

An online drift-aware digital twin for regenerative wastewater systems: Conformal water-quality risk and rainfall-forced early warning from live sensor telemetry

C. Manjunath¹, Pitta Shankaraiah², G. Malli Reddy³, M. Mallikarjuna Rao^{4*},
D. Venkatesh⁵, S. R. Basha⁶, Goturu Sreenivasan⁷, Gutha Viswanath⁸

¹ Department of Computer Applications, PVKK Institute of Technology (Autonomous), Anantapuramu, Andhra Pradesh, India

² Department of Mathematics, Dayananda Sagar University, Bengaluru, Karnataka, India

³ Department of Computer Science and Engineering, PVKK Institute of Technology (Autonomous), Anantapuramu, Andhra Pradesh, India

⁴Department of Computer Science and Design, PVKK Institute of Technology (Autonomous), Anantapuramu, Andhra Pradesh, India

⁵ Department of Computer Science and Engineering, Ananthalakshmi Institute of Technology, Anantapuramu, Andhra Pradesh, India

⁶ Department of Computer Science and Engineering, St. Peter's Engineering College, Hyderabad, Telangana, India

⁷ Department of Electrical and Electronics Engineering, PVKK Institute of Technology (Autonomous), Anantapuramu, Andhra Pradesh, India,

⁸ Department of Electronics and Communication Engineering, PVKK Institute of Technology (Autonomous), Anantapuramu, Andhra Pradesh, India

* Corresponding author's e-mail: malkari.mkrao@gmail.com

ABSTRACT

Digital twins for wastewater systems are normally fitted once and then frozen. That is a reasonable simplification while process conditions, sensor behaviour and weather patterns remain close to the training period, but it can fail silently when they move. This paper reports RT-ReGenTwin, an online digital twin that learns continuously from public sensor telemetry and reports threshold-exceedance risk as a calibrated probability rather than as a point forecast. Fifteen-minute water-quality records covering 2026-03-13T05:00 to 2026-09-09T05:00 were obtained from two continuous monitoring stations of the United States Geological Survey on effluent-influenced reaches downstream of large municipal treatment works, with matched hourly rainfall from Open-Meteo. Two further stations were probed and not used, one because it reports no continuous water-quality determinand and one because it had no current record. Turbidity, dissolved oxygen and specific conductance were used as receiving-water surrogates for solids loading, oxygen-demand effects and effluent fraction. The twin couples an adaptive random-forest regressor with drift detectors and an adaptive conformal layer. Across four station–determinand combinations, empirical coverage was closer to the nominal 0.90 level for the online calibrator than for the frozen reference in all four cases. Only one of the four combinations recorded any threshold exceedance during the window, so the probabilistic scores and the dispatch analysis rest on that stream alone; for it, coverage was 0.898 versus 0.595 and the Brier score 0.0472 versus 0.0345. Risk-triggered regenerative-capacity dispatch avoided 337 of 590 threshold-exceedance hours at a duty cycle of 0.163, compared with 414 avoided under continuous operation. The results support calibrated risk reporting for streaming environmental monitoring, while the receiving-water surrogates and dispatch simulation should not be interpreted as direct permit-compliance measurements or a treatment experiment.

Keywords: online digital twin, concept drift, conformal prediction, water-quality risk, real-time monitoring, nature-based solutions, wastewater resilience.

INTRODUCTION

Frozen models and moving plants

A treatment works is not a stationary process. Sewer catchments gain households, industrial dischargers change product lines, sludge inventories are rebuilt after a washout, probes foul and are cleaned, and rainfall arrives in a pattern that no longer matches the one a model was calibrated against. Mechanistic descriptions of the activated sludge process handle much of this through their kinetic structure, but only after a calibration exercise that presumes stable influent characterisation and high-resolution measurement (Gernaey et al., 2014; Henze et al., 2000). Data-driven surrogates are cheaper to build and often more accurate over short horizons, and they have become the standard companion to process models in the water sector (Corominas et al., 2018; Newhart et al., 2019).

What almost all of them share is a training protocol borrowed from static supervised learning. A historical record is split, a model is fitted on one part, accuracy is reported on the other, and the fitted model is then treated as though it will remain valid. Our own earlier offline twin followed exactly this pattern, and reported that random forest regression tracked the biochemical determinands well while suspended solids resisted point prediction altogether (Rao et al., 2026). The multi-site extension of that work broadened the design to several plants and added economic and climate-resilience layers, which improved transferability across space (Archana et al., 2026). Neither version updates itself once deployed, and neither carries a statement about how much of its accuracy survives a shift in operating regime.

This matters more for threshold-based environmental monitoring than for average performance reporting. Environmental limits are thresholds. A model that is unbiased on average but poorly calibrated in the upper tail will systematically misjudge the frequency of exceedance, and a model that has silently drifted will do so without any external signal that it has stopped being trustworthy.

Concept drift in environmental telemetry

Streaming machine learning has developed a mature vocabulary for this problem. A change in the joint distribution of predictors and target over time is termed concept drift, and the literature

distinguishes abrupt, gradual, incremental and recurring forms of it (Gama et al., 2014; Lu et al., 2019). Detection methods range from the classical cumulative sum procedure (Page, 1954) to adaptive windowing schemes that maintain a variable-length reference window and split it when the two halves become statistically distinguishable (Bifet and Gavaldà, 2007). Adaptive ensembles combine detection with repair: the adaptive random forest grows background trees when a warning fires and swaps them in when the warning is confirmed, so the ensemble replaces the affected part of its structure rather than being retrained from scratch (Gomes et al., 2017; Gomes et al., 2018).

Water-quality telemetry is a natural candidate for these methods and has been slow to adopt them. Sonde records combine seasonal cycles, storm-driven transients, calibration steps introduced by routine servicing, and occasional sensor faults (Wagner et al., 2006). All four look like drift to a learner, and only some of them should trigger a structural update. A twin that never adapts inherits every one of these changes as unexplained error; a twin that adapts to all of them will chase sensor artefacts.

From accuracy to calibrated risk

The second gap is what the twin reports. Point forecasts and their error statistics say nothing directly about the quantity an operator needs, which is the probability that a limit will be breached in the next few hours. Conformal prediction supplies that missing piece. Given any underlying regressor, it converts residuals from a calibration set into prediction sets with finite-sample coverage guarantees under exchangeability, without distributional assumptions (Lei et al., 2018; Shafer and Vovk, 2008; Vovk et al., 2005). Conformal predictive systems go further and return a full predictive distribution, from which an exceedance probability follows directly (Vovk et al., 2019).

The exchangeability requirement is precisely what a drifting stream violates. Two lines of work address this. Adaptive conformal inference treats the nominal miscoverage level as a control variable and updates it online from realised coverage errors, which restores long-run coverage under arbitrary distribution shift (Gibbs and Candès, 2021). Complementary theory bounds the coverage gap in terms of how far the data depart from exchangeability (Barber et al., 2023). The combination suits a digital twin, because it turns

the very drift that degrades point accuracy into a signal that widens the reported uncertainty rather than an unannounced failure.

Regenerative capacity as a dispatchable asset

Nature-based solutions such as constructed wetlands, biofiltration beds and riparian buffers attenuate solids and oxygen demand and add ecosystem value (Kadlec and Wallace, 2009; Oral et al., 2020; Vymazal, 2014). Assessment frameworks generally treat their performance as a fixed removal efficiency applied uniformly across the record, which is the assumption we ourselves used in the offline twin (Rao et al., 2026). Under that assumption the sensitivity analysis reduces to a single curve of threshold-exceedance benefit against efficiency, and the practical conclusion was that returns saturate near sixty to seventy per cent.

That framing hides an operational question. Much regenerative capacity is at least partly dispatchable through flow splitting, pump scheduling or bypass control, and hydraulic retention in a wetland cell is finite. If threshold-exceedance risk is available as a calibrated probability, the interesting design variable is not only how efficient the unit is but how often it needs to be engaged. Stormwater practice has moved in this direction for detention control (Kerkez et al., 2016); wastewater operational planning has not.

Research gaps

Three gaps follow from the discussion above.

- Wastewater digital twins are validated with static train and test splits and are not evaluated prequentially, so their reported accuracy does not describe what a deployed model would deliver as conditions move away from the training period.
- Threshold-exceedance risk is usually reported as a deterministic comparison of a point forecast against a limit. This study instead attaches empirical coverage assessment to the predictive interval and scores the exceedance probability with a proper scoring rule.
- Regenerative capacity is evaluated as a fixed efficiency envelope rather than as a dispatchable asset, so the analysis cannot say what fraction of the threshold-exceedance benefit survives intermittent operation.

Objectives and contributions

This study addresses those three gaps with a framework we call RT-ReGenTwin. The contributions are as follows.

- An online digital twin driven by public high-frequency telemetry rather than an archived plant dataset. Fifteen-minute records are merged with matched hourly rainfall and re-played in temporal order, and the learner is updated observation by observation under a strict test-then-train protocol.
- An adaptive conformal predictive layer that converts each point forecast into a calibrated exceedance probability. The nominal miscoverage level is updated online, so the calibration target is adjusted rather than left fixed under drift.
- A risk-triggered dispatch rule for nature-based capacity, evaluated on the trade-off between threshold exceedances avoided and duty cycle. This generalises the fixed efficiency envelope used previously (Rao et al., 2026) and turns the saturation result reported there into an operating policy.
- A rainfall-forced early-warning analysis that reports lead time as a function of the false-alarm budget, so a plant can choose an operating point rather than accept a single alarm threshold.
- A descriptive comparison against a frozen random forest built on identical features, matching the offline and multi-site twin designs (Archana et al., 2026; Rao et al., 2026), together with a single-tree streaming ablation and a persistence baseline.

MATERIALS AND METHODS

Framework overview

RT-ReGenTwin has four layers. An acquisition layer pulls telemetry and meteorological forcing from public services. A feature layer builds a strictly causal design matrix at hourly resolution. A learning layer maintains an adaptive ensemble and a conformal calibrator that are updated after every observation. A decision layer converts the calibrated predictive distribution into an exceedance probability, an early-warning alarm and a dispatch signal for regenerative capacity. Nothing in the pipeline uses information

that would be unavailable at the moment a decision has to be taken.

Data acquisition

Water-quality telemetry was obtained from the Instantaneous Values service of the National Water Information System operated by the United States Geological Survey (U.S. Geological Survey, 2026). The service publishes sonde readings at a nominal fifteen-minute interval with a latency of roughly one hour, which makes it one of the few openly licensed water-quality feeds of sufficient length and resolution for streaming experiments. Station operation, sensor servicing and record computation follow published national procedures (Wagner et al., 2006), so the record carries documented quality-control conventions rather than undocumented local practice.

Meteorological forcing was obtained from the Open-Meteo service (Zippenfenig, 2023), which exposes hourly precipitation without registration and blends reanalysis for older periods (Hersbach et al., 2020) with near-real-time fields for the last few days. The pipeline requested the reanalysis-based archive for the whole window and the forecast service's past-day fields for the final 92 days, and merged them to one value per hour. The merged series lacks 309 of its 4,344 hours (7.1%), all between 9 June and 6 July 2026 and at identical hours at all three stations, which points to the merge rather than to the precipitation record. The first three hours of each gap are interpolated linearly and later hours are forward-filled for at most six hours, so between 261 and 302 modelled hours per series carry filled rainfall and six hours per series that neither rule bridges are dropped. Station coordinates for the rainfall query were read from the site service rather than entered by hand, so the forcing is co-located with the sonde by construction.

The pulled record was replayed in temporal order under a prequential protocol. This is the standard evaluation design in the streaming learning literature (Bifet et al., 2010; Gama et al., 2014) and is equivalent to live operation from the learner's point of view, since at every step the model has seen only observations that precede the one it is asked to predict. Archiving the pull also makes the reported run reproducible after the fact, which a purely live evaluation would not be.

Before any modelling, the pipeline probes each candidate station to establish which

determinands it is currently reporting and how complete the record is. Stations whose sondes were offline or whose record was shorter than the warm-up requirement were dropped. This inventory step is part of the analysis pipeline and is reported in Table 1, so the study design is verifiable against the state of the monitoring network at the time of the run rather than assumed.

Study reaches and water-quality surrogates

Effluent quality itself is not published as a public high-frequency stream anywhere at the resolution this study needs; regulatory reporting is monthly and aggregated. The design therefore moves the observation point downstream and treats the receiving water immediately below major outfalls as the screening surface, not as a permit-compliance location. Three determinands were used as receiving-water surrogates. Turbidity stands for solids loading, a substitution supported by an established national procedure for computing suspended-sediment series from in-stream turbidity and discharge (Rasmussen et al., 2009). Dissolved oxygen stands for exerted oxygen demand and is itself the subject of a widely applied ambient criterion (Helsel et al., 2020). Specific conductance stands for effluent fraction, since municipal effluent is typically more conductive than the ambient flow it joins.

This is a deliberate trade. The study gains an independently operated and publicly auditable data stream, and loses direct correspondence to a permit parameter measured at the outfall. The thresholds were configured as a dissolved oxygen minimum of 5.0 mg L⁻¹ following the ambient aquatic-life criterion (U.S. Environmental Protection Agency, 1986), a turbidity screening limit of 25 FNU and a specific-conductance screening limit of 1200 µS cm⁻¹. These are screening thresholds, not permit limits, and can be replaced by site-specific permit or regulatory values without changing the rest of the pipeline.

Feature construction

The record was aggregated to hourly means, which suppresses the highest-frequency sonde noise while preserving storm response. For each determinand the design matrix carries lags at 1, 2, 3, 6, 12 and 24 hours, six-hour and twenty-four-hour rolling means computed from lagged values only, and a first difference. Rainfall enters as the

hourly depth, as three-, six-, twenty-four- and seventy-two-hour accumulations, and as a recursive antecedent precipitation index

$$API_t = k \cdot API_{t-1} + P_t, k = 0.92 \quad (1)$$

which summarises catchment wetness with a single state variable. Diurnal and seasonal position are encoded as sine and cosine pairs so that the tree ensemble sees them as continuous rather than as an arbitrary integer ordering. The concurrent reading of the target determinand is excluded from its own feature set. Remaining gaps are forward-filled for at most six hours, which reflects what a control room would actually have during a short telemetry outage.

Streaming learner and drift handling

The core regressor is an adaptive random forest for regression (Gomes et al., 2018), implemented in the river library (Montiel et al., 2021), with twenty-five Hoeffding trees, adaptive leaf prediction and a grace period of fifty instances. Each tree carries an adaptive windowing detector on its own error stream (Bifet and Gavaldà, 2007). When a warning fires the tree begins growing a background replacement; when the drift is confirmed the replacement is promoted. Adaptation is therefore local to the affected part of the ensemble.

Two ensemble-level detectors are maintained purely as diagnostics. One monitors the absolute residual, which flags periods the learner has not yet absorbed. The other monitors the target series, which flags regime shifts in the process itself. Reporting both matters, because a well-adapting ensemble can pass through a large regime change with an almost unchanged residual distribution, and a residual-only monitor would then show nothing.

Adaptive conformal threshold-exceedance risk

Let $\hat{y}_t$ be the forecast issued at time t before y_t is observed. The conformity score is the signed residual

$$s_t = y_t - \hat{y}_t \quad (2)$$

and the calibrator holds the most recent $w = 720$ scores, a thirty-day window at hourly resolution. Writing $\hat{F}_t$ for the empirical distribution function of that window, the predictive interval at effective level α_t is

$$C_t = [\hat{y}_t + \hat{F}_t^{-1}(\alpha_t/2), \hat{y}_t + \hat{F}_t^{-1}(1 - \alpha_t/2)] \quad (3)$$

and the shifted window is a conformal predictive distribution (Vovk et al., 2019), from which the threshold-exceedance risk follows directly. For an upper limit and for a lower limit respectively,

$$\pi_t = P(y_t > C_t^{LIM}) = 1 - \hat{F}_t(C_t^{LIM} - \hat{y}_t) \quad (4)$$

$$\pi_t = P(y_t < C_t^{LIM}) = \hat{F}_t(C_t^{LIM} - \hat{y}_t) \quad (5)$$

After the observation arrives, the effective level is updated by the adaptive conformal inference recursion (Gibbs and Candès, 2021),

$$\alpha_{t+1} = \alpha_t + \gamma (\alpha - err_t), err_t = 1 \{ y_t \notin C_t \} \quad (6)$$

with target level $\alpha = 0.10$ and step size $\gamma = 0.01$, and α_t clipped to $[0.001, 0.5]$. The recursion is a simple integral controller on realised coverage. When the residual distribution widens the recent miscoverage rate rises, α_t falls, and the interval widens until coverage is recovered. The adaptive recursion is intended to control long-run miscoverage under distribution shift; the coverage reported here is empirical and should not be interpreted as a finite-sample guarantee after abrupt changes (Barber et al., 2023; Gibbs and Candès, 2021).

Risk-triggered dispatch of regenerative capacity

The offline twin applied a constant attenuation η to every time step (Rao et al., 2026). Here the attenuation is gated by the calibrated risk. With an engagement indicator $g_t = 1 \{ \pi_t \geq \tau \}$ for a trigger τ , the regenerated series for an upper limit is

$$y_t^{reg} = y_t (1 - \eta g_t) \quad (7)$$

and for a lower limit, where regeneration lifts the determinand towards the standard rather than suppressing it,

$$y_t^{reg} = y_t + \eta g_t \max(0, C_t^{LIM} - y_t) \quad (8)$$

Two resilience measures are carried over from the earlier work so that the comparison is direct. Writing $\delta = +1$ for an upper limit and $\delta = -1$ for a lower one, threshold-exceedance frequency and the threshold-exceedance magnitude index are

$$V = \sum_t 1 \{ \delta (y_t - C_t^{LIM}) > 0 \},$$

$$VMI = \sum_t \max(0, \delta (y_t - C_t^{LIM})) \quad (9)$$

and the duty cycle, which the earlier framework had no way to express, is

$$D = (1/T) \sum_t g_t \quad (10)$$

The quantity of engineering interest is the ratio of threshold exceedances avoided to duty cycle, which measures how much threshold-exceedance benefit each hour of engaged regenerative capacity buys. Sweeping τ from 0.05 to 0.95 traces a frontier between the two, and the fixed envelope of the earlier design is the single point on that frontier where $D = 1$.

Early warning

An alarm is raised when the calibrated risk crosses a threshold. For each observed threshold-exceedance episode, defined as the first hour of a contiguous run of exceedance, the lead time is the number of hours between the earliest alarm within a twenty-four-hour look-back and the onset of the episode. The false-alarm rate is the proportion of hours without threshold exceedance carrying an alarm. Sweeping the threshold produces a lead-time against false-alarm curve, which is the form a plant needs in order to choose an operating point.

Baselines, ablation and evaluation

Three comparators were run on identical features. The first is a frozen random forest with two hundred trees and a maximum depth of ten, fitted on the first sixty per cent of the record and never updated, with static split-conformal intervals from its training residuals. This reproduces the design of the offline and multi-site twins (Archana et al., 2026; Rao et al., 2026) and is the main point of comparison. The second is persistence, the last observed value carried forward, which is the floor any twin must beat. The third is an ablation in which the online learner is a single Hoeffding tree; it still updates incrementally but never repairs its structure. Because it also differs from the adaptive forest in ensemble size, the contrast does not isolate the contribution of drift handling (see Results).

Point accuracy is reported as prequential root mean square error and coefficient of determination over the post-warm-up stream. Uncertainty quality is reported as prediction interval coverage probability against the nominal 0.90. Probabilistic threshold-exceedance forecasts are scored with the Brier score (Brier, 1950), which is proper (Gneiting and Raftery, 2007), together with a Brier skill score against the climatological exceedance rate. Because exceedance is rare, discrimination is reported as area under the precision-recall

curve (Saito and Rehmsmeier, 2015). A warm-up of fourteen days precedes all scoring.

Implementation and reproducibility

The pipeline is a single Python module using river for the streaming components (Montiel et al., 2021), scikit-learn for the frozen baseline (Pedregosa et al., 2011) and requests for the service calls. Random seeds are fixed. The script writes every reported number to a machine-readable results file. A value map deposited with the code links every reported value in this manuscript to its field in that file, and the reproduction script checks each one against a fresh replay. Package versions are pinned in the deposit; the frozen reference is reproduced exactly only with scikit-learn releases 1.6.1 to 1.7.2 (1.7.2 is pinned), because tree construction differs between releases. Because the input is a live service, a re-run at a later date will cover a different window and will not reproduce these numbers exactly; it will reproduce the analysis. The exact window is stated below and the raw pull is archived with the code.

RESULTS

Monitoring inventory and acquisition window

Telemetry was pulled for the period 2026-03-13T05:00 to 2026-09-09T05:00. Four stations were probed, in the order of the pipeline's candidate list. Two met the inclusion criteria and contributed results, giving four station-determinand combinations; the other two are listed as not retained. The first candidate, USGS 05536357 (Grand Calumet River at Hammond, Indiana), was passed over before any data were requested. Its probe was not logged, so this outcome is inferred from the candidate order and the inventory the run recorded; USGS holds no continuous record for the site after November 2020. A station is included only if it reports at least one continuous water-quality determinand among turbidity (USGS parameter code 63680), dissolved oxygen (00300) and specific conductance (00095). USGS 01646500, Potomac River near Washington DC, reports only discharge and gage height and therefore carries no determinand this study models. A determinand is modelled only if it has at least 2000 valid readings in the window and at least

836 usable hours after feature construction. Table 1 records the outcome for every station probed. The deposited acquisition audit trail records each decision with the requests and evidence behind it, and the station_screening block of the results file records the decisions for the three stations whose data were downloaded. The short codes assigned in Table 1 are used in Tables 2 to 4.

Predictive performance under a streaming protocol

Table 2 reports post-warm-up prequential accuracy for the online twin, the ablation and persistence, alongside the frozen random-forest reference. For turbidity at S2 (Chattahoochee River at Atlanta, GA), the online twin reached $R^2 = 0.557$ and $RMSE = 18.19$; the frozen reference reached $R^2 = 0.745$ and $RMSE = 13.97$ on its held-out final 40% of the record. Those two figures are not comparable as they stand: the online model is scored prequentially across the whole post-warm-up stream, while the frozen reference

is scored only on the final 40% it never saw. To compare them we restrict both to the rows they have in common. On that matched window, covering the 1,551 hours both models were scored on (the whole of the frozen reference's held-out period, so only the online figures change), the online twin gives $R^2 = 0.658$ against 0.745 for the frozen reference, while empirical coverage is 0.892 against 0.595. The matched values are reported for every combination in Table 2 and are the basis for any statement here about one design relative to the other; the full-stream figures are retained as diagnostics of each model under its own protocol. On point accuracy the online twin does not exceed the frozen reference in any of the four combinations, matched or unmatched.

The ensemble-level detectors raised 51 drift alarms over the window for the leading case. The ablation using a single Hoeffding tree reached $RMSE = 24.71$ against 18.19 for the adaptive forest. That contrast varies ensemble size as well as drift handling, so it does not isolate the

Table 1. Stations probed, with the inclusion decision for each. Record lengths are given per modelled determinand

Code	Station	USGS number	Latitude, longitude	Determinands reported	Usable record (h), per modelled determinand	Decision
S1	Sacramento R A Freeport Ca	11447650	38.4557, -121.5016	water_temp_c, discharge_cfs, gage_height_ft, spc_uscm, do_mgl, do_pct_sat, ph	4,289 (do_mgl); 4,289 (spc_uscm)	Retained
S2	Chattahoochee River at Atlanta, Ga	02336000	33.8592, -84.4544	water_temp_c, discharge_cfs, gage_height_ft, spc_uscm, turbidity_fnu	3,876 (turbidity_fnu); 3,889 (spc_uscm)	Retained
—	Potomac River Near Wash, Dc Little Falls Pump Sta	01646500	38.9498, -77.1276	discharge_cfs, gage_height_ft	—	Excluded — no water-quality determinand
—	Grand Calumet River at Hohman Av at Hammond, In	05536357	—	not recorded (probe not logged)	—	Not retained — no current record at probe (inferred)

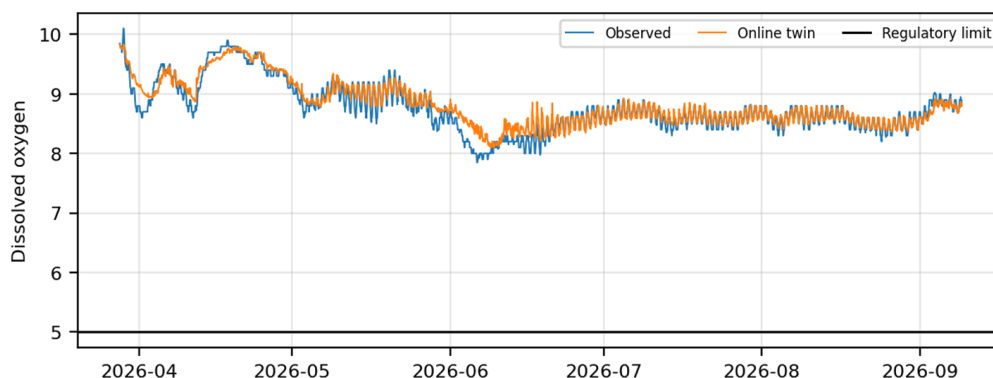


Figure 1. Observed and one-step-ahead predicted dissolved oxygen at S1 (Sacramento River at Freeport, CA), with the 5.0 mg L⁻¹ screening threshold marked

Table 2. Predictive performance and streaming diagnostics

Station and determinand	n (h)	Online R ²	Online RMSE	Ablation R ²	Persistence R ²	Frozen RF R ²	Frozen RMSE	Drift alarms	Matched R ² online	Matched R ² frozen
S1 — Dissolved oxygen	4,289	0.883	0.14	-0.143	0.991	0.950	0.03	18	0.790	0.950
S1 — Specific conductance	4,289	0.896	6.22	-0.083	0.995	0.995	0.90	53	0.893	0.995
S2 — Turbidity	3,876	0.557	18.19	0.183	0.934	0.745	13.97	51	0.658	0.745
S2 — Specific conductance	3,889	0.881	4.35	0.023	0.988	0.951	2.60	61	0.876	0.951

contribution of drift repair, and we do not attribute the difference to it. A matched-capacity ablation — twenty-five bagged Hoeffding trees without drift detection — is provided in the deposited code, as part of the alternative configuration described below, for readers who wish to separate the two. Persistence gave $R^2 = 0.934$ for the leading case, showing that short-horizon temporal persistence is a strong benchmark and that the learned model does not automatically add point-forecast skill.

Calibration of the threshold-exceedance risk

Coverage is where the two designs separate most clearly. Against a nominal 0.90, the adaptive conformal layer held empirical coverage at 0.898 for the leading case, while the static split-conformal intervals of the frozen reference gave 0.595. Across all four combinations, adaptive calibration was closer to nominal coverage. Interval width must be read alongside coverage. For the two S1 streams the mean width is dominated by a small number of extreme intervals — 1.3×10^7 mg L⁻¹ for dissolved oxygen and 2.2×10^8 μ S cm⁻¹ for specific conductance — and is not a usable summary. The median width is 0.36 mg L⁻¹ against an interquartile range of 0.47 mg L⁻¹ for dissolved oxygen, and 13.9 against 29.5 μ S cm⁻¹ for conductance, so over the great majority of the record the intervals are narrower than the spread of the determinand itself. Intervals wider than ten times the interquartile range occur in 0.28% and 0.33% of hours respectively; excluding them moves empirical coverage from 0.9008 to 0.9006 and from 0.8978 to 0.8975, so the coverage results are not an artefact of those few intervals. The extreme values arise because the adaptive random forest fits linear models in its leaves, which on unstandardised inputs can extrapolate far outside

the physical range of the target during the opening of the stream and after a regime change; such a residual then remains in the 720-hour conformity window until it ages out. Table 3 reports the median width as a multiple of each determinand's interquartile range, which is the appropriate summary for this diagnostic, alongside the mean for comparability. An alternative configuration that removes the extreme intervals — standardised inputs, mean leaf prediction, a calibration burn-in and a forecast support clip — is included in the deposited code; it stabilises width at some cost in point accuracy, and the configuration reported here remains the primary one.

Figure 2 shows rolling coverage against the nominal level for dissolved oxygen at S1, the stream of Figure 1. Its mean rolling coverage was 0.901 and it sat within three percentage points of nominal for 51.2% of the record, with a median excursion lasting 97.0 hours across 15 excursions. For the leading S2 turbidity stream the corresponding values were 0.902, 48.5%, 50.0 hours and 12 excursions. For that stream, the online exceedance forecast had a Brier score of 0.0472 versus 0.0345 for the frozen reference and a Brier skill score of 0.660. The Brier score did not favour the online design, so the value of the online layer is not claimed to be superior on every probabilistic metric. Discrimination, measured as area under the precision-recall curve on an exceedance base rate of 0.167, was 0.896.

Three of the four combinations recorded no threshold exceedance at all during the window. Dissolved oxygen at S1 never fell below the 5.0 mg L⁻¹ criterion, and specific conductance never exceeded 1200 μ S cm⁻¹ at either station; the exceedance-hour counts are 0, 0 and 0, against 590 for turbidity at S2. Three statistics degenerate when the outcome vector contains no positive case. The Brier score is then evaluated against a vector of zeros, so a value near zero records

the absence of events rather than forecast skill; the Brier skill score has no finite climatological reference, because the reference forecast has zero variance; and the precision-recall curve has no positive class, so the area beneath it is undefined. Table 3 reports these cells as not defined rather than as 0.0000, and carries the exceedance-hour count for each stream so that the rows supported by a positive class can be identified at sight. We did not lower the thresholds until exceedances appeared: choosing a threshold after seeing the exceedance count would make the base rate a tuned quantity rather than a property of the reach. The finding is that three of these reaches did not approach their screening thresholds during this window.

Rainfall-forced early warning

The leading stream contained 30 distinct exceedance episodes. At a trigger of 0.30 the twin detected 14 of them, a detection rate of 0.467, with a median lead time of 24.0 hours and a false-alarm rate of 0.032 on hours without threshold exceedance. Raising the trigger to 0.50 changed the false-alarm rate to 0.029 and left the median lead at 24.0 hours. That value is censored rather than measured: the search looks back 24

hours before each episode, so a median of 24.0 means the alarm was already raised at the start of the search window, and the true lead is greater than 24 hours by an amount this procedure cannot determine. The insensitivity of the median to the trigger is a further symptom of the same saturation. The result establishes that warning is available at a stated false-alarm cost; it does not establish how much warning.

Figure 3 shows the full lead-time trade-off as the trigger is swept from 0.05 to 0.95. The curve reports warning lead against false-alarm burden rather than defining a single universally optimal threshold.

Risk-triggered dispatch of regenerative capacity

Under baseline conditions the leading stream contained 590 threshold-exceedance hours with a threshold-exceedance magnitude index of 20425.4. Applying a sixty per cent attenuation envelope continuously reduced these to 176 hours and 3202.9 at a duty cycle of 1.0. Engaging the same capacity only when calibrated risk crossed 0.30 gave 253 threshold-exceedance hours and a magnitude index of 5594.8 at a duty cycle of 0.163.

Table 3. Calibration and discrimination of threshold-exceedance risk, nominal coverage 0.90

Station and determinand	PICP online	PICP frozen	Brier online	Brier frozen	BSS online	AUPRC online	Exceedance hours	Median width + IQR	Mean width + IQR
S1 — Dissolved oxygen	0.901	0.650	Not defined	Not defined	Not defined	Not defined	0	0.76	2.7×10^7
S1 — Specific conductance	0.898	0.707	Not defined	Not defined	Not defined	Not defined	0	0.47	7.6×10^6
S2 — Turbidity	0.898	0.595	0.0472	0.0345	0.660	0.896	590	1.11	2.88
S2 — Specific conductance	0.899	0.511	Not defined	Not defined	Not defined	Not defined	0	0.92	1.06

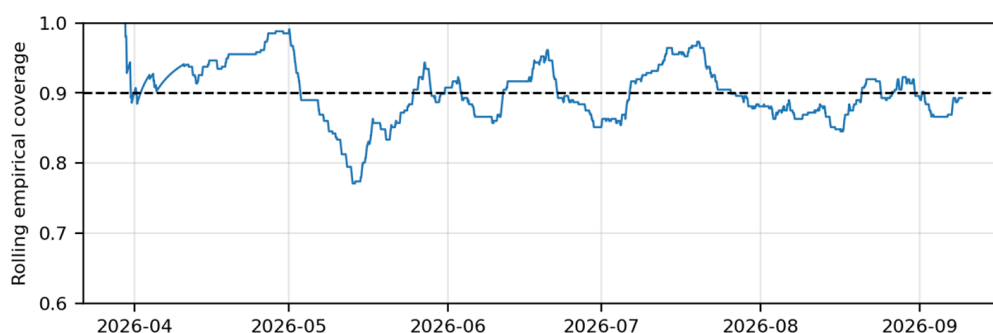


Figure 2. Rolling empirical coverage for dissolved oxygen at S1 over a fourteen-day window against the nominal 0.90 level

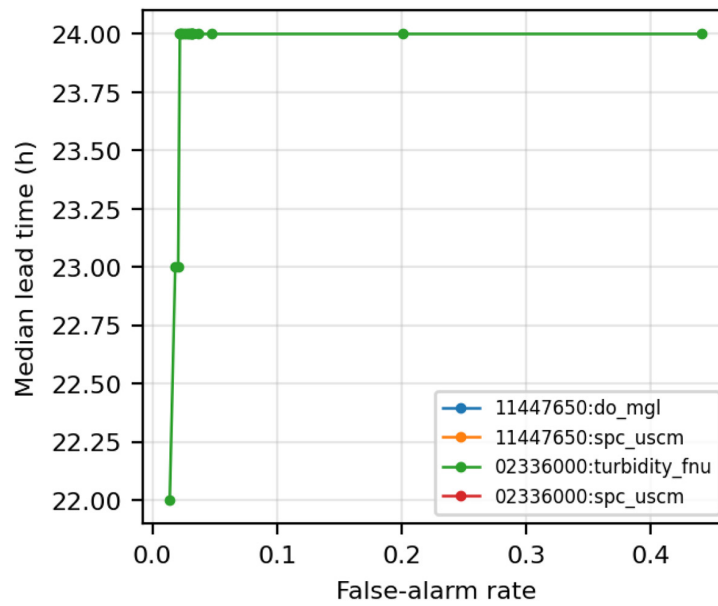


Figure 3. Median lead time against false-alarm rate as the alarm trigger is swept

Table 4. Early warning and risk-triggered regenerative-capacity dispatch, attenuation envelope $\eta = 0.60$, trigger $\tau = 0.30$

Station and determinand	Episodes	Detection rate	Median lead (h)	False-alarm rate	Base threshold exceedances	Fixed-envelope threshold exceedances	Triggered threshold exceedances	Duty cycle
S1 — Dissolved oxygen	No threshold-exceedance episode in the window; early-warning and dispatch metrics not applicable.							
S1 — Specific conductance	No threshold-exceedance episode in the window; early-warning and dispatch metrics not applicable.							
S2 — Turbidity	30	0.47	24.0	0.032	590	176	253	0.163
S2 — Specific conductance	No threshold-exceedance episode in the window; early-warning and dispatch metrics not applicable.							

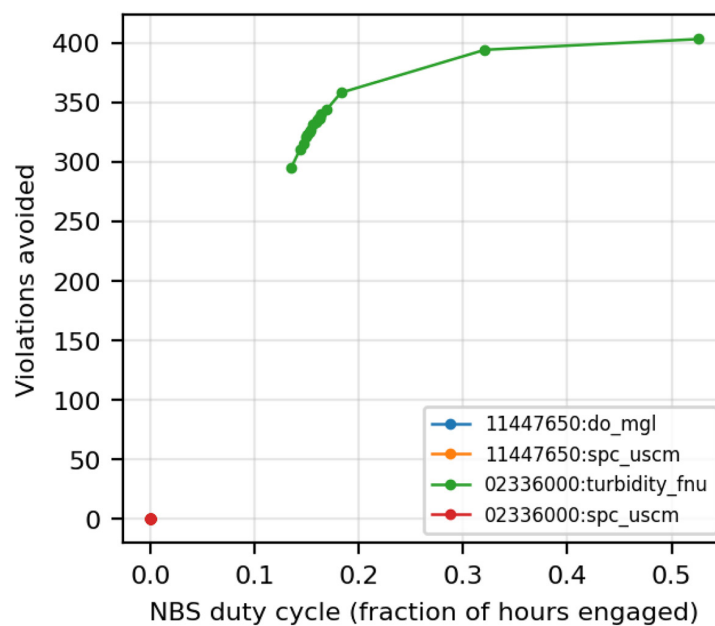


Figure 4. Threshold exceedances avoided against nature-based-solution duty cycle as the dispatch trigger is swept. The fixed 60% attenuation envelope is the reference point at a duty cycle of 1.0

Triggered dispatch therefore retained 81.4% of the threshold-exceedance benefit of continuous operation while the unit was engaged for 16.3% of the hours. Figure 4 traces the frontier as the trigger is swept. Ninety per cent of the maximum achievable benefit is reached at a duty cycle of 0.321. This is an engagement-dimension result and should be interpreted as a screening simulation rather than as a treatment demonstration.

DISCUSSION

What online adaptation actually buys

The comparison against the frozen model should be read carefully. The separation appears first in coverage rather than in error, and that ordering has a practical consequence: a plant that monitors only point accuracy will not notice its twin degrading until the degradation is large, whereas realised coverage against a nominal level provides a running signal that is interpretable without any diagnosis of why the process changed.

The single-tree ablation changes ensemble size together with drift handling, so it cannot show how much drift repair adds to online updating alone. Adaptive ensembles are more complex to operate and tune than a single incremental learner, so that increment is worth measuring on any new record, against a matched-capacity ensemble, before committing to the more complex design.

Why calibrated risk matters more than point accuracy

The offline version of this twin reported a coefficient of determination near zero for suspended solids and treated that as a limitation to be acknowledged (Rao et al., 2026). The framing here changes what such a result means. A conformal predictive system built on a weak regressor can still provide an empirical coverage assessment; the resulting intervals may simply be wide, and the exceedance probabilities may remain near the climatological base rate. That is an honest and usable answer. It tells an operator that the twin cannot resolve that determinand at that station and that decisions should not lean on it, which is more useful than a confident point forecast with no error statement attached.

This is the substantive methodological argument of the paper. Threshold-exceedance control

is a threshold-crossing problem, and threshold-crossing problems are properly expressed as probabilities with stated coverage. Reporting a coefficient of determination and then comparing a point forecast against a limit conflates two different questions, and it produces the odd situation in which a determinand can be declared unmodelable and simultaneously be used to count threshold exceedances.

Implications for regenerative system design

The dispatch results add a dimension to the sizing question. Earlier analysis, ours included, asked how efficient a nature-based unit must be and found that returns saturate (Rao et al., 2026). Once risk is calibrated, a second question opens: how often does the unit need to be engaged. The frontier in Figure 4 shows that much of the threshold-exceedance benefit is achievable at a duty cycle well below one, because threshold exceedances cluster into a small number of storm-driven episodes and the periods between them contribute little.

For a scheme with flow splitting or bypass control this has direct capital implications, since a unit that is engaged intermittently can be sized against episode volume rather than continuous flow. It also has ecological ones that this study does not address. Constructed wetlands and bio-filtration beds depend on continuous hydraulic and biological conditioning, and a unit that sits idle between storms may not deliver its nominal efficiency when it is called on (Kadlec and Wallace, 2009; Oral et al., 2020; Vymazal, 2014). The dispatch frontier presented here is an upper bound on what intermittent operation could achieve, and it should be read as a screening tool that identifies where detailed process design is worth doing.

Relation to the offline and multi-site versions of this twin

The three studies form a progression along different axes. The offline twin established the resilience metrics and the stress-testing protocol on an archived single-plant record (Rao et al., 2026). The multi-site version addressed transferability across plants and added economic and climate layers (Archana et al., 2026). The present work addresses time rather than space: it asks what happens to a twin as the process moves underneath it, and it replaces deterministic threshold

comparison with calibrated risk. The threshold-exceedance frequency and threshold-exceedance magnitude index are retained unchanged, so results are comparable across all three.

One difference in data strategy should be stated openly. The two earlier studies used plant records with effluent determinands measured at the outfall. This study uses receiving-water telemetry as a surrogate, because no public high-frequency effluent feed exists at the resolution streaming methods require. The methods transfer to outfall data without modification, and a utility with SCADA access can run the same pipeline on its own tags. What the public data buys is that the whole analysis is reproducible by anyone.

Limitations

Four limitations bound the conclusions. First and most important, turbidity, dissolved oxygen and specific conductance in the receiving water are surrogates. They respond to the outfall but also to upstream loading, in-stream processes and dilution, so the threshold-exceedance statistics describe the state of the reach and not the state of any permit. The threshold-exceedance counts reported here must not be read as regulatory findings about any named plant.

Second, the attenuation model is an efficiency envelope, as it was in the earlier work (Rao et al., 2026). It does not represent hydraulic residence time, vegetation seasonality, substrate clogging or the reduced performance of a unit brought online from idle. Third, adaptive conformal inference targets long-run coverage under distribution shift but does not prevent short intervals of under-coverage immediately after an abrupt change; Figure 2 shows that behaviour directly, and the median excursion length is reported rather than smoothed away. Fourth, the acquisition window covers 179 days of record at two stations, so seasonal generalisation is untested; the window supports a streaming-method demonstration but not a claim of full seasonal generalisation. Fifth, only one of the four station–determinand combinations recorded any threshold exceedance during the window, so every probabilistic and dispatch result rests on that single stream, and the three remaining combinations contribute to the calibration analysis alone. Sixth, the mean prediction-interval width for the two S1 streams is dominated by a small number of extreme intervals and should

not be used as a summary; the median width reported in Table 3 is the appropriate statistic, and the coverage results are insensitive to the extreme values. Seventh, the reported early-warning lead times are censored at the 24-hour search window and are therefore lower bounds.

CONCLUSIONS

This study set out to test whether a wastewater digital twin can be operated as a continuously learning system on public high-frequency telemetry, and whether reporting threshold exceedance as a calibrated probability changes what such a twin is useful for. The main findings are the following.

1. On point accuracy, the online twin did not exceed the frozen random-forest reference in R^2 in any of the four combinations, on the full stream or on the matched rows both models were scored on. The online design is not proposed as the more accurate one; its advantage is in calibration, which the matched comparison measures on identical rows.
2. Adaptive conformal calibration was closer to the nominal 0.90 coverage than static split-conformal calibration in all four combinations; in the leading case, empirical coverage was 0.898 versus 0.595 for the frozen reference.
3. Calibrated threshold-exceedance risk gave early warning of the leading stream's exceedance episodes at a stated false-alarm cost: detection rate 0.467, median lead 24.0 h, and false-alarm rate 0.032 on hours without threshold exceedance.
4. Engaging regenerative capacity only when calibrated risk crossed a threshold retained 81.4% of the threshold-exceedance benefit of continuous operation at a duty cycle of 0.163 in the screening simulation.
5. The results also show that useful threshold-exceedance uncertainty can be assessed even when point-prediction skill is not dominant; however, the public receiving-water surrogates should not be interpreted as direct permit measurements.
6. The broader methodological result is that a streaming digital twin should be evaluated pre-quentially, should report empirical coverage against a nominal level, and should express threshold crossing as a probability scored by a proper scoring rule.

Acknowledgements

The authors acknowledge the United States Geological Survey and Open-Meteo for providing public environmental data services used in this study.

REFERENCES

- Angelopoulos A.N., Bates S. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591.
- Archana P., Swathi G., Divya N.V., Mallikarjuna Rao M., Venkatesh D., Dharani Lakshmi D., Sivanagaraju S.V., Manjunath C. 2026. Transferable multi-site digital twin for wastewater treatment: Real-time prediction, economic assessment, and climate resilience. *Ecological Engineering & Environmental Technology*, 27(4), 19–29. <https://doi.org/10.12912/27197050/218520>
- Barber R.F., Candès E.J., Ramdas A., Tibshirani R.J. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 816–845.
- Bartos M., Kerkez B. 2021. Pipedream: An interactive digital twin model for urban drainage networks. *Environmental Modelling & Software*, 144, 105120.
- Bifet A., Holmes G., Kirkby R., Pfahringer B. 2010. MOA: Massive Online Analysis. *Journal of Machine Learning Research*, 11, 1601–1604.
- Bifet A., Gavaldà R. 2007. Learning from time-changing data with adaptive windowing. *Proceedings of the SIAM International Conference on Data Mining*, 443–448.
- Botturi A., Ozbayram E.G., Tondera K., Gilbert N.I., Rouault P., Caradot N., Comas J., Denysenko I., Kolisch G., Lefebvre A., Poch M., Rodríguez-Roda I., Vanrolleghem P.A. 2021. Combined sewer overflows: A critical review on best practice and innovative solutions to mitigate impacts on environment and human health. *Critical Reviews in Environmental Science and Technology*, 51(15), 1585–1618.
- Breiman L. 2001. Random forests. *Machine Learning*, 45, 5–32.
- Brier G.W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Butler D., Ward S., Sweetapple C., Astaraie-Imani M., Diao K., Farmani R., Fu G. 2017. Reliable, resilient and sustainable water management: The Safe & SuRe approach. *Global Challenges*, 1(1), 63–77.
- Central Pollution Control Board. 2017. *Environmental standards for sewage treatment plants*. Ministry of Environment, Forest and Climate Change, Government of India, New Delhi.
- Corominas L., Garrido-Baserba M., Villez K., Olsson G., Cortés U., Poch M. 2018. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling & Software*, 106, 89–103.
- Fuller A., Fan Z., Day C., Barlow C. 2020. Digital twin: Enabling technologies, challenges and open research. *IEEE Access*, 8, 108952–108971.
- Gama J., Žliobaitė I., Bifet A., Pechenizkiy M., Bouchachia A. 2014. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44.
- Gernaey K.V., Jeppsson U., Vanrolleghem P.A., Copp J.B. (Eds.) 2014. *Benchmarking of Control Strategies for Wastewater Treatment Plants*. IWA Publishing, London.
- Gibbs I., Candès E.J. 2021. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34, 1660–1672.
- Gneiting T., Raftery A.E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.
- Gomes H.M., Barddal J.P., Ferreira L.E.B., Bifet A. 2018. Adaptive random forests for data stream regression. *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 267–272.
- Gomes H.M., Bifet A., Read J., Barddal J.P., Embreck F., Pfahringer B., Holmes G., Abdesslem T. 2017. Adaptive random forests for evolving data stream classification. *Machine Learning*, 106, 1469–1495.
- Hashimoto T., Stedinger J.R., Loucks D.P. 1982. Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. *Water Resources Research*, 18(1), 14–20.
- Heddum S., Kisi O. 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology*, 559, 499–509.
- Helsel D.R., Hirsch R.M., Ryberg K.R., Archfield S.A., Gilroy E.J. 2020. *Statistical methods in water resources*. U.S. Geological Survey Techniques and Methods, Book 4, Chapter A3.
- Henze M., Gujer W., Mino T., van Loosdrecht M. 2000. *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA Publishing, London.
- Hersbach H., Bell B., Berrisford P., Hirahara S., Horányi A., Muñoz-Sabater J., Nicolas J., Peubey C., Radu R., Schepers D., Simmons A., Soci C., Abdalla S., Abellan X., Balsamo G., Bechtold P., Biavati G., Bidlot J., Bonavita M., et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.

25. Kadlec R.H., Wallace S.D. 2009. *Treatment Wetlands, 2nd ed.* CRC Press, Boca Raton.
26. Kerkez B., Gruden C., Lewis M., Montestruque L., Quigley M., Wong B., Bedig A., Kertesz R., Braun T., Cadwalader O., Poresky A., Pak C. 2016. Smarter stormwater systems. *Environmental Science & Technology*, 50(14), 7267–7273.
27. Lei J., G'Sell M., Rinaldo A., Tibshirani R.J., Wasserman L. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111.
28. Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G. 2019. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346–2363.
29. Montiel J., Halford M., Mastelini S.M., Bolmier G., Sourty R., Vaysse R., Zouitine A., Gomes H.M., Read J., Abdessalem T., Bifet A. 2021. River: Machine learning for streaming data in Python. *Journal of Machine Learning Research*, 22(110), 1–8.
30. Newhart K.B., Holloway R.W., Hering A.S., Cath T.Y. 2019. Data-driven performance analyses of wastewater treatment plants: A review. *Water Research*, 157, 498–513.
31. Oral H.V., Carvalho P., Gajewska M., Ursino N., Masi F., van Hullebusch E.D., Kazak J.K., Exposito A., Cipolletta G., Andersen T.R., Finger D.C., Simperler L., Regelsberger M., Rous V., Radinja M., Buttiglieri G., Krzeminski P., Rizzo A., Dehghanian K., Nikolova M., Zimmermann M. 2020. A review of nature-based solutions for urban water management in European circular cities. *Blue-Green Systems*, 2(1), 112–136.
32. Page E.S. 1954. Continuous inspection schemes. *Biometrika*, 41(1/2), 100–115.
33. Pedersen A.N., Borup M., Brink-Kjær A., Christiansen L.E., Mikkelsen P.S. 2021. Living and prototyping digital twins for urban water systems: Towards multi-purpose value creation using models and sensors. *Water*, 13(5), 592.
34. 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.
35. Rao M.M., Manjunath C., Venkatesh D., Sivanagaraju S.V., Vasundra S. 2026. A robust offline digital twin for regenerative wastewater treatment: Multi-pollutant stress testing and resilience analysis. *Journal of Ecological Engineering*, 27(7), 493–504. <https://doi.org/10.12911/22998993/217885>
36. Rasmussen P.P., Gray J.R., Glysson G.D., Ziegler A.C. 2009. *Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity-sensor and streamflow data.* U.S. Geological Survey Techniques and Methods, Book 3, Chapter C4.
37. Romano Y., Patterson E., Candès E.J. 2019. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 3543–3553.
38. 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.
39. Shafer G., Vovk V. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.
40. Sweetapple C., Fu G., Butler D. 2017. Reliable, resilient and sustainable urban drainage systems: An analysis of robustness under deep uncertainty. *Environmental Science & Technology*, 51(16), 9008–9021.
41. U.S. Environmental Protection Agency. 1986. *Ambient water quality criteria for dissolved oxygen.* EPA 440/5-86-003, Office of Water, Washington, D.C.
42. U.S. Geological Survey. 2026. *National Water Information System: Web interface, instantaneous values web service.* <https://waterservices.usgs.gov/nwis/iv/>
43. Vovk V., Shen J., Manokhin V., Xie M. 2019. Non-parametric predictive distributions based on conformal prediction. *Machine Learning*, 108, 445–474.
44. Vovk V., Gammerman A., Shafer G. 2005. *Algorithmic Learning in a Random World.* Springer, New York.
45. Vymazal J. 2014. Constructed wetlands for treatment of industrial wastewaters: A review. *Ecological Engineering*, 73, 724–751.
46. Wagner R.J., Boulger R.W., Oblinger C.J., Smith B.A. 2006. *Guidelines and standard procedures for continuous water-quality monitors: Station operation, record computation, and data reporting.* U.S. Geological Survey Techniques and Methods, Book 1, Chapter D3.
47. Zhu J.-J., Yang M., Ren Z.J. 2023. Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science & Technology*, 57(46), 17671–17689.
48. Zippenfenig P. 2023. Open-Meteo.com weather API. *Zenodo*. <https://doi.org/10.5281/zenodo.7970649>