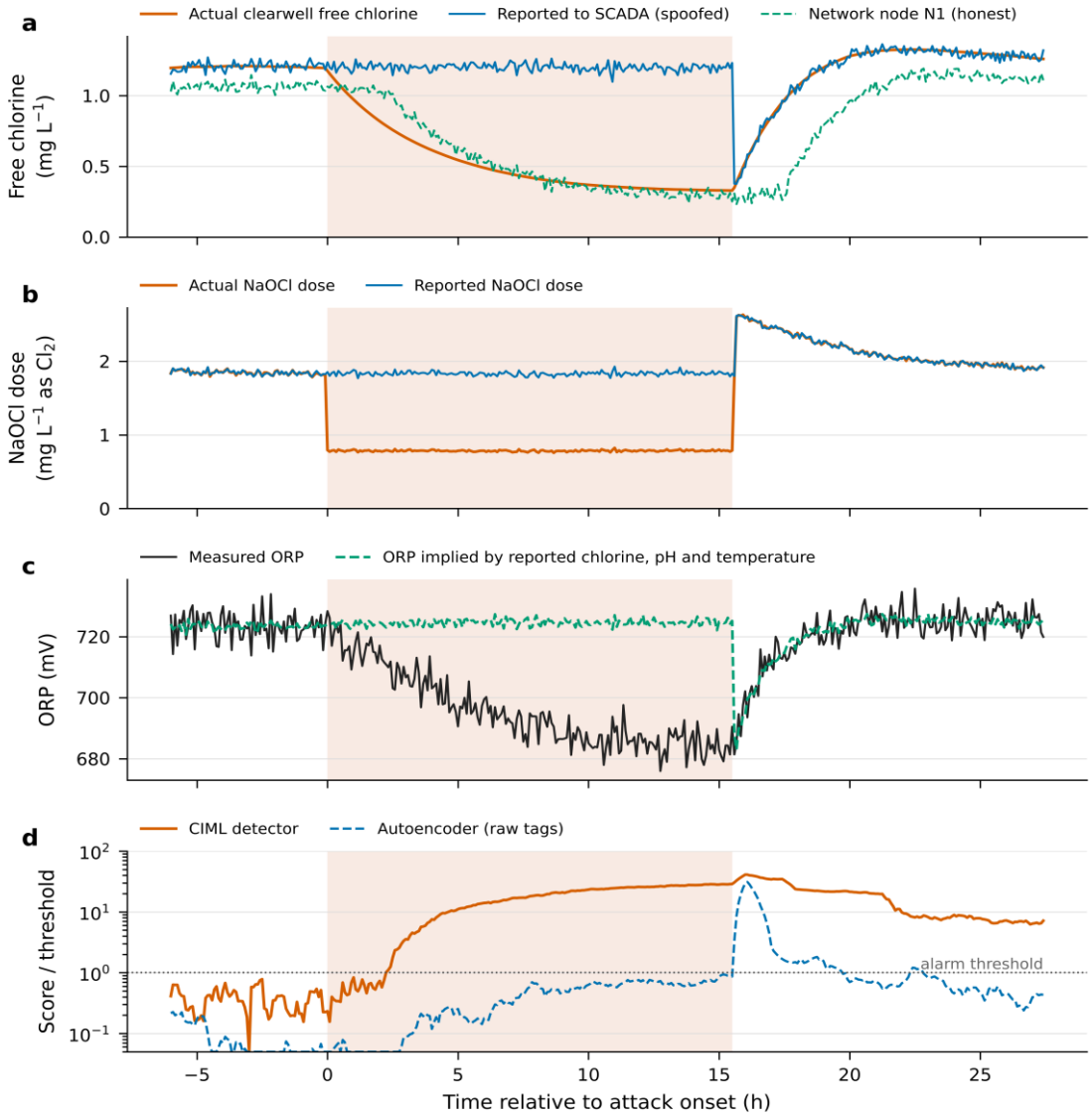


Fig. 3. A masked hypochlorite underdose (scenario A3, a test event lasting 15.5 h). (a) True clearwell free chlorine, the value reported to SCADA and the honest reading at network node N1. (b) True and reported hypochlorite dose. (c) Measured ORP and the ORP implied by the reported chlorine, pH and temperature through Eq. (3). (d) Smoothed detector scores divided by their alarm thresholds, on a logarithmic scale. Shading marks the attack.

The raw tag autoencoder never crossed its threshold during the attack: a gradual 40 mV drift in ORP, within the range produced by temperature and setpoint changes, did not stand out among 21 normal looking signals. Its score rose only when the dose was restored and the controller overshot, many hours after the network had received underchlorinated water; the first network node recorded 0.24 mg L^{-1} and the most distant node fell below 0.05 mg L^{-1}.

3.4. Detection performance

Across the full test period, the chemistry informed detector with triage achieved the highest AUROC, AUPRC, precision and F1 score of all methods and the lowest rate of false alarm episodes (Fig. 4, Table 3). Its AUROC was 0.950 and its AUPRC 0.673, compared with 0.261 for the raw tag autoencoder and 0.483 for the same autoencoder with a triage classifier trained on raw tag features. At the operating threshold it reached a precision of 0.79 and a recall of 0.79, giving an F1 score of 0.79, roughly double the best reference method. It also raised the fewest false alarms: 0.19 false alarm episodes per day, or about one every five days, against 0.86 to 1.94 per day for the detectors without triage.

The behaviour of the reference methods is instructive. With thresholds set on storm free validation data, the data driven detectors alarmed on 15 % to 48 % of samples within benign anomalies, mostly during storms, and the purely mechanistic residual detector alarmed on 44 % of them, almost entirely during instrument faults. The chemistry informed detector without triage had the highest recall of any method (0.91) and detected every attack event, but its false positive rate within benign anomalies was similar to that of the mechanistic detector. Triage reduced that rate from 45.5 % to 0.6 % while retaining 87 % of detected attack samples. Adding the same triage to the raw tag autoencoder also suppressed benign alarms but cut its recall to 0.32, because raw tag features separate attacks from benign anomalies less cleanly (Section 3.6). The precision recall curves show the practical consequence: across the range of recall from 0.2 to 0.8 the chemistry informed detector with triage keeps precision between 0.70 and 0.80, whereas every other method stays below 0.61 (Fig. 4b).

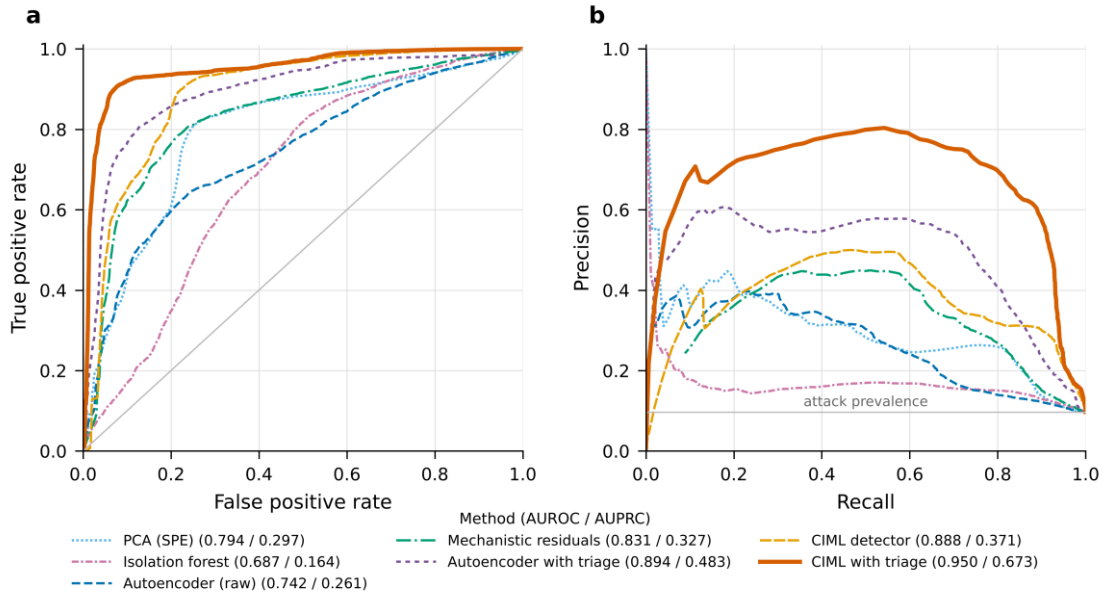


Fig. 4. (a) Receiver operating characteristic and (b) precision recall curves of all methods over the 110 day test period, with benign anomalies counted as negatives. The legend gives AUROC and AUPRC.

3.5. Scenario level behaviour and timeliness

Event level results differed sharply between scenarios (Fig. 5). The raw tag autoencoder detected the overt attacks A1 and A4 at once and the scenarios with large raw signal excursions, A2, A6 and A7, quickly. It detected only one of four analyser bias attacks (A5), after more than 10 h, and missed one of four masked underdoses (A3), detecting two of the others only 6 to 8 h after onset. The mechanistic residual detector showed the opposite profile. It caught every deception attack but only half of the overt caustic attacks and three quarters of the overt hypochlorite attacks, since these violate no balance.

Table 3. Detection performance on the 110 day test period (28 attack events). Benign anomalies are counted as negatives. FPR, false positive rate; FA, false alarm episodes.

Method	AUROC	AUPRC	Prec.	Recall	F1	FPR normal	FPR benign	Event recall	FA per day
PCA (squared prediction error)	0.794	0.297	0.26	0.78	0.39	0.12	0.48	1.00	1.48
Isolation forest	0.687	0.164	0.15	0.19	0.17	0.04	0.28	0.64	1.02
Autoencoder, raw tags	0.742	0.261	0.32	0.45	0.37	0.08	0.15	0.86	0.86
Mechanistic residuals	0.831	0.327	0.28	0.79	0.41	0.12	0.44	0.89	1.94
Autoencoder, raw tags, with triage	0.894	0.483	0.57	0.32	0.41	0.04	0.00	0.75	0.29
CIML detector	0.888	0.371	0.29	0.91	0.44	0.13	0.46	1.00	1.45
CIML with triage (proposed)	**0.950**	**0.673**	**0.79**	**0.79**	**0.79**	**0.03**	**0.01**	**0.93**	**0.19**

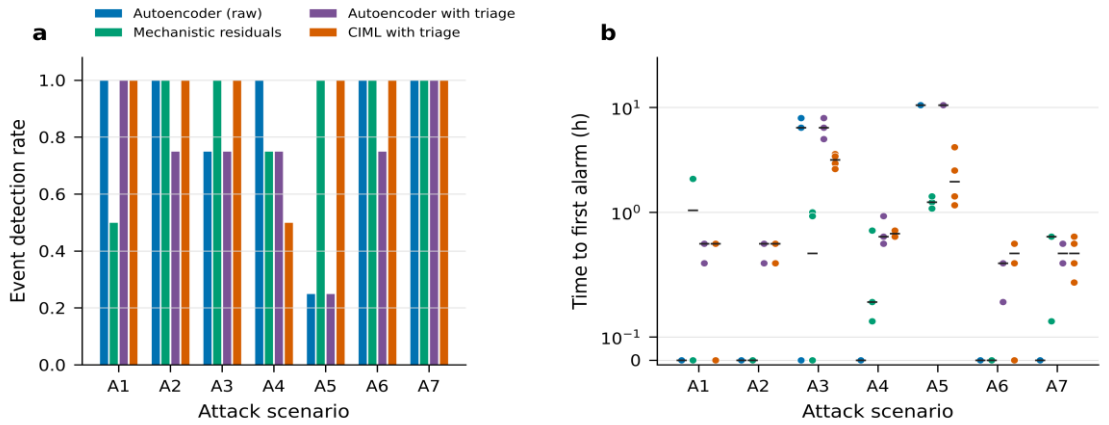


Fig. 5: Scenario level performance. (a) Fraction of the four test events of each scenario that raised at least one alarm. (b) Time from attack onset to first alarm for each detected event; horizontal bars mark medians, and the axis is linear below 0.5 h and logarithmic above.

The chemistry informed detector with triage detected all four events in six of the seven scenarios, combining the strengths of both. Its median detection time was about 0.5 h for the replay attacks on caustic, coagulant and level A2, A6, A7), 2.0 h for the analyser bias (A5) and 3.2 h for the masked underdose (A3).

The last figure is set by chemistry rather than by the detector: the clearwell behaves as a mixed tank with a residence time of about 4 h, so the true residual, and with it the ORP, declines gradually. Alarms for A3 came soon after the first underchlorinated water was detected at the nearest monitored node, but well before the residual was depleted at distant nodes, leaving ample opportunity for corrective dose(s) and/or flushing.

There are two caveats for this picture. The initial triage stage is time consuming first of all. The attack probability is smoothed over an hour, resulting in a median delay to the first alarm of ~0.5 h for detectors subjected to overt attacks, with a median delay of near zero for detectors with no triage. It does not matter if the chemical attack is slow or gradual, but for a sudden chemical attack of a caustic type, the sum of the time to detect the attack and the response time determines the exposure, and a utility may decide not to go through the triage process whenever the Langelier index or pH crosses the regulatory envelope. Second, that there is zero delay for some of the reference detectors is also due in part to the high false alarm rates of these detectors; an alarm that is on at the start of the attack is classified as a real-time detection.

The least favorable event was overt hypochlorite overdose (A4) which occurred in two of four events following triage. The chemistry was not significantly affected by the overdose; the beer was still within normal chemistry, despite the high residual (about 2 mg/L) in the clearwell. 43 % of the flagged samples were assigned to the attack class by triage.

3.6. Separating attacks from benign anomalies

Residual signatures discriminated attacks from benign events far better than raw signals (Fig. 6a). Among anomalous samples flagged by the detector, the triage classifier labelled none of the 3777 samples from instrument faults and algal blooms as attacks, and only 13 % of flagged storm samples. Every attack class except A4 had more than 70 % of its flagged samples labelled correctly. The raw tag classifier achieved comparable rejection of benign events only by labelling far fewer attack samples as attacks, which is why its pairing with the autoencoder lost recall.

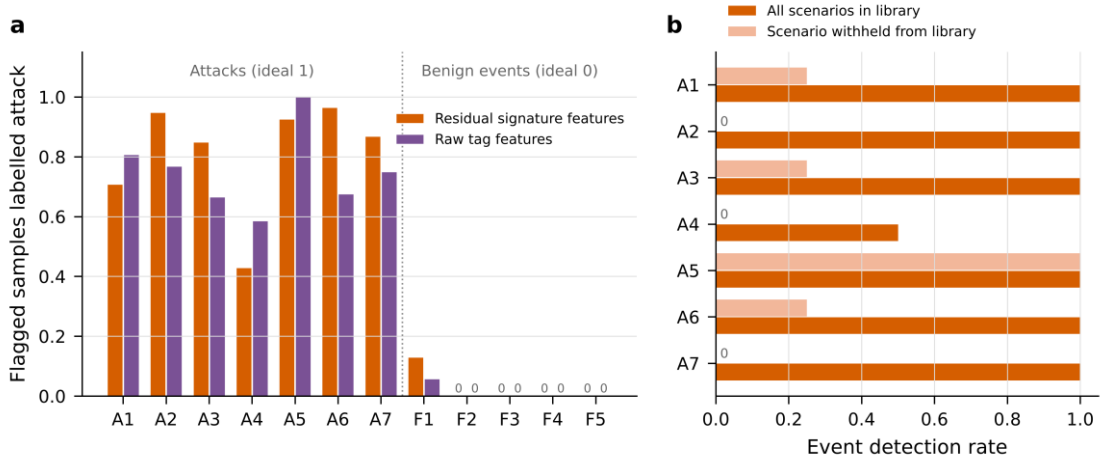


Fig. 6: Triage of detector alarms. (a) Fraction of flagged samples labelled as attacks by classifiers using residual signature features or raw tag features, by event class (n.f., not flagged). (b) Event detection rate of CIML with triage when all scenarios are in the twin library and when the scenario in question is withheld.

The leave one scenario out test exposed the main weakness of supervised triage (Fig. 6b). When an attack type was withheld from the library, detection of that type fell from 100 % to 25 % for A1, A3 and A6, from 100 % to zero for A2 and A7, and from 50 % to zero for A4. Only the analyser bias attack (A5) was still recognised, because its signature, a

kinetic residual that grows while the redox residual stays modest, resembles no benign event. The result for A2 is particularly telling. With A2 removed, the classifier had learned that a large redox residual with an otherwise quiet plant usually means a fouled electrode, which is exactly how a stealthy caustic attack presents. Supervised triage therefore generalises to new instances and magnitudes of known attack types but not reliably to new types, and the library must be broad.

3.7. Robustness to noise and to informed adversaries

Raising sensor noise eroded every method, but it did not change their ranking (Fig. 7a). The F1 score of the chemistry informed detector with triage was 0.74, 0.81 and 0.76 at noise multipliers of 0.5, 1 and 1.5, then fell to 0.67 at twice and 0.56 at three times the nominal noise. The best reference method at each level reached only 0.39 to 0.44. The slightly lower score at half the nominal noise than at nominal noise is within the variation expected between independently simulated records containing only 14 attack events each, and it is a useful reminder that single test records give F1 scores with an uncertainty of several hundredths. The reference detectors fluctuated without a consistent trend, which suggests that their performance depends more on how storms and faults happen to fall relative to their thresholds than on noise as such. At three times the nominal noise, residuals built on ORP and on network chlorine become noisy enough that small deceptions disappear, and the decline of the chemistry informed detector is driven mainly by later detection of the analyser bias and underdose scenarios.

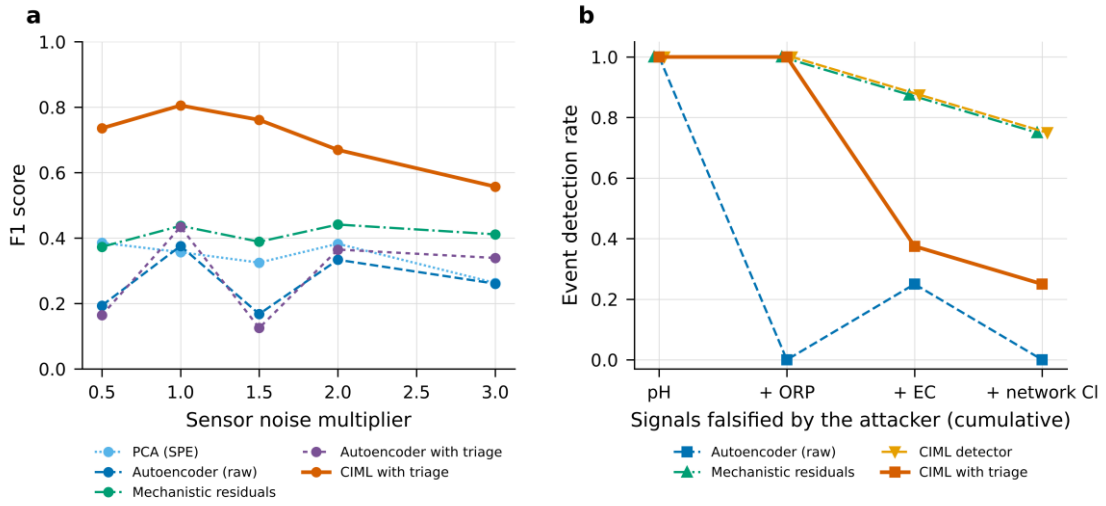


Fig. 7. Robustness. (a) F1 score against sensor noise multiplier for five independently simulated 110 day records. (b) Event detection rate for eight stealthy caustic overdoses (scenario A2) as the attacker replays progressively more signals.

The informed adversary experiment gave a more nuanced picture than we had expected (Fig. 7b). When only the finished pH was replayed, every detector caught all eight stealthy caustic attacks, because the true pH of about 9.5 depressed the ORP by more than 150 mV, a change that even the raw tag autoencoder recognised. Once the attacker replayed the ORP as well, the autoencoder detected none of the eight events. The chemistry informed detector and the mechanistic residual detector still caught all of them, and they continued to catch 7 of eight when conductivity was also replayed and 6 of eight when the chlorine readings at all three network nodes were added. The reason is that a replayed value is consistent with the plant as it was 24 h earlier, not as it is now. Temperature, chlorine setpoint, raw water conductivity and organic load all change from one day to the next, and the honest signals that report them contradict the replayed ones

through the redox, conductivity and kinetic balances. Triage, however, labelled most of these multisignal replays as benign, and detection after triage fell to 3 and 2 of eight. The replay signatures had no counterpart in the library, which is the same failure seen in the leave one scenario out test. An attacker who fabricated every linked signal in real time from a process model, rather than replaying recorded values, was not simulated; such an attacker would leave no chemical contradiction and represents the limit of what any consistency based detector can achieve (Folorunso et al., 2025).

4. 0 Discussion

4.1. Why chemistry helps

The results confirm an easy argument. In a treatment plant, the envelope of normal behaviour is broad – the raw water is never the same twice; a data driven detector learns the envelope, and the broader it is, the more active the detector will be. Internal attacks are not visible and natural excursions outside are visible as attacks. Chemical balances are not the same type of balances. They are independent of the source water, since they refer to what happens to any water in a specific unit, and they join signals which are measured by independent instruments. A storm alters almost all the raw signal but leaves almost all the others signals at normal levels, a masked underdose alters almost no raw signal other than the normal range but alters the redox and network balance. The detector then becomes more sensitive when data driven monitoring is no longer effective, and more specific when it has noise.The detector then becomes more sensitive at the points where data driven monitoring is ineffective, and more specific at the points where it is noisy, as shown in Figs. 2 to 5.

The same argument has been made for hydraulic consistency in distribution networks (Housh and Ohar, 2018), and treatment chemistry extends it to the stages of a utility where the most dangerous manipulations, of coagulant, caustic and disinfectant, take place.

The residuals also make alarms explainable. When the detector fires, the operator sees which balance failed and by how much, for example that the ORP is 30 mV below the value consistent with the reported chlorine and pH. That statement points to a short list of explanations, a fouled electrode, a falsified chlorine reading or a falsified pH reading, each of which can be checked with a grab sample and a handheld meter in minutes. Adoption is at stake as such an explanation has been found absent in previously surveyed data driven detectors (Giraldo et al., 2018; Addeen et al., 2021), and operators who don't understand an alarm learn to ignore it.

Another benefit is that special data is not required for calibration. The residuals in Table 2 were obtained during 56 days of normal operation, and were selected from the data for which the normal setpoint was periodically changed. Therefore, for utilities with historians, the residual layer can be constructed from existing data and the calibrated coefficients serve as a health check for the dosing equipment and instruments: If the apparent alkalinity consumption of the coagulant or the electrode slope changes, it indicates a process problem or a problem with the electrode or coagulant that should be addressed as well

4.2. Limits of the approach

Three restrictions need to be highlighted. The first is the basic one. A falsified detection of any kind of consistency based detector is possible when an attacker is able to manipulate all the signals connected by a balance and maintain the consistency among all of them (Urbina et al., 2016). The informed adversary experiment demonstrates that this is not the case for the simplest and most common deception, which is replay, as values recorded yesterday do not match with the signals that report the temperature, setpoints and raw water that are honest today. It's possible for a more sophisticated attacker to calculate values in real time from a process model. Chemistry increases this cost significantly as the attacker has to recreate

carbonate equilibria, electrode behaviour and conductivity to a satisfactory level for a number of instruments, possibly from various manufacturers, and configured to various controllers. But this doesn't make it impossible. The practical solution is to introduce some witnesses which cannot be accessed from the control network: a pH or conductivity sensor in the distribution network reporting via a telemetry path which is not connected to the control network, or an enclosed sensor which operates via a data logger that is not accessible via the control network. The residual framework would describe which other witness would fill the biggest gap, which would be a formal design question and could be solved as an optimal sensor placement problem.

The second constraint is related to triage. As demonstrated by the leave one scenario out results, a supervised classifier can recognize the attack types very similar to those in the library and can misidentify new attack types as familiar fault types. Its job ought to therefore be to prioritize rather than to let alarms go unheard. Benign anomaly models can be transferred better via open set and one class formulations, which explicitly model benign anomalies, rather than attacks. Additionally, digital twin-generated libraries can be systematically enhanced by sampling attack strategies, instead of building attack scenarios by hand.

The third restriction is the evidence base. All the results are from a simulated plant. The simulator incorporates typical chemistry, realistic noise and controller behaviour, natural storms, set-point changes and instrument failures, but it also included many other complications that are not modeled in the simulator, but are found in real plants: analyser lag, maintenance gaps, manual dosing changes that are not logged, multiple filter trains with staggered backwashes, mixed disinfectant regimes including chloramination, and source waters whose organic matter does not correlate as well with UV absorbance as assumed. Some of these will cause the residual distributions to become broader and less sensitive. The balances, however, are not simulator artefacts, but fundamental relations as will be found in any plant; and the calibration step is intended to include plant-specific coefficients. The next step is to validate historian records from operating plants, preferably combined with a red team exercise in which the utility staff makes a setpoint change (in a benign manner) and simulates attacks on a test train. This released code is able to take an export of a historian with the tag names indicated in Fig. 1, and executes the entire pipeline without modification.

4.3. Implications for utilities

Much of the time, small municipal companies and those in fast-growing cities in Africa and South Asia have limited instrumentation and security staff. For them the framework has a practical appeal: it uses analysers that are already standard for regulatory compliance (pH, turbidity, conductivity, free chlorine) together with ORP and UV254 probes that are inexpensive by comparison, and it requires no change to control logic. If the values of alkalinity are not available online, then daily laboratory data might be used for the carbonate balances (with lower time resolution). The same logic applies in wastewater treatment, where it is possible to enforce balances on such parameters as ammonium, nitrate, alkalinity, oxygen uptake and redox, for nitrification, denitrification and chemical phosphorus removal that are as rich as the ones used here, and where it is possible to cause permit violations or even the failure of the process by manipulation of aeration or chemical phosphorus removal

Finally, chemistry informed monitoring is a complement to, not a substitute for, conventional cyber hygiene. The Oldsmar intrusion succeeded because of shared passwords and exposed remote access software (CISA, 2021). Network segmentation, authentication and patching remain the first line of defence. Process level detection of the kind developed here is the last

line, the one that still works when the others have failed and the attacker is already moving dosing pumps.

5.0 Conclusions

Treatment chemistry supplies independent witnesses that can test whether SCADA data are telling the truth. Recast as calibrated residuals, carbonate equilibrium, ionic conductivity, redox potential, chlorine kinetics and mass balance were insensitive to storms yet broke in distinct patterns under attack. A detector built on them, with a triage stage trained on a digital twin, outperformed widely used data driven monitors in precision, F1 score and false alarm rate, and it explained each alarm in chemical terms. Its calibration needs only routine historian data. Three limits stand out: attacks that leave the chemistry consistent, attack types absent from the triage library, and the reliance of this study on simulated data. Validation on operating plant records, open set triage and the placement of tamper resistant witness sensors are the priorities for further work. The same reasoning applies directly to wastewater treatment, where nutrient and redox balances are equally rich.

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

Adeyemi, A.B., Onyenachi, J., Salone, A. 2025. Integrating machine learning with environmental chemistry to forecast pollutant releases in coatings production. World Journal of Advanced Research and Reviews, 2025, 27(03), 1061-1072 https://doi.org/10.30574/wjarr.2025.27.3.3153

Adeyemi, A.B., Adebara, M. R., Chukwuma, O.J., Jackas, O. 2025. Coupling environmental chemistry with Artificial Intelligence for coastal hazard risk assessment under climate change. International Journal of Science and Research Archive, 2025, 16(03), 498-507 https://doi.org/10.30574/ijsra2025.16.3.2572

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). ACM, Pittsburgh, pp. 25–28. https://doi.org/10.1145/3055366.3055375

Benjamin, M.M., 2015. Water Chemistry, 2nd ed. Waveland Press, Long Grove, IL.

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). ACM, Hong Kong, pp. 355–366. https://doi.org/10.1145/1966913.1966959

Chandran, P., Sunil, K. S., 2025. Safeguarding water distribution systems: Cyber-physical attack detection using rectilinear hybrid belief classifier network. Journal of Water Process Engineering, 76, Article 108131. https://doi.org/10.1016/j.jwpe.2025.108131

CISA (Cybersecurity and Infrastructure Security Agency), 2021. Compromise of U.S. Water Treatment Facility. Joint Cybersecurity Advisory AA21-042A, 11 February 2021. https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-042

Copeland, A., Lytle, D.A., 2014. Measuring the oxidation–reduction potential of important oxidants in drinking water. Journal AWWA 106 (1), E10–E20. https://doi.org/10.5942/jawwa.2014.106.0002

Crittenden, J.C., Trussell, R.R., Hand, D.W., Howe, K.J., Tchobanoglous, G., 2012. MWH's Water Treatment: Principles and Design, 3rd ed. John Wiley & Sons, Hoboken, NJ. https://doi.org/10.1002/9781118131473

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

Folorunso, T. N., Idowu, K., & Balogun, A. (2025). AI-driven financial risk forecasting: A multi-model ensemble approach for enterprise decision intelligence. *Journal of Computational Analysis and Applications, 34*(3). https://eudoxuspress.com/index.php/pub/article/view/5080

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., Cardenas, 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), 76. https://doi.org/10.1145/3203245

Goh, J., Adepu, S., Junejo, K.N., Mathur, A., 2017. A dataset to support research in the design of secure water treatment systems, in: Havarneanu, G., Setola, R., Nassopoulos, H., Wolthusen, S. (Eds.), Critical Information Infrastructures Security. CRITIS 2016. Lecture Notes in Computer Science, vol. 10242. Springer, Cham, pp. 88–99. https://doi.org/10.1007/-3-319-71368-7_8

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), 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., Cherotich, V., Aniebonam, E. E., Odozor, L., & During, A. (2025). Integrating AI and machine learning into cyber risk management for critical utility systems. International Journal of Artificial Intelligence Engineering and Technology. https://doi.org/10.54660/IJAIET.2025.6.2.32-48

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). IEEE, New Orleans, pp. 1058–1065. https://doi.org/10.1109/ICDMW.2017.149

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics 3, 422–440. https://doi.org/10.1038/s42254-021-00314-5

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering 29 (10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

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). ACM, Toronto, pp. 72–83. https://doi.org/10.1145/3264888.3264896

Langelier, W.F., 1936. The analytical control of anti-corrosion water treatment. Journal AWWA 28 (10), 1500–1521. https://doi.org/10.1002/j.1551-8833.1936.tb13 785.x

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

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). IEEE, Vienna, pp. 31–36. https://doi.org/10.1109/CySWater.2016.7469060

Morris, J.C., 1966. The acid ionization constant of HOCl from 5 to 35°. The Journal of Physical Chemistry 70 (12), 3798–3805. https://doi.org/10.1021/j100884a007

Odozor, L., Berko, S. N., Obu, A. U., & Idowu, K. (2025). Building a resilient U.S. cybersecurity workforce through practical training. *International Journal of Scientific and Management Research, 8*(8), 30–36.

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.

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

Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Read, J.S., Jia, X., Willard, J., Appling, A.P., Zwart, J.A., Oliver, S.K., Karpatne, A., Hansen, G.J.A., Hanson, P.C., Watkins, W., Steinbach, M., Kumar, V., 2019. Process-guided deep learning predictions of lake water temperature. Water Resources Research 55 (11), 9173–9190. https://doi.org/10.1029/2019WR024922

Rossman, L.A., Clark, R.M., Grayman, W.M., 1994. Modeling chlorine residuals in drinking-water distribution systems. Journal of Environmental Engineering 120 (4), 803–820. https://doi.org/10.1061/(ASCE)0733-372(1994)120:4(803)

Sayghe, A. (2025). Digital twin-driven intrusion detection for industrial SCADA: A cyber-physical case study. Sensors, 25(16), Article 4963. https://doi.org/10.3390/s25164963

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), e0118432. https://doi.org/10.1371/journal.pone.0118432

Stumm, W., Morgan, J.J., 1996. Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters, 3rd ed. John Wiley & Sons, New York.

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), 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), 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 Jr., E., 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., Biswas, P., Pasha, M.F.K., Kc, B., Somasundaram, S.L., Housh, M., Ohar, Z., 2018. Battle of the attack detection algorithms: disclosing cyber attacks on water distribution networks. Journal of Water Resources Planning and Management 144 (8), 04018048. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000969

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

Urbina, D.I., Giraldo, J.A., Cardenas, 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). ACM, Vienna, pp. 1092–1105. https://doi.org/10.1145/2976749.2978388

Vasconcelos, J.J., Rossman, L.A., Grayman, W.M., Boulos, P.F., Clark, R.M., 1997. Kinetics of chlorine decay. Journal AWWA 89 (7), 54–65. https://doi.org/10.1002/j.1551-8833.1997.tb08259.x

WHO (World Health Organization), 2022. Guidelines for Drinking-water Quality: Fourth Edition Incorporating the First and Second Addenda. World Health Organization, Geneva. ISBN 978-92-4-004506-4.

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V., 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Computing Surveys 55 (4), 66. https://doi.org/10.1145/3514228

Declarations

Ethical Approval and Consent to Participate

This study did not involve human participants, animals, or human biological materials. Therefore, ethical approval and informed consent were not required.

Consent for Publication

Not applicable.

Funding

The authors received no financial support from any funding agency in the public, commercial, or not-for-profit sectors for the conduct of this research.

Data Availability Statement

The data supporting the findings of this study are included within the article. Additional information may be made available by the corresponding author upon reasonable request.

Authors' Contributions

The manuscript was written by the author/