Expert Systems With Applications 325 (2026) 132705

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Constraint-aware reinforcement learning for energy-efficient and regulation-compliant wastewater treatment

Omid Sobhani [a,*], Thomas Huybrechts [a], Hamid Toliati [a], Cristian Camilo Gomez Cortes [b], Kevin Mets [a], Siegfried Mercelis [a]

[a] University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Sint-Pietersvliet 7, 2000 Antwerp, Belgium
[b] BIOMATH - Model-based analysis and optimization of bioprocesses, Faculty of Bioscience, Department of data analysis and mathematical modeling, Ghent University, 9000 Ghent, Belgium

ARTICLE INFO

Keywords:
Reinforcement learning
Artificial intelligence
Optimization
Wastewater treatment

ABSTRACT

This paper addresses the sustainable and regulation-compliant control of wastewater treatment plants (WWTPs) characterized by nonlinear process dynamics, external disturbances, and multiple, often competing, effluent-quality constraints (e.g., biochemical oxygen demand, ammonia, nitrate, and phosphorus limits). We propose an Adaptive Lagrangian Soft Actor–Critic (AL-SAC) algorithm that incorporates a Lagrangian relaxation term into the maximum-entropy SAC objective in order to impose hard bounds on these key effluent variables. Lagrange multipliers are updated online based on instantaneous constraint violations, eliminating manual penalty tuning and enhancing convergence stability. The proposed AL-SAC controller is evaluated on the Benchmark Simulation Model No. 1 (BSM1) and provides up to a 23.9% reduction in energy consumption (i.e., aeration and pumping), improves the effluent quality index (EQI) by 2.92%, and significantly reduces the period over which total nitrogen (TN) and soluble ammonium nitrogen (SNH) concentrations exceed regulatory thresholds. These results demonstrate AL-SAC's promise for energy-efficient and fully compliant WWTP operation.

1. Introduction

Wastewater treatment plants (WWTPs) contribute to protecting public health and maintaining environmental quality by eliminating organic content, nutrients, and pathogens. The activated sludge process is the central element in most WWTPs. It features strongly nonlinear and time-variant behavior promoted by variable influent composition, microbial kinetics, and ambient conditions (Dickin et al., 2020). These complex dynamics make it difficult to operate WWTPs efficiently, especially under changing conditions. In addition to maintaining stable process performance, operators must optimize energy usage, particularly for aeration, which is one of the most energy-intensive components (Jamaludin et al., 2024). Furthermore, they have to ensure compliance with strict and often conflicting regulatory limits on effluent quality, such as biochemical oxygen demand (BOD), ammonia, nitrate, and phosphorus. This creates a multi-objective optimization challenge that requires balancing energy efficiency with reliable environmental compliance (Muloiwa et al., 2022; Sandino et al., 2010).

Traditional control approaches mainly rely on feedback proportional–integral (PI) loops that regulate fixed dissolved oxygen (DO) setpoints. However, under variable loading conditions, this often leads to inefficient over- or under-aeration (Holman & Wareham, 2003). Ammonia-based aeration control (ABAC) improves energy efficiency by adjusting airflow based on real-time ammonium levels, yet it still depends on manual setpoint tuning and cannot fully adapt to rapid or unexpected changes in influent characteristics (Budzynski, 2024; Schraa et al., 2019). More advanced strategies, such as ABAC-SRT, coordinate aeration with solids retention time (SRT) to stabilize biomass concentration. However, this approach introduces complexity that leads to solving a challenging multi-variable optimization problem (Gu et al., 2023). While these approaches overcome certain limitations of classical control, they continue to struggle with adaptability and scalability.

Motivated by these limitations, recent research has increasingly explored data-driven and learning-based strategies for WWTP operation. In particular, reinforcement learning (RL) offers a promising framework for sequential decision-making under dynamic and uncertain conditions, as it enables control policies to be learned directly through interaction with the process (Sutton & Barto, 2018). Despite its potential, applying RL to wastewater treatment remains challenging because operational

* Corresponding author.
E-mail addresses: Seyedomid.Sobhani@uantwerpen.be (O. Sobhani), Thomas.Huybrechts@uantwerpen.be (T. Huybrechts), Hamidreza.Toliati@uantwerpen.be (H. Toliati), CristianCamilo.GomezCortes@UGent.be (C.C. Gomez Cortes), Kevin.Mets@uantwerpen.be (K. Mets), Siegfried.Mercelis@uantwerpen.be (S. Mercelis).

https://doi.org/10.1016/j.eswa.2026.132705
Received 24 November 2025; Received in revised form 29 April 2026; Accepted 30 April 2026
Available online 5 May 2026

objectives must be balanced against strict effluent-quality requirements. Existing RL-based approaches have largely addressed such requirements through reward shaping, where regulatory violations are penalized indirectly via manually tuned terms in the reward function. Although intuitive, this strategy is often highly sensitive to hyperparameter selection and may lead to unstable trade-offs between energy use, effluent quality, and compliance. Poorly chosen penalty coefficients can allow temporary violations during exploration or distort the optimization landscape, ultimately limiting policy performance. These limitations motivate the use of constrained reinforcement learning, in which effluent-quality requirements are incorporated explicitly into the learning problem rather than indirectly through manually tuned penalties.

To the best of our knowledge, constrained RL has not yet been investigated for wastewater treatment control with explicit effluent-quality constraints while simultaneously optimizing energy efficiency. In this paper, we introduce the Adaptive Lagrangian Soft Actor–Critic (AL-SAC) algorithm to address these challenges. AL-SAC extends the standard Soft Actor–Critic (SAC) (Haarnoja et al., 2018) approach by incorporating a mechanism inspired by constrained optimization to explicitly enforce water-quality limits. Specifically, it introduces additional variables, known as dual variables, which quantify the severity of constraint violations and guide the learning process toward feasible operation. Through a primal–dual update, AL-SAC simultaneously adjusts both the control policy (the primal component) and these dual variables (the dual component) to balance performance objectives with regulatory compliance. This formulation leverages Lagrange multipliers, a classical tool from optimization theory, to dynamically adjust the penalty applied to effluent violations, allowing the algorithm to learn effective control strategies without manual tuning of penalty weights. By doing so, AL-SAC provides a systematic way to ensure that wastewater treatment plants operate efficiently while maintaining compliance with effluent quality standards. The contributions of this work can be summarized as:

- Constrained-RL formulation: We formulate activated-sludge control as a constrained Markov decision process (CMDP) and impose discharge constraints directly within the learning objective.
- Normalized logarithmic barrier cost: We apply a normalized logarithmic barrier cost function throughout the environment to provide a consistent signal across different constraints. This function penalizes the agent as the effluent quality approaches regulatory limits.
- Adaptive Lagrangian SAC: We develop on top of the maximum-entropy SAC architecture (Haarnoja et al., 2018) by introducing a primal-dual update procedure with dual variables to ensure near-constraint satisfaction in each policy update (Achiam et al., 2017).
- Sample-efficient off-policy updates: Our approach dynamically adjusts Lagrange multipliers from off-policy samples while ensuring energy efficiency and regulatory compliance without manual penalty tuning.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on data-driven, reinforcement learning, and constrained reinforcement learning approaches for wastewater treatment control. Section 3 presents the wastewater treatment benchmark, the problem formulation, the proposed AL-SAC method, and the training setup. Section 4 reports the experimental results. Section 5 discusses the main assumptions, limitations, and practical implications of the proposed approach. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Related work

Recent years have seen increasing interest in data-driven strategies for wastewater treatment control, motivated by the limitations of conventional control approaches under highly variable influent and operating conditions. Recent reviews have further emphasized the importance of dynamic modelling, monitoring, and control frameworks that can respond effectively to disturbances and changing operating conditions in real time rather than only under nominal conditions (Parsa et al., 2024). In this context, machine learning methods have been explored to capture the nonlinear dynamics of wastewater treatment plants (WWTPs) directly from operational data and to support more adaptive and efficient control strategies.

Early data-driven research in wastewater treatment focused primarily on prediction, soft sensing, and decision support rather than direct control. Zhou (2019) examined data-driven modelling for influent prediction, highlighting the potential of learning-based methods to capture plant variability from operational data. Nourani et al. (2021) and Alali et al. (2023) further applied supervised learning techniques, including artificial neural networks and support vector regression, to estimate effluent quality and other process variables relevant to plant operation. Related work also explored soft-sensing approaches to support supervision and control. Hernández-del Olmo et al. (2019) developed a machine-learning-based weather soft sensor to infer weather-related operating conditions from routinely measured plant data for advanced control support. Subsequent studies extended the use of artificial intelligence beyond prediction toward broader tasks such as process analysis, monitoring, optimization, and control support. Wang et al. (2021) proposed an interpretable machine-learning framework for effluent quality control that characterized the influence of operational variables on effluent parameters while accounting for process time lags. Nevertheless, these approaches remain primarily supportive in nature and do not directly address the sequential decision-making problem required for real-time control.

Reinforcement learning has attracted increasing attention as a framework for wastewater treatment control because it enables the learning of control policies through direct interaction with the process environment. An early application was presented by Hernandez-del Olmo and Gaudioso (2011), who developed an RL-based control strategy for the BSM1 benchmark and demonstrated the feasibility of learning-based control in this setting. Chen et al. (2021) applied multi-agent reinforcement learning to wastewater treatment control and demonstrated its potential for improving plant-wide operational performance. Hu et al. (2024) investigated deep reinforcement learning for real-time aeration control and reported energy savings relative to conventional control strategies. More recently, Croll et al. (2023b) systematically evaluated several RL algorithms on the BSM1 benchmark and showed that RL can discover energy-efficient operating policies across different control scenarios. However, their results also revealed that improved energy performance may come at the expense of effluent quality. Croll et al. (2025) further showed that the design of the reward function strongly affects training behavior, control performance, and treatment risk. Mohammadi et al. (2025) also applied Soft Actor-Critic to wastewater treatment control with delayed dynamics, highlighting both the potential of entropy-regularized RL and the practical challenges associated with such process control problems.

A common feature of existing RL-based wastewater treatment studies is the use of reward shaping to address effluent-quality requirements. Under this formulation, regulatory violations are penalized through manually designed terms added to the reward function (Croll et al., 2023a; Mohammadi et al., 2025). While this approach is intuitive, its effectiveness depends strongly on the selection of penalty weights and often requires extensive hyperparameter tuning to balance competing objectives such as energy consumption, effluent quality, and regulatory compliance (Croll et al., 2025). As shown by Croll et al. (2023b), different reward formulations and agent architectures can lead to markedly different trade-offs, ranging from substantial energy savings accompanied by severe deterioration in effluent quality to only modest improvements in treatment performance.

Constrained reinforcement learning provides a more principled alternative by incorporating operational requirements directly into the policy optimization problem. Constrained and safe formulations have attracted increasing attention in reinforcement learning for applications in which control performance must be achieved under explicit safety or resource limitations. Liu et al. (2021) reviewed policy learning under constraints in model-free reinforcement learning and summarized the main formulations based on constrained Markov decision processes. Achiam et al. (2017) introduced Constrained Policy Optimization as a practical approach for policy learning under explicit safety constraints, whereas Paternain et al. (2019) provided theoretical foundations through a zero-duality-gap formulation for constrained reinforcement learning. In addition to cumulative-constraint formulations, methods based on stability guarantees and barrier certificates have been developed for constrained dynamical systems (Han et al., 2021; Zhao et al., 2022). These developments indicate that constrained reinforcement learning has become a well-established direction for control problems involving explicit operational and safety requirements.

Despite these developments, the application of constrained reinforcement learning to wastewater treatment control remains limited. Existing studies have largely relied on unconstrained formulations based on reward shaping, while the explicit treatment of effluent-quality requirements within policy optimization remains largely unexplored.

3. Methods and materials

3.1. Benchmark simulation model no. 1

The Benchmark Simulation Model No. 1 (BSM1) is a standardized and extensively validated simulation framework developed to assess control strategies in conventional activated sludge wastewater treatment systems. It provides a dynamic, mechanistic representation of the biological treatment process, enabling reproducible evaluation of performance across different control strategies (Alex et al., 2008).

BSM1 represents the secondary treatment section of a municipal wastewater treatment plant, consisting of five continuous stirred-tank reactors (CSTRs) in series followed by a single secondary clarifier (Fig. 1). Each CSTR is modeled as a completely mixed reactor governed by dynamic mass-balance equations in Eqs. (1) and (2), with biochemical reactions described by the Activated Sludge Model No. 1 (ASM1) (Henze et al., 2006). Aeration in the aerobic reactors is represented via oxygen transfer coefficients (K_{La}), which control dissolved oxygen dynamics, while influent, recycle, and sludge flows determine hydraulic interactions between units.

The biological sequence begins with two unaerated (anoxic) reactors that facilitate denitrification, followed by three aerated reactors where nitrification and carbon oxidation occur. Reactor interconnections are established through two primary recirculation streams. The first is an internal recycle flow (IRF) that returns mixed liquor from the bottom of the fifth (last) aerobic reactor back to the head of the first anoxic reactor, promoting effective anoxic/aerobic cycling for nitrogen removal. The second is an external recycle, in which settled sludge from the clarifier underflow is conveyed back to the first reactor. Waste Activated Sludge (WAS) is continuously withdrawn from the clarifier underflow for sludge management, while return activated sludge (RAS) drawn from the same underflow stream maintains biomass concentration in the reactor train.

The state of each soluble component S_i and particulate component X_i in any CSTR is governed by mass-balance equations of the form:

$$\frac{dS_i}{dt} = \frac{Q_{\text{in}}}{V}\left(S_{\text{in},i} - S_i\right) + \sum_j R_{ij}(S, X), \tag{1}$$

$$\frac{dX_i}{dt} = \frac{Q_{\text{in}}}{V}\left(X_{\text{in},i} - X_i\right) + \sum_j R_{ij}(S, X) - \frac{Q_{\text{w}}}{V} X_i, \tag{2}$$

where V is the volume of reactor. The net inflow Q_{in} includes influent, internal recycle, and RAS, while the wastage flow Q_{w} represents WAS removal. The reaction terms R_{ij} follow the ASM1 as detailed by Henze et al. (2006), incorporating 13 state variables and eight kinetic processes covering microbial growth (heterotrophic and autotrophic), decay, hydrolysis, nitrification, denitrification, and ammonification. Reaction rates are expressed via Monod-type kinetics with stoichiometric and half-saturation parameters.

The secondary clarifier is represented by a one-dimensional, ten-layer model incorporating flux-limited solids transport dynamics based on Takács et al. (1991). A concentration-dependent settling velocity function captures three distinct settling regimes: discrete, hindered, and compressive. Solids flux between layers is described by Eq. (3):

$$J_k = v(\alpha_k)\left(X_k - X_{k+1}\right), \quad k = 1, \ldots, 9, \tag{3}$$

where X_k denotes the concentration of solids in layer k and $v(\alpha_k)$, the corresponding settling velocity. Clarifier effluent is drawn from the top layer, while sludge underflow supplies both RAS and WAS streams. Fig. 1 provides a schematic overview of the reactor train, recirculation loops, and clarifier layering.

The coupling of hydrodynamics, microbial conversions, oxygen transfer, and settling dynamics generates a high-dimensional, nonlinear, and time-varying system. Operational variations in aeration intensity (via oxygen-transfer coefficient adjustments), recycle ratios, and sludge-wasting rates exert divergent effects on effluent quality, nitrogen removal performance, and energy consumption, posing a complex, non-convex control challenge.

As a baseline for comparison, we use the standard closed-loop control strategy defined in the original BSM1 benchmark. In this configuration, dissolved oxygen in the final aerobic reactor is regulated through feedback PI control of the oxygen transfer coefficient K_{La}, while nitrate in the last anoxic reactor is regulated through feedback PI control of the internal recycle flow rate. This benchmark controller serves as the conventional reference policy against which the proposed RL agent is evaluated.

For simulation and control experiments, we utilize the open-source Python implementation of BSM1 by Miederer et al. (2025). This implementation is wrapped within an OpenAI Gymnasium environment (Towers et al., 2025), enabling episodic interactions, reward shaping, and constraint-handling mechanisms essential for RL-based controller development. The Gymnasium interface abstracts action and observation spaces, facilitating seamless integration with standard RL algorithms while preserving the modularity and extensibility of the underlying BSM1 simulator.

3.2. Constraint reinforcement learning framework

The control problem considered in this study is formulated as a Constrained Markov Decision Process (CMDP), represented by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \{C_i\}_{i=1}^m, \gamma)$ (Altman, 1999). Here, $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ the transition probability, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, $\{C_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_{i=1}^m$ a set of cost functions associated with m separate constraints, and $\gamma \in [0, 1)$ the discount factor that determines the relative importance of future rewards and costs. This formulation is well-suited to wastewater treatment control, as it captures its sequential and stochastic dynamics while allowing effluent quality constraints to be explicitly incorporated into the decision-making framework. The objective is to learn a stochastic policy $\pi(a \mid s)$ that maximizes the expected cumulative discounted reward:

$$J_R(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right], \tag{4}$$

subject to the constraint expectations:

$$J_{C_i}(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) \right] \leq d_i, \quad \forall i \in \{1, \ldots, m\}. \tag{5}$$

Eqs. (4) and (5) together define the constrained optimization problem that the policy must solve. Specifically, Eq. (4) represents the expected cumulative discounted reward under policy π, while Eq. (5) enforces

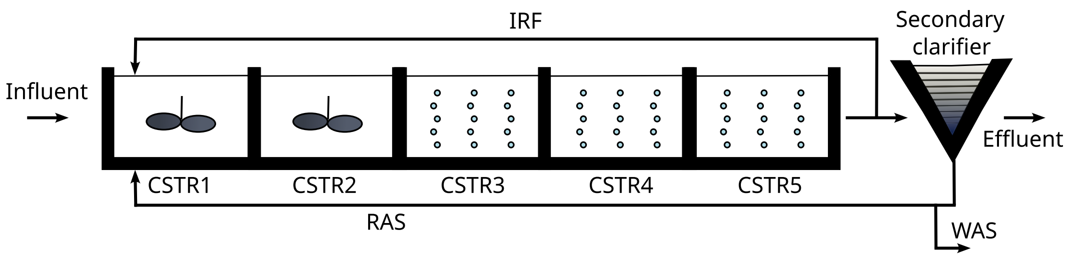


Fig. 1. Schematic overview of BSM1 (Alex et al., 2008).

that the expected discounted cost associated with each constraint remains below its corresponding limit d_i. In the context of wastewater treatment control, these constraints correspond to regulatory limits on effluent quality, ensuring that operational objectives are pursued without violating environmental requirements.

3.3. Constraint Markov decision process based on BSM1

Building on the BSM1 process description in Section 3.1 and the CMDP formulation introduced in Section 3.2, the wastewater treatment control problem is represented as a CMDP defined over the BSM1 environment. The process variables simulated by BSM1 constitute the system state, while the control inputs define the action space. Operational objectives are encoded through the reward function, and regulatory requirements are represented as constraint cost functions. The specific definitions of the state space, action space, reward, and constraint costs used in this study are presented in the following subsections.

3.3.1. State space

The state vector comprises selected variables relevant for capturing the time-dependent nitrogen dynamics of the process and informing aeration control decisions. Specifically, it includes the time of day encoded using `sine` and `cosine` functions to represent diurnal patterns, the influent ammonium concentration, the ammonium concentrations in the three aerobic reactors, and two deviation signals that quantify the extent to which the current effluent ammonium and total nitrogen concentrations exceed their respective regulatory limits. The state at time t is therefore defined as:

$$s_t = \left[\sin(2\pi t), \cos(2\pi t), \mathrm{NH}_4^{+,\mathrm{inf}}, \mathrm{NH}_4^{+,\mathrm{R3}}, \mathrm{NH}_4^{+,\mathrm{R4}}, \mathrm{NH}_4^{+,\mathrm{R5}}, \Delta\mathrm{NH}_4^{+,\mathrm{eff}}, \Delta\mathrm{TN}^{\mathrm{eff}}\right]_t \quad (6)$$

In Eq. (6), the deviation terms are defined as $\Delta\mathrm{NH}_4^{+,\mathrm{eff}} = \mathrm{NH}_4^{+,\mathrm{eff}} - \mathrm{NH}_4^{+,\mathrm{limit}}$ and $\Delta\mathrm{TN}^{\mathrm{eff}} = \mathrm{TN}^{\mathrm{eff}} - \mathrm{TN}^{\mathrm{limit}}$. This state representation is fully observable and compactly encodes temporal variation, spatial distribution of ammonium across the aerobic zone, and real-time compliance with effluent nitrogen constraints, which are critical features for learning effective aeration control policies.

3.3.2. Actions

In this study, the control policy manipulates four variables: the oxygen transfer coefficients (K_{La}) in aerobic reactors 3 to 5, and the internal recycle flow rate, expressed as a ratio relative to the influent flow ($Q_{\mathrm{int}}/Q_{\mathrm{in}}$). These variables represent key control inputs governing the biological treatment process. Regulation of K_{La} directly influences dissolved oxygen concentrations, which determine the rates of nitrification and heterotrophic activity. Independent adjustment of K_{La} in each reactor therefore enables spatial control of aerobic capacity, reducing the risk of oxygen limitation or energy losses caused by excessive aeration.

The internal recycle ratio controls the transport of nitrate and organic carbon between aerobic and anoxic zones and therefore strongly affects denitrification performance and total nitrogen removal. These variables are frequently considered in wastewater treatment control studies due to their significant impact on energy consumption, effluent quality, and process stability. Moreover, they correspond to actuators that are available in full-scale facilities, making them suitable candidates for the practical implementation of learned control policies. The manipulated variables and their ranges are shown in Table 1.

Table 1
Manipulated variables (action space) and operational bounds.

Variable	Description	Range
$K_{La,3}$	Oxygen transfer coefficient in CSTR 3	[0, 350] d^{-1}
$K_{La,4}$	Oxygen transfer coefficient in CSTR 4	[0, 350] d^{-1}
$K_{La,5}$	Oxygen transfer coefficient in CSTR 5	[0, 350] d^{-1}
$Q_{\mathrm{int}}/Q_{\mathrm{in}}$	Internal recycle ratio	[0, 5]

3.3.3. Reward and cost functions

The control objective in this study is to minimize the total energy consumption associated with wastewater treatment while ensuring compliance with effluent quality standards. Specifically, the reward function is designed to penalize excessive energy use by incorporating two major contributors to energy demand: aeration energy and pumping energy. Aeration energy, which typically constitutes the largest share of total energy usage in activated sludge systems (Jamaludin et al., 2024), is modeled as being directly proportional to the oxygen transfer coefficients $K_{La,3}$, $K_{La,4}$, and $K_{La,5}$ applied in the aerobic reactors. Similarly, the energy required for internal recirculation is modeled as a function of the internal recycle flow rate, represented by the ratio $Q_{\mathrm{int}}/Q_{\mathrm{in}}$.

By formulating the reward as the negative sum of these energy terms, the RL agent is encouraged to learn control strategies that minimize operational energy consumption while respecting effluent constraints. This approach aligns with practical operational goals in full-scale wastewater treatment plants, where energy efficiency must be balanced against process stability and regulatory compliance.

In addition to minimizing energy consumption, the agent is tasked with maintaining effluent quality within acceptable regulatory limits. To encode these requirements, we define a continuous, differentiable cost function that penalizes violations of effluent constraints while preserving numerical stability and sensitivity near the constraint boundaries.

We adopt a normalized log-barrier formulation in which the cost increases as the effluent concentration x approaches its prescribed threshold from below, while a linear penalty is imposed when the threshold is exceeded. The function is defined in Eq. (7).

$$\mathrm{cost}(x) = \begin{cases} -\log\left(\max\left(\epsilon, \frac{\tau_x - x}{\tau_x}\right)\right), & \text{if } x < \tau_x \\ -\log(\epsilon) - k \cdot \frac{\tau_x - x}{\tau_x}, & \text{if } x \geq \tau_x \end{cases} \quad (7)$$

Here, τ_x is the threshold associated with variable x and the term $(\tau_x - x)/\tau_x$ represents the normalized relative distance to the constraint boundary. The log-barrier ensures that the cost grows sharply as the variable approaches the threshold from the feasible side, promoting conservative operation near constraint limits. The overflow term introduces a linear penalty beyond the threshold, scaled by a factor k, to discourage violations while maintaining continuity and differentiability at the boundary. The constant ϵ ensures numerical stability by avoiding undefined values in the logarithm.

Table 2
Effluent nitrogen limits based on BSM1 benchmark.

Variable	Description	Limit [g/m^3]
SNH_{eff}	Effluent ammonium nitrogen	4
TN_{eff}	Effluent total nitrogen	18

This formulation yields a smooth and convex surrogate for hard constraints and is particularly suitable for integration into constrained RL frameworks, where gradient-based optimization and cost shaping are critical for stable learning.

In this work, we impose constraints on two key effluent quality indicators: ammonium nitrogen (SNH_{eff}) and total nitrogen (TN_{eff}). These variables are selected due to their high regulatory importance and environmental impact. Ammonium is directly toxic to aquatic organisms and must be effectively nitrified, while total nitrogen reflects the aggregate nutrient load, including ammonium, nitrate, and organic nitrogen species. Both metrics are routinely monitored in wastewater treatment operations and are also used within the BSM1 benchmark to assess plant performance. Accordingly, we adopt the standard BSM1 thresholds for these variables, as shown in Table 2.

These constraints ensure that the learned control policies do not compromise effluent quality while optimizing energy efficiency. Notably, these two indicators frequently exceed their respective thresholds in the BSM1 baseline operation, whereas other regulated components (e.g., BOD, COD, TSS) typically remain within acceptable limits. Furthermore, SNH and TN exhibit a trade-off: increasing aeration promotes nitrification and reduces SNH, but may elevate TN due to nitrate accumulation if denitrification is insufficient. Conversely, promoting denitrification can reduce TN, but may compromise nitrification, leading to higher SNH levels. This trade-off highlights the importance of explicitly incorporating both variables as constraints to guide balanced and environmentally compliant control strategies.

3.4. Adaptive Lagrangian soft actor-critic

To address the constrained control problem formulated as a CMDP, we propose an Adaptive Lagrangian Soft Actor-Critic (AL-SAC) algorithm. This method integrates a Lagrangian relaxation mechanism into the Soft Actor-Critic (SAC) framework, enabling principled constraint handling in continuous state-action spaces. SAC is particularly well-suited as a foundation for this work due to its sample efficiency, off-policy learning, and inherent stability, making it highly effective in complex, high-dimensional control environments such as wastewater treatment plants.

The SAC algorithm optimizes a stochastic policy by maximizing a reward signal augmented with an entropy term, promoting both efficient exploration and convergence stability. In addition, SAC's off-policy nature allows efficient reuse of experience data, which is critical when interacting with computationally expensive simulators, such as BSM1. Its twin Q-network architecture and use of target networks further reduce overestimation bias and stabilize value estimation.

To incorporate constraints into SAC, we adopt a Lagrangian dual formulation that introduces a set of non-negative multipliers $\lambda_i \geq 0$, one for each constraint C_i. This reformulates the CMDP objective defined in Eqs. (4) and (5) as the following Lagrangian function:

$$\mathcal{L}(\pi, \lambda) = J_R(\pi) - \sum_{i=1}^{m} \lambda_i \left(J_{C_i}(\pi) - d_i \right), \tag{8}$$

where $J_R(\pi)$ denotes the expected discounted return and $J_{c_i}(\pi)$ denotes the expected discounted cost associated with constraint i. The constant d_i represents the allowable upper bound for constraint i. Building on Eq. (8), the AL-SAC algorithm optimizes the entropy-regularized objective in Eq. (9):

$$J_{\text{AL-SAC}}(\pi) = \sum_{t=0}^{\infty} \mathbb{E}_{(s_t,a_t)\sim\pi} \left[r(s_t, a_t) - \sum_{i=1}^{m} \lambda_i c_i(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot \mid s_t)) \right], \tag{9}$$

where $c_i(s_t, a_t)$ is the instantaneous cost associated with constraint i, $\mathcal{H}(\pi(\cdot \mid s_t)) = -\mathbb{E}_{a_t\sim\pi}[\log \pi(a_t \mid s_t)]$ denotes the policy entropy, and $\alpha > 0$ is the temperature parameter, which is automatically tuned to achieve a target entropy. The resulting optimization problem corresponds to the saddle-point formulation in Eq. (10), which is commonly used in constrained optimization (Giorgi et al., 2023):

$$\min_{\lambda \geq 0} \max_{\pi} \mathcal{L}(\pi, \lambda). \tag{10}$$

Here, the policy π is updated to maximize the Lagrangian objective, while the multipliers (λ_i) are adjusted to enforce constraint satisfaction. The multipliers are updated using projected gradient ascent:

$$\lambda_i \leftarrow \left[\lambda_i + \eta \left(J_{C_i}(\pi) - d_i \right) \right]_+, \quad \forall i, \tag{11}$$

Algorithm 1 Adaptive Lagrangian Soft Actor-Critic (AL-SAC).

Initialize policy parameters ϕ, reward critic parameters $\theta_r^{(1,2)}$, cost critic parameters $\theta_c^{(1,2)}$ with m heads (one per constraint $i = 1 \ldots m$), target networks, dual variables $\lambda_i \geq 0$, and replay buffer $\mathcal{D}$
for each iteration **do**
  for environment step **do**
    Sample $a_t \sim \pi_\phi(\cdot|s_t)$, observe r_t, $\{c_{i,t}\}$, s_{t+1},
    store $(s_t, a_t, r_t, \{c_{i,t}\}, s_{t+1})$ in $\mathcal{D}$
  end for
  for gradient step **do**
    Sample minibatch $(s, a, r, c_i, s') \sim \mathcal{D}$
    Reward critic:
    $y_r = r + \gamma(\min_j Q_r^{(j)}(s', a') - \alpha \log \pi(a'|s'))$;
    update $Q_r^{(j)}$ minimizing $\|Q_r^{(j)}(s, a) - y_r\|^2$
    Cost critics:
    for $i = 1 \ldots m$ **do**
      $y_{c_i} = c_i + \gamma \max_j Q_{c_i}^{(j)}(s', a')$;
      update $Q_{c_i}^{(j)}$ minimizing $\|Q_{c_i}^{(j)}(s, a) - y_{c_i}\|^2$
    end for
    Policy: $\nabla_\phi J_\pi = \mathbb{E}[\alpha \log \pi_\phi(a|s) - \min_j Q_r^{(j)}(s, a) + \sum_i \lambda_i Q_{c_i}(s, a)]$
    for $i = 1 \ldots m$ **do**
      $\lambda_i \leftarrow [\lambda_i + \eta(\mathbb{E}[Q_{c_i}(s, a)] - d_i)]_+$
    end for
  end for
end for

where $\eta > 0$ is the learning rate, and $[x]_+ = \max(x, 0)$ ensures non-negativity. No additional stopping criterion or upper bound is imposed on the multipliers; they are updated throughout training and may increase or decrease adaptively depending on the degree of constraint satisfaction. In practice, the expectation $J_{C_i}(\pi)$ in Eq. (11) is approximated using mini-batch samples drawn from the replay buffer. Specifically, during each gradient update step, the dual variables are updated using the mean predicted constraint cost over the sampled batch, as shown in Algorithm 1. The update therefore relies on the critic estimates $Q_{c_i}(s, a)$ as an approximation of the expected discounted cost for constraint i. This results in an online primal–dual update in which the policy parameters are optimized via gradient descent on the Lagrangian objective, while the dual variables λ_i are updated via projected gradient ascent to penalize constraint violations. This adaptive update eliminates the need for manual penalty tuning and allows the agent to dynamically balance performance against constraint satisfaction.

The training architecture maintains the following components:

- A stochastic actor $\pi_\phi(a|s)$, parameterized by ϕ, representing the control policy.

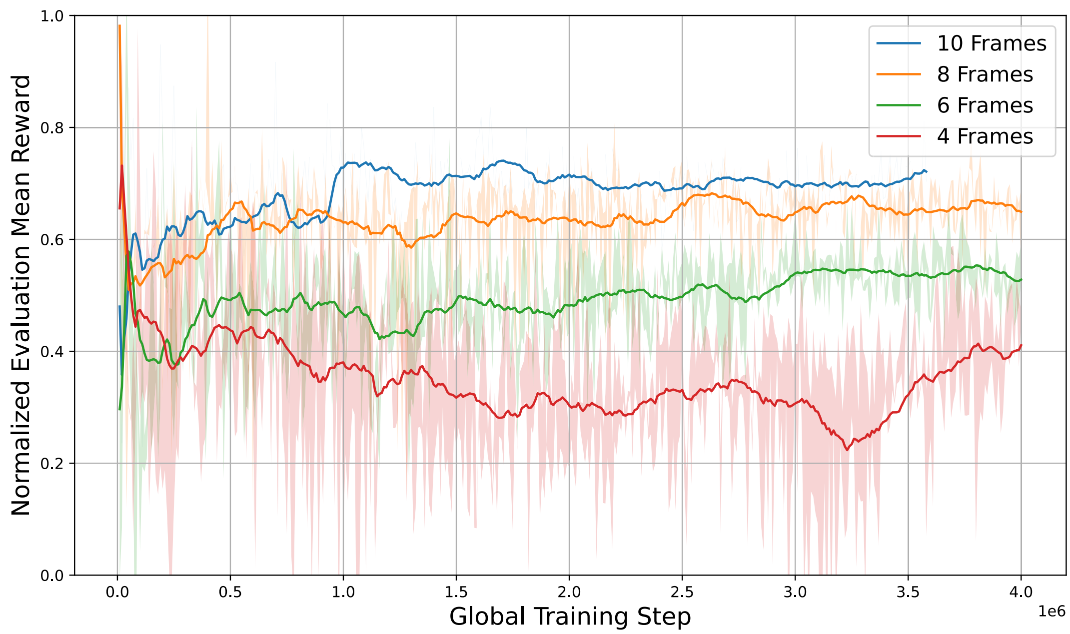


Fig. 2. Normalized moving average of evaluation mean rewards across different frame stack size.

- Two reward critics $Q^{(1)}_{\theta_r}$, $Q^{(2)}_{\theta_r}$, parameterized by θ_r to estimate the soft value of the reward objective.
- Two cost critics $Q^{(1)}_{\theta_c}$, $Q^{(2)}_{\theta_c}$, parameterized by θ_c, each with multiple heads corresponding to cost functions c_i, which estimate vectors of expected constraint violations.

Critics are trained via Bellman targets for both reward and cost objectives (Sutton & Barto, 2018). Specifically, the reward critics minimize the temporal-difference (TD) error with targets defined in Eq. (12):

$$y_r = r(s,a) + \gamma\, \mathbb{E}_{s'\sim P}\left[\min_{j\in\{1,2\}} Q_r^{(j)}\left(s', a' \sim \pi\right) - \alpha \log \pi(a' \mid s')\right]. \tag{12}$$

For the cost critics, the maximum over the two target Q-functions is used to obtain a conservative estimate of future constraint violations, thereby promoting safety by reducing the risk of underestimating potential costs. The cost critics are trained by minimizing the TD error with targets defined in Eq. (13):

$$y_{c_i} = c_i(s,a) + \gamma\, \mathbb{E}_{s'\sim P}\left[\max_{j\in\{1,2\}} Q_{C_i}^{(j)}\left(s', a' \sim \pi\right)\right], \quad \forall i \in \{1,\dots,m\}. \tag{13}$$

The actor is optimized using the gradient expression in Eq. (14), derived from the adjusted Q-values:

$$\nabla_\phi J_\pi = \mathbb{E}_{s_t\sim \mathcal{D}}\left[\nabla_\phi \alpha \log \pi_\phi(a_t|s_t) - Q_{\text{lagr}}(s_t,a_t)\right], \tag{14}$$

where $\mathcal{D}$ denotes the replay buffer and the adjusted critic is defined in Eq. (15):

$$Q_{\text{lagr}}(s_t,a_t) = Q_{\theta_r}(s_t,a_t) - \sum_{i=1}^{m} \lambda_i Q_{\theta_{c_i}}(s_t,a_t), \tag{15}$$

This formulation encourages the policy to optimize reward while simultaneously accounting for the specific constraints. As a result, the approach provides a robust and adaptive solution framework for constrained control in wastewater treatment systems. A step-by-step summary of the AL-SAC training loop is provided in Algorithm 1.

3.5. Training protocol and hyperparameters

The control agent is trained in an episodic simulation environment based on BSM1 dynamics. Each training episode begins from a predefined steady-state operating condition and continues for a simulated duration of two weeks. Unlike the standard BSM1 protocol that samples one of three fixed influent profiles (dry, rain, storm) at episode start, for the training step, we generate a new unique influent trajectory for every episode using the dynamic generator ensuring broad coverage of operating conditions during training.

The generator combines three main components: diurnal profiles, stochastic fluctuations, and weather events. Diurnal patterns provide the baseline variation in flow and pollutant concentrations across the day. Random fluctuations are then added to mimic short-term variability and measurement noise. Finally, weather events such as rain or storms are introduced at random times with varying intensity and duration. The generator produces time series for flow, total COD, total nitrogen, and total phosphorus, which are subsequently converted into the full BSM1 influent state vector. This conversion involves dilution effects, redistribution of COD fractions during first-flush events, and nitrogen fractionation according to BSM1 conventions. The resulting profiles maintain the internal consistency of BSM1 while introducing realistic variability. The time resolution is fixed at 15 minutes, in line with the BSM1 benchmark definition (Alex et al., 2008) and consistent with typical monitoring and control intervals in full-scale plants (Åmand et al., 2013). This design provides a balance between computational feasibility and practical relevance (Gernaey et al., 2011; Saagi, 2017).

The environment is discretized into 15-minute intervals, meaning that the agent receives a new observation and selects a control action every 15 minutes of simulated time. Consequently, the learned control policy operates on a practically relevant timescale. At the end of each episode, the environment is reset to its steady-state configuration to ensure consistent initialization. This training protocol enables stable learning under realistic process dynamics while facilitating efficient exploration across operational scenarios.

To enhance temporal representation and improve the agent's ability to capture delayed system dynamics, a frame-stacking strategy is employed. Specifically, the input to the policy network comprises the eight most recent observation vectors, which together represent two hours of historical data. The choice of using eight frames was guided by empirical evaluations in which several stack sizes, including 4, 6, 8, and 10, were systematically compared. As shown in Fig. 2, these experiments demonstrated that increasing the number of stacked frames led to improvements in training stability and overall control performance. However, further increasing the stack size from 8 to 10 provided only marginal gains. Consequently, a stack size of 8 was selected as the minimal con-

Table 3
Main hyperparameters used in the AL-SAC controller.

Parameter	Value
Actor and critic hidden layers	[256, 256] (ReLU)
Cost critic hidden layers	[256, 256] (ReLU) + *linear* output
Discount factor for reward, γ	0.99
Discount factor for cost, γ_c	0.90
Learning rate (actor and critics)	1×10^{-3}
Dual learning rate, η	1×10^{-3}
Batch size	128

figuration that yields reliable performance benefits without incurring unnecessary computational cost.

We adopt the default settings of Stable-Baselines3 (SB3, v2.3) for Soft Actor-Critic (Raffin et al., 2021). Both the actor and critic share a two-layer perceptron with ReLU activations, whereas the cost critic uses the same structure of hidden layers, but a linear output layer to keep the cost signal unbounded. We retain SB3's default optimizer and entropy-regularization schedule; the only deviations are the discount factor for costs ($\gamma_c = 0.90$) and the aforementioned linear activation. Table 3 represents the main hyperparameters used for the training.

4. Results

All simulations are initialized from a common steady-state solution of the activated sludge reactor model. Starting from these conditions, each control strategy is simulated over a two-week period (14 days). For the evaluation phase, only the three influent profiles provided in the BSM1 benchmark are used, namely the dry-weather, rain-weather, and storm-weather scenarios, each defined by its characteristic flow and pollutant-load time series. By applying the same initialization and two-week evaluation window across all profiles, we ensure that observed differences in performance arise solely from the control logic rather than from transient start-up effects. Results for each profile are reported relative to the standard BSM1 baseline control strategy, in which dissolved oxygen and internal recycle flow are regulated by conventional PI controllers.

To evaluate and compare control strategies in an activated-sludge plant, we employ three well-established performance metrics plus a newly proposed "integral above threshold”. First, the energy consumption is reported in kWh/d and combines the electricity used by the aeration blowers and pump units. Since aeration is the dominant contributor to energy consumption in wastewater treatment plants, reducing the specific energy consumption of the process can significantly decrease operational costs.

Second, we calculate the Effluent Quality Index (EQI) as defined in BSM1. EQI is defined in Eq. (16) as the flow- and pollutant-weighted mass discharge over the evaluation period as follows:

$$\mathrm{EQI} = \frac{1}{1000\,(t_f - t_0)} \int_{t_0}^{t_f} \left(\sum_k \beta_k\, C_k(t) \right) Q_e(t)\, \mathrm{d}t\,, \tag{16}$$

where t_0 and t_f denote the start and end of the simulation window (in days), $Q_e(t)$ [m^3.d^{-1}] is the instantaneous effluent flow rate, $C_k(t)$ [g.m^{-3}] is the concentration of pollutant component k (e.g., TSS, COD, BOD, TKN, NO_3^-, P_{tot}) in the effluent. The prefactor $1/[\,1000\,(t_f - t_0)\,]$ converts the integral into average kg/d over the period. Finally, β_k is a dimensionless weighting factor for component k, and in BSM1 are $\beta_{\mathrm{TSS}} = 2$, $\beta_{\mathrm{COD}} = 1$, $\beta_{\mathrm{BOD}} = 2$, $\beta_{\mathrm{TKN}} = 20$, $\beta_{\mathrm{NO_3}} = 20$, and $\beta_{P_{\mathrm{tot}}} = 20$.

Third, we report the time during which each effluent parameter p surpasses its imposed limit L_p. Specifically, we count the number of 15-minute samples for which $C_p(t) > L_p$ and convert this count to hours (since each sample corresponds to $\Delta t = 0.25\,\mathrm{h}$). This value, expressed either in hours or as a percentage of the total samples, quantifies the duration for which the effluent exceeded its regulatory threshold. This continuous metric captures both short-duration spikes and longer excursions, rather than treating exceedance as a simple Boolean condition.

Finally, we introduce the integral above threshold (IAT), defined in Eq. (17), to capture both the duration and magnitude of non-compliance:

$$\mathrm{IAT}_p = \sum_{t\,:\,C_p(t) > L_p} \left[C_p(t) - L_p \right] \Delta t, \quad \Delta t = 0.25\ \mathrm{h}. \tag{17}$$

By summing $\left[C_p(t) - L_p\right]$ for each time step above the limit and multiplying by Δt, IAT (in $\mathrm{mg \cdot h\, L^{-1}}$) penalizes both large deviations and extended durations of exceedance. Consequently, IAT can distinguish between strategies that exhibit the same total "hours above limit” but differ in the magnitude of exceedance, as well as those with similar exceedance magnitudes but differing durations. This makes IAT a more informative indicator of potential downstream environmental impact, as it accounts for both the severity and duration of exceedances, unlike time-above-limit which only reflects duration. All evaluation metrics are defined such that lower values correspond to better performance.

4.1. Energy consumption

Over the 14-day simulation horizon across the three influent profiles, the RL-based control strategy achieves an average daily energy consumption (aeration and pumping) of 3034.4 kWh/d, representing a 23.2% reduction relative to the baseline of 3950.1 kWh/d. Table 4 provides a breakdown of the separate aeration and pumping energy demands for each profile under both control strategies.

4.2. Effluent quality and compliance

Table 5 compares the influent-quality index (IQI) and effluent-quality index (EQI) for the baseline controller and the RL agent, and reports the RL agent's percentage improvement. Across all three weather profiles, the RL agent consistently reduces EQI, achieving gains that range from roughly 0.64% to 2.92%.

Table 6 summarizes the RL agent's performance in meeting regulatory effluent-quality limits relative to the baseline controller. Across all three influent conditions, the RL-based controller substantially reduces both the duration and magnitude of effluent exceedances for total nitrogen (TN) and soluble ammonia (SNH). Under dry-weather conditions, TN exceedances are eliminated entirely, while SNH exceedance time decreases from 55.75 h to 20.75 h and the corresponding IAT decreases from 116.94 to 11.06 mg·h L^{-1}. Under rain-event conditions, TN exceedance time decreases from 38.50 h to 2.50 h, with the IAT reduced from 42.20 to 0.21 mg·h L^{-1}. Over the same profile, SNH exceedance time decreases from 71.25 h to 16.50 h, while the IAT is reduced from 140.89 to 8.31 mg·h L^{-1}. Storm-event results show similarly strong improvements: TN exceedances are again eliminated entirely, and SNH exceedance time decreases from 71.50 h to 21.25 h, with the corresponding IAT reduced from 150.89 to 10.72 mg·h L^{-1}. Effluent solids (TSS) remain essentially within limits in both cases, with only 0.50 h of exceedance under storm conditions; however, the cumulative TSS exceedance remains lower under RL than under the baseline. COD and BOD do not exceed their respective limits under any profile.

In the violin plots of Fig. 3a, the TN violins reveal that the baseline scenario consistently produces higher median concentrations and heavier upper tails than the RL agent across all three hydrologic profiles. For baseline, the 95$^{\mathrm{th}}$-percentile TN values are 19.71 mg/L in dry, 19.30 mg/L in rain, and 19.65 mg/L in storm profiles-each exceeding the 18 mg/L limit, whereas the RL agent's corresponding 95$^{\mathrm{th}}$-percentiles are all below the limit with, respectively, 17.31 mg/L, 17.32 mg/L, and 17.21 mg/L. This indicates that extreme TN excursions are more frequent under baseline operation.

In Fig. 3b, SNH distributions show a similar pattern: baseline's 95$^{\mathrm{th}}$-percentile SNH reaches 6.35 mg/L (dry), 6.64 mg/L (rain), and 6.61 mg/L (storm), which is well above the 4 mg/L threshold. In contrast, the RL agent's upper-tail SNH distribution remains much tighter at 4.20 mg/L, 3.98 mg/L, and 4.12 mg/L, respectively. Taken together, the violins and 95$^{\mathrm{th}}$-percentile metrics demonstrate that the RL agent

Table 4
Daily aeration and pumping energy consumption (kWh/d) for the BSM1 baseline and RL control.

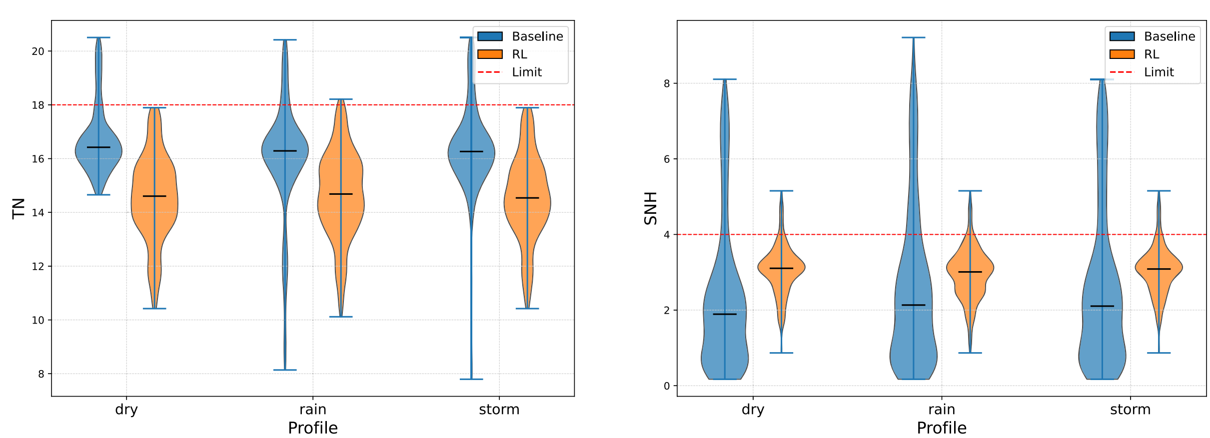
Profile	Baseline Aero (kWh/d)	Baseline Pump (kWh/d)	RL Aero (kWh/d)	RL Pump (kWh/d)	Total Reduction (%)
Dry-weather	3696.43	240.65	2779.74	294.70	21.9%
Rain-weather	3679.56	269.76	2621.94	387.23	23.8%
Storm-weather	3703.42	260.61	2687.05	332.53	23.9%

Table 5
Comparison of influent-quality index (IQI), EQI and removal efficiency for baseline vs. RL control.

Profile	Strategy	IQI (kg/d)	EQI (kg/d)	Improvement over baseline (%)
Dry-weather	Baseline	52 081.46	5 320.10	–
	RL		**5 164.67**	**2.92%**
Rain-weather	Baseline	52 081.46	6 121.41	–
	RL		**6 082.10**	**0.64%**
Storm-weather	Baseline	52 953.34	5 748.10	–
	RL		**5 607.00**	**2.45%**

Table 6
Effluent exceedance metrics for TN, SNH and TSS under baseline and RL control across influent profiles.

Profile	Parameter	Run	Hours Exceeded	Time Percent Exceeded (%)	IAT($mg.hL^{-1}$)
Dry-weather	TN	Baseline	47.25	15.06%	60.23
		RL	**0.00**	**0.00%**	**0.00**
	SNH	Baseline	55.75	16.59%	116.94
		RL	**20.75**	**6.17%**	**11.06**
Rain-weather	TN	Baseline	38.50	11.48%	42.20
		RL	**2.50**	**0.74%**	**0.21**
	SNH	Baseline	71.25	21.20%	140.89
		RL	**16.5**	**4.92%**	**8.31**
Storm-weather	TN	Baseline	44.75	13.32%	56.19
		RL	**0.00**	**0.00%**	**0.00**
	SNH	Baseline	71.50	21.28%	150.89
		RL	**21.25**	**6.32%**	**10.72**
	TSS	Baseline	0.50	0.15%	0.12
		RL	0.50	0.15%	**0.03**



(a) (b)

Fig. 3. Violin plots of (a) Total Nitrogen (TN) and (b) Soluble Ammonium-Nitrogen (SNH, NH_4^+-N) distributions under Baseline (blue) and RL (orange) configurations across dry, rain, and storm profiles. Dashed lines mark regulatory limits.

substantially reduces the frequency of extreme nutrient concentrations compared to the baseline in all flow regimes.

As shown in Fig. 4, over the 24-h dry-weather effluent record, the baseline scenario exhibits distinct midday peaks with TN (top) and SNH (bottom) exceeding their respective limits of 18 and 4 mg/L. In contrast, the RL agent attenuates these peaks. TN briefly approaches the 18 mg/L limit without exceeding it, while SNH surpasses 4 mg/L with reduced magnitude and shorter duration compared to the baseline scenario. The overnight and early-morning concentrations are similar for both scenarios, but from 08:00 onward, the RL agent maintains a smoother profile, reducing both the magnitude and duration of regulatory-exceeding loads and thereby improving effluent compliance.

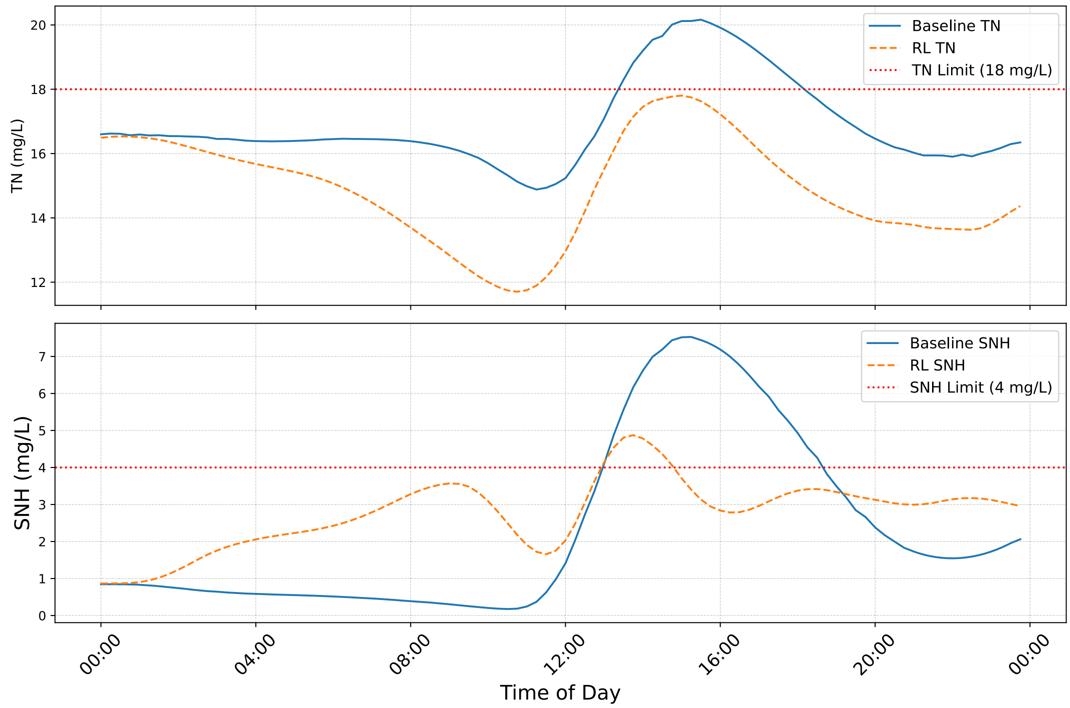


Fig. 4. Effluent TN (top) and SNH (bottom) over 24 h (first 96 samples at 15 min intervals) for Baseline (blue solid) and RL (orange dashed).

Table 7
Comparison of control performance under the one-year influent profile.

Scenario	Aero + Pump (kWh/d)	EQI (kg/d)	Energy saving	EQI improvement
Benchmark	3834	6078	–	–
ABAC[a]	3307	5897	13.7%	3.0%
TD3-0[a]	3310	5980	13.7%	1.6%
AL-SAC	**3049**	**5445**	**20.5%**	**10.4%**

[a] Values reported by Croll et al. (2023b) for the BSM1 benchmark configuration.

4.3. Control strategy evaluation

Fig. 5 compares the aeration profiles (K_{La}) in CSTR 3–5 and the internal recycle ratio under the baseline and RL-controlled scenarios over a five-day period. Under the baseline, K_{La} in CSTR3 and CSTR4 remains approximately constant at 240 h^{-1}, while CSTR5 shows a smoother time-varying profile within a lower operating range. In contrast, the RL controller generates markedly more dynamic and demand-responsive aeration patterns. In CSTR3 and CSTR4, the RL policy frequently switches between very low aeration and values near the upper operating range. A similar but even more pronounced switching behavior is observed in CSTR5, where K_{La} alternates rapidly between near-zero and high aeration levels. Overall, these profiles indicate that the RL controller applies aeration more selectively and intermittently than the baseline strategy.

Furthermore, the internal recycle ratio behavior differs substantially between the baseline and RL-controlled cases. The baseline profile varies continuously over time, whereas the RL policy keeps the recycle ratio close to 1 for extended periods and introduces brief spikes up to approximately 5 at selected intervals. This suggests that the RL controller uses internal recycle more selectively, in combination with adaptive K_{La} adjustments, rather than following the smoother modulation observed in the baseline case. Such behavior is consistent with a control strategy that concentrates corrective action during selected periods while allowing lower-intensity operation otherwise.

4.4. Long-term operation and robustness

In addition to the standard BSM1 evaluation scenarios, we perform a long-term simulation to assess the behavior of the AL-SAC controller over extended continuous operation. For this purpose, a one-year influent profile is generated following the methodology of Croll et al. (2023b).

Table 7 summarizes the performance of the proposed AL-SAC controller against the conventional ammonia-based aeration control (ABAC) and the TD3-0 agent reported by Croll et al. (2023b). For a fair comparison, all three strategies are evaluated under the same one-year influent profile. Relative to the benchmark PI control, ABAC and TD3 achieve moderate energy savings of approximately 13–14%, with only limited improvements in effluent quality (about 1–3%). In contrast, the AL-SAC controller reduces total energy consumption by 20.5% while improving the EQI by 10.4%. Notably, the TD3 agent employs reward shaping to indirectly balance regulatory compliance with operational cost, whereas the AL-SAC controller applies an adaptive Lagrangian optimization approach that enforces effluent-quality constraints explicitly during training. This difference in formulation enables AL-SAC to maintain regulatory compliance throughout the entire year without ad-hoc reward tuning, while achieving superior energy efficiency and effluent performance.

To further assess robustness, we evaluate the trained AL-SAC policy under five observation-uncertainty scenarios during the one-year simulation. In these experiments, relative Gaussian noise is applied to the

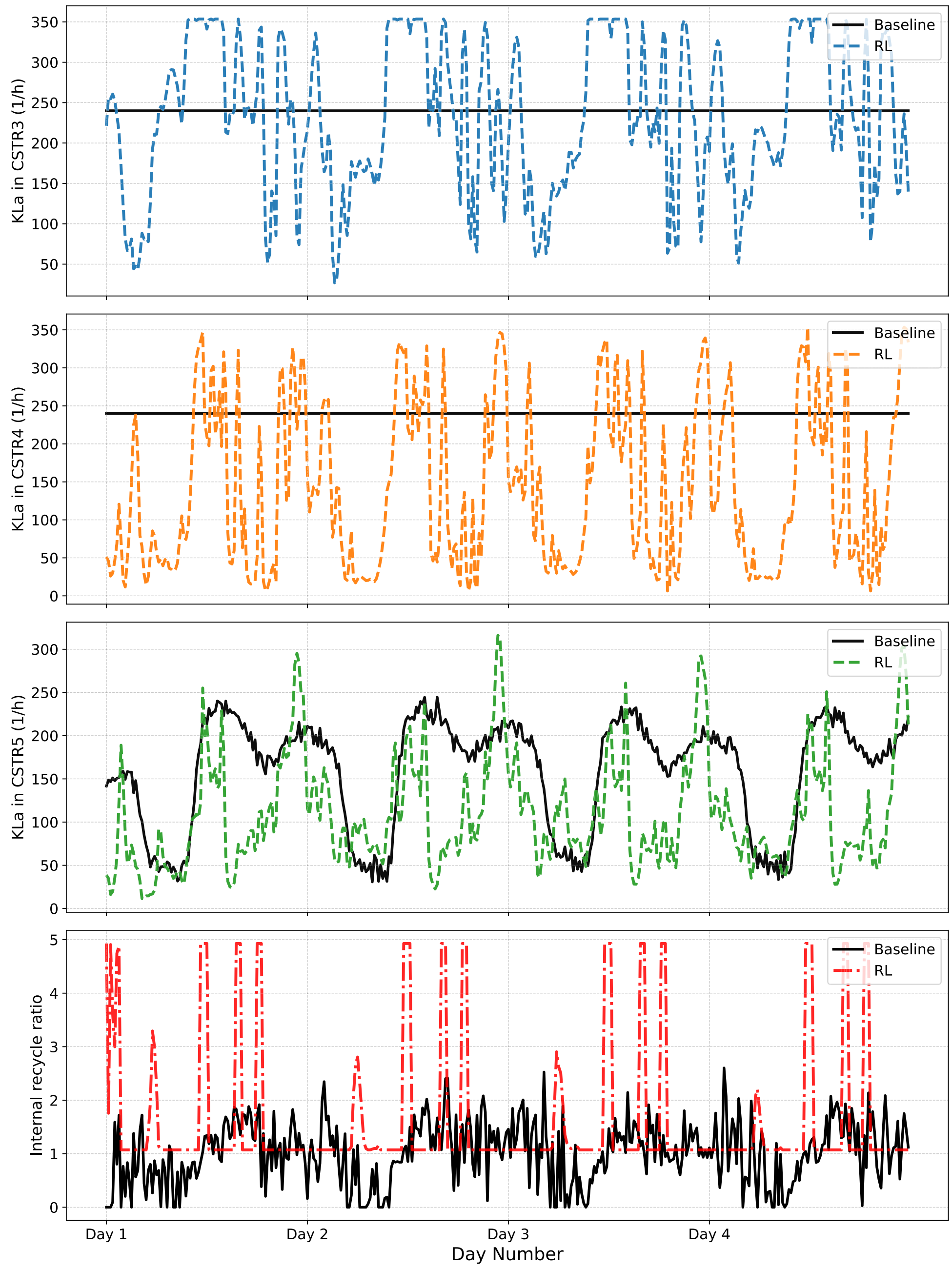


Fig. 5. Comparison of K_{La} (h^{-1}) in CSTRs 3–5 and the internal recycle ratio for the reinforcement learning (RL) and baseline scenarios over a five-day period.

concentration-related observation channels, and selected scenarios additionally include additive bias shifts. The time-of-day features remain unchanged. These scenarios are intended as robustness stress tests and should not be interpreted as an exact implementation of the default BSM1 sensor model.

The results shown in Table A.2 indicate that AL-SAC remains robust under all tested scenarios. Relative to the noise-free case, the increase in average daily energy consumption is minor, while EQI changes only marginally. Similarly, the exceedance percentages for SNH and TN remain close to their baseline values. These findings suggest that the learned policy is not overly sensitive to the tested levels of measurement noise and moderate bias. The full specification of the scenarios and the corresponding results are provided in Appendix A.

5. Discussions

The practical deployment of the proposed controller raises several implementation considerations that are relevant for real wastewater treatment plants. While the results obtained in the BSM1 benchmark demonstrate the potential of constrained reinforcement learning for improving operational efficiency and regulatory compliance, translating such approaches to full-scale facilities requires careful consideration of plant control architectures, instrumentation availability, and process uncertainties.

First, the controller in this study directly manipulates the oxygen transfer coefficients (K_{La}) in the simulation environment. This design choice simplifies the control interface during training and reduces computational overhead by bypassing the intermediate dissolved oxygen (DO) control loops commonly present in real plants. In practice, however, the controller would be deployed within a hierarchical control architecture. A practical implementation would involve a cascaded structure in which the RL agent provides optimal DO setpoints based on the current process state, while conventional PI controllers regulate aeration intensity to track these setpoints. Such a supervisory control configuration preserves the robustness of existing low-level controllers while allowing the RL agent to optimize plant-wide operational objectives.

Second, the state representation used in this study assumes the availability of ammonium measurements at multiple locations within the aeration tanks, as well as influent ammonium concentrations. While effluent ammonium monitoring is common in modern treatment facilities, distributed ammonium sensing throughout the biological reactors is not yet universally available. Two practical solutions may therefore be considered. One option is the installation of additional online sensors as part of plant digitalization and automation efforts. Another is the use of soft-sensing approaches, based on either data-driven or mechanistic models, to estimate unmeasured variables from available signals. The integration of such soft sensors with RL-based control represents a promising direction for enabling advanced control without extensive instrumentation upgrades.

A related issue concerns the reliability of these measurements. The robustness experiments in this study considered noisy and biased observations, but not complete sensor loss or sensor-dropout scenarios. This distinction is important for practical deployment, since the loss of a critical measurement may have a more severe impact than moderate measurement uncertainty. In the context of RL-based wastewater control, (Croll et al., 2023b) showed that severely corrupted NHx sensor readings can lead to loss of effluent compliance and emphasized the need for reliable sensors, diversified observations, and additional safety mechanisms around the RL controller. These findings suggest that fault detection, state estimation, or supervisory fallback strategies may be required for safe real-world implementation.

Beyond implementation considerations, the learned control policy also provides insight into the operational dynamics of the treatment process. As illustrated in Fig. 5, the RL controller adopts a demand-responsive aeration strategy, characterized by pronounced increases in aeration intensity during peak loading periods and extended periods of low aeration during lower influent loads. This behavior effectively aligns energy consumption with treatment demand, allowing the system to maintain nitrification performance while reducing unnecessary aeration. In parallel, the internal recycle ratio is dynamically adjusted to support denitrification when nitrate concentrations increase. These coordinated adjustments illustrate how the RL agent implicitly learns the coupling between nitrification, denitrification, and aeration energy consumption, enabling more adaptive control strategies than fixed setpoint approaches.

Finally, it should be noted that the controller was developed and evaluated using the BSM1 simulation benchmark. Although BSM1 provides a widely accepted and mechanistically grounded representation of activated sludge processes, it cannot fully capture all sources of uncertainty present in full-scale wastewater treatment plants. Factors such as actuator delays, seasonal biomass shifts, and unmodeled disturbances may affect real-world performance. These considerations highlight the importance of validating the proposed control strategy under realistic plant conditions, such as pilot-scale facilities or digital twin environments, before large-scale deployment.

6. Conclusions and future work

This study presents Adaptive Lagrangian Soft Actor-Critic (AL-SAC) as an effective constrained RL approach for the real-time control of activated-sludge wastewater treatment plants. By framing aeration and recycle operations within a Constrained Markov Decision Process and embedding effluent-quality requirements via adaptive Lagrange multipliers, the AL-SAC controller achieves substantial reductions in energy consumption with an average of 3 034.4 kWh/day, corresponding to a 23.2% reduction relative to the BSM1 baseline. The AL-SAC controller achieves these improvements while reliably respecting ammonium and total nitrogen discharge limits. Furthermore, the AL-SAC controller markedly reduces both the duration and magnitude of permit-threshold exceedances, demonstrating that advanced RL can reconcile stringent environmental regulations with operational efficiency. Overall, the results substantiate the potential of AL-SAC to bridge the divide between advanced learning-based control and the practical demands of industrial wastewater treatment.

The proposed learning framework operates by embedding discharge limits for ammonium nitrogen and total nitrogen directly into the objective of a constrained Markov decision process. This design allows new safety or quality constraints to be introduced simply by adding their cost functions and updating the associated dual variables during training. Consequently, the controller can be extended to accommodate additional effluent metrics without requiring any extra hyperparameter tuning.

From a computational perspective, the proposed controller is well suited for real-time deployment. The policy network consists of a lightweight multilayer perceptron with two hidden layers of 256 neurons each, resulting in a small number of parameters and low inference cost. Once training is completed offline, online operation only requires a forward pass through the actor network to compute the control action. Given that control decisions are applied every 15 minutes in the BSM1 environment, the inference time is negligible compared to the process dynamics and typical industrial control intervals. Consequently, the computational requirements of the proposed AL-SAC controller are compatible with standard industrial control hardware and do not pose a barrier for real-time deployment.

Future work will focus on validating the approach in pilot- and full-scale treatment plants under variable influent conditions and sensor uncertainties, while leveraging extensive historical process data for pre-training and warm-starting the controllers. We will also investigate the integration of plant-wide control strategies-linking unit processes through hierarchical or distributed RL architectures-to coordinate aeration, nutrient removal, and sludge handling for optimized overall performance. In parallel, we intend to expand the set of water-quality objectives to include BOD, TSS, and emerging micropollutants, and to develop robust RL variants that explicitly handle model mismatch, time-varying dynamics, and intermittent sensor failures. By combining data-driven pre-training, scalable constraint handling, and system-level coordination, this framework paves the way toward the next generation of intelligent, self-tuning, and sustainable wastewater treatment systems.

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

During the preparation of this work, the authors used OpenAI ChatGPT to proofread. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Omid Sobhani: Writing – original draft, Visualization, Validation, Conceptualization, Methodology, Investigation, Formal analysis, Software; **Thomas Huybrechts:** Writing – review & editing, Visualization, Validation, Conceptualization, Methodology, Investigation, Formal analysis; **Hamid Toliati:** Writing – review & editing, Methodology, Investigation, Conceptualization; **Cristian Camilo Gomez Cortes:** Resources, Conceptualization; **Kevin Mets:** Writing – review & editing, Methodology, Supervision; **Siegfried Mercelis:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability

This study used standard benchmark simulation and referred to it in the text. For specific data we provide the reference in the text.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Omid Sobhani reports financial support was provided by European Union. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been co-funded by the European Union under the Horizon Europe research and innovation programme through the DARROW project (Grant Agreement No. 101070080). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

The authors would like to thank Dr. Kimberly Solon (Ghent University) for the fruitful discussions that contributed to this work.

Appendix A. Observation-noise robustness scenarios

To evaluate robustness to imperfect measurements, observation noise was injected directly into the concentration-related channels of the agent input during the one-year test simulations. Let o_i denote the ith observation channel. For the selected channels, the noisy observation was generated according to Eq. (A.1):

$$\tilde{o}_i = o_i + b_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \tag{A.1}$$

where b_i is an additive bias term and ϵ_i is Gaussian noise. In the relative-noise mode used in this study, the standard deviation was scaled with the magnitude of the corresponding observation according to Eq. (A.2):

$$\sigma_i = \alpha_i |o_i|, \tag{A.2}$$

where α_i is the relative noise coefficient specified for each scenario. Noise was applied only to the concentration-related observation channels, while the time-of-day encoding features were kept noise-free. The exact values of α_i and b_i for all scenarios are listed in Table A.1 .

Table A.1
Observation-noise scenarios used in the robustness evaluation. Channel indices 2–7 correspond to influent SNH, reactor 3 SNH, reactor 4 SNH, reactor 5 SNH, effluent SNH deviation, and effluent TN deviation, respectively.

Scenario	Name	Relative noise std (channels 2–7)	Bias (channels 2–7)
A	Mild zero-bias noise	{0.10, 0.10, 0.10, 0.10, 0.10, 0.20}	{0, 0, 0, 0, 0, 0}
B	Moderate positive-bias shift	{0.20, 0.20, 0.20, 0.20, 0.20, 0.30}	{+0.10, +0.10, +0.20, +0.20, +0.10, +0.20}
C	Severe positive-bias shift	{0.30, 0.30, 0.30, 0.30, 0.30, 0.50}	{+0.20, +0.30, +0.50, +0.50, +0.20, +0.50}
D	Moderate mixed-bias shift	{0.20, 0.20, 0.20, 0.20, 0.20, 0.30}	{+0.10, −0.10, +0.20, −0.20, +0.10, −0.20}
E	Severe mixed-bias shift	{0.30, 0.30, 0.30, 0.30, 0.30, 0.50}	{+0.20, −0.30, +0.50, −0.50, +0.20, −0.50}

Table A.2
Performance of the trained controller under the considered observation-noise scenarios during the one-year evaluation.

Scenario	EQI	Energy	SNH exceed. [days]	SNH exceed. [%]	TN exceed. [days]	TN exceed. [%]
Noise-free baseline	5445.01	3049.94	25.08	6.89	0.57	0.16
A	5442.52	3057.29	25.10	6.90	0.55	0.15
B	5441.37	3057.44	25.01	6.87	0.50	0.14
C	5443.25	3058.10	25.22	6.93	0.50	0.14
D	5441.50	3058.41	24.92	6.85	0.49	0.13
E	5440.48	3061.56	24.89	6.84	0.60	0.17

References

Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. In D. Precup, & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 22–31). PMLR (*vol. 70*). Proceedings of Machine Learning Research.

Alali, Y., Harrou, F., & Sun, Y. (2023). Unlocking the potential of wastewater treatment: Machine learning based energy consumption prediction. *Water*, *15*(13). https://doi.org/10.3390/w15132349

Alex, J., Benedetti, L., Copp, J. B., Gernaey, K. V., Jeppsson, U., Nopens, I., Pons, M.-N., Rieger, L., Rosen, C., Steyer, J. P. et al. (2008). Benchmark simulation model no. 1 (BSM1): Technical Report TEIE-7229. Lund University. https://www2.iea.lth.se/publications/Reports/LTH-IEA-7229.pdf.

Altman, E. (1999). Constrained Markov decision processes. Routledge. https://doi.org/10.1201/9781315140223

Åmand, L., Olsson, G., & Carlsson, B. (2013). Aeration control–a review. *Water Science and Technology*, *67*(11), 2374–2398. https://doi.org/10.2166/wst.2013.139

Budzynski, G. (2024). Total ammonia aeration control (TAAC) theory - an innovative ammonia-based aeration controller. *AQUA - Water Infrastructure, Ecosystems and Society*, *73*(3), 396–406. https://doi.org/10.2166/aqua.2024.209

Chen, K., Wang, H., Valverde-Pérez, B., Zhai, S., Vezzaro, L., & Wang, A. (2021). Optimal control towards sustainable wastewater treatment plants based on multi-agent reinforcement learning. *Chemosphere*, *279*, 130498. https://doi.org/10.1016/j.chemosphere.2021.130498

Croll, H. C., Ikuma, K., Ong, S. K., & Sarkar, S. (2023a). Reinforcement learning applied to wastewater treatment process control optimization: Approaches, challenges, and path forward. *Critical Reviews in Environmental Science and Technology*, *53*(20), 1775–1794. https://doi.org/10.1080/10643389.2023.2183699

Croll, H. C., Ikuma, K., Ong, S. K., & Sarkar, S. (2023b). Systematic performance evaluation of reinforcement learning algorithms applied to wastewater treatment control optimization. *Environmental Science & Technology*, *57*(46), 18382–18390. https://doi.org/10.1021/acs.est.3c00353

Croll, H. C., Ikuma, K., Ong, S. K., & Sarkar, S. (2025). Reinforcement learning optimization of a water resource recovery facility: Evaluating the impact of reward function design on agent training, control optimization, and treatment risk. *Journal of Water Process Engineering*, *69*, 106658. https://doi.org/10.1016/j.jwpe.2024.106658

Dickin, S., Bayoumi, M., Giné, R., Andersson, K., & Jiménez, A. (2020). Sustainable sanitation and gaps in global climate policy and financing. *NPJ Clean Water*, *3*(1), 24. https://doi.org/10.1038/s41545-020-0072-8

Gernaey, K. V., Flores-Alsina, X., Rosen, C., Benedetti, L., & Jeppsson, U. (2011). Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. *Environmental Modelling & Software*, *26*(11), 1255–1267. https://doi.org/10.1016/j.envsoft.2011.06.001

Giorgi, G., Jiménez, B., & Novo, V. (2023). Convex optimization: Saddle points characterization and introduction to duality. In *Basic mathematical programming theory*, pp. 243–273). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-30324-1_8

Gu, Y., Li, Y., Yuan, F., & Yang, Q. (2023). Optimization and control strategies of aeration in WWTPs: A review. *Journal of Cleaner Production*, *418*, 138008. https://doi.org/10.1016/j.jclepro.2023.138008

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 1861–1870). PMLR (*vol. 80*). Proceedings of Machine Learning Research.

Han, M., Tian, Y., Zhang, L., Wang, J., & Pan, W. (2021). Reinforcement learning control of constrained dynamic systems with uniformly ultimate boundedness stability guarantee. *Automatica*, *129*, 109689. https://doi.org/10.1016/j.automatica.2021.109689

Henze, M., Gujer, W., Mino, T., & Van Loosedrecht, M. (2006). Activated sludge models ASM1, ASM2, ASM2d and ASM3. IWA publishing.

Holman, J. B., & Wareham, D. G. (2003). Oxidation–reduction potential as a monitoring tool in a low dissolved oxygen wastewater treatment process. *Journal of Environmental Engineering*, *129*(1), 52–58. https://doi.org/10.1061/(ASCE)0733-9372(2003)129:1(52)

Hu, F., Zhang, X., Lu, B., & Lin, Y. (2024). Real-time control of A2O process in wastewater treatment through fast deep reinforcement learning based on data-driven simulation model. *Water*, *16*(24). https://doi.org/10.3390/w16243710

Jamaludin, M., Tsai, Y.-C., Lin, H.-T., Huang, C.-Y., Choi, W., Chen, J.-G., & Sean, W.-Y. (2024). Modeling and control strategies for energy management in a wastewater center: A review on aeration. *Energies*, *17*(13). https://doi.org/10.3390/en17133162

Liu, Y., Halev, A., & Liu, X. (2021). Policy learning with constraints in model-free reinforcement learning: A survey. In Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 4508–4515). International Joint Conferences on Artificial Intelligence Organization. Survey Track. https://doi.org/10.24963/ijcai.2021/614

Miederer, J., Meier, L., Elhaus, N., Markthaler, S., & Karl, J. (2025). Energy management model for wastewater treatment plants. *Energy Reports*, *13*, 6349–6361. https://doi.org/10.1016/j.egyr.2025.05.045

Mohammadi, E., Ortiz-Arroyo, D., Hansen, A. A., Stokholm-Bjerregaard, M., Gros, S., Anand, A. S., & Durdevic, P. (2025). Application of soft actor-critic algorithms in optimizing wastewater treatment with time delays integration. *Expert Systems with Applications*, *277*, 127180. https://doi.org/10.1016/j.eswa.2025.127180

Muloiwa, M., Dinka, M. O., & Nyende-Byakika, S. (2022). Modelling and optimization of energy consumption in the activated sludge biological aeration unit. *Water Practice and Technology*, *18*(1), 140–158. https://doi.org/10.2166/wpt.2022.154

Nourani, V., Asghari, P., & Sharghi, E. (2021). Artificial intelligence based ensemble modeling of wastewater treatment plant using jittered data. *Journal of Cleaner Production*, *291*, 125772. https://doi.org/10.1016/j.jclepro.2020.125772

Hernandez-del Olmo, F., & Gaudioso, E. (2011). Reinforcement learning techniques for the control of wastewater treatment plants. In J. M. Ferrández, J. R. Álvarez Sánchez, F. de la Paz, & F. J. Toledo (Eds.), *New challenges on bioinspired applications* (pp. 215–222). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21326-7_24

Hernández-del Olmo, F., Gaudioso, E., Duro, N., & Dormido, R. (2019). Machine learning weather soft-sensor for advanced control of wastewater treatment plants. *Sensors*, *19*(14). https://doi.org/10.3390/s19143139

Parsa, Z., Dhib, R., & Mehrvar, M. (2024). Dynamic modelling, process control, and monitoring of selected biological and advanced oxidation processes for wastewater treatment: A review of recent developments. *Bioengineering*, *11*(2), 189. https://doi.org/10.3390/bioengineering11020189

Paternain, S., Chamon, L. F. O., Calvo-Fullana, M., & Ribeiro, A. (2019). Constrained reinforcement learning has zero duality gap. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 7555–7565). Red Hook, NY, USA: Curran Associates Inc.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, *22*(268), 1–8. http://jmlr.org/papers/v22/20-1364.html.

Saagi, R. (2017). Generating wastewater treatment plant influent data for realistic evaluation of future scenarios – case study for WWTPs in stockholm, Sweden. Technical Report Division of Industrial Electrical Engineering and Automation, Lund University.

Sandino, J., Crawford, G., Fillmore, L., & Kinnear, D. (2010). Energy efficiency in wastewater treatment in north america: A WERF compendium of best practices and case studies of novel approaches. In *Weftec 2010* (pp. 3333–3346). Water Environment Federation.

Schraa, O., Rieger, L., Alex, J., & Miletić, I. (2019). Ammonia-based aeration control with optimal SRT control: Improved performance and lower energy consumption. *Water Science and Technology*, *79*(1), 63–72. https://doi.org/10.2166/wst.2019.032

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. Cambridge, MA, USA: A Bradford Book.

Takács, I., Patry, G. G., & Nolasco, D. (1991). A dynamic model of the clarification-thickening process. *Water Research*, *25*(10), 1263–1271. https://doi.org/10.1016/0043-1354(91)90066-Y

Towers, M., Kwiatkowski, A., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Andreas, K., Krimmel, M., Arjun, K. G., De Lazcano Perez-Vicente, R., Terry, J. K., Pierré, A., Schulhoff, S. V., Tai, J. J., Tan, H., & Younis, O. G. (2025). Gymnasium: A standard interface for reinforcement learning environments. In *The thirty-ninth annual conference on neural information processing systems datasets and benchmarks track*. https://openreview.net/forum?id=qPMLvJxtPK.

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. *Science of The Total Environment*, *784*, 147138. https://doi.org/10.1016/j.scitotenv.2021.147138

Zhao, Q., Zhang, Y., & Li, X. (2022). Safe reinforcement learning for dynamical systems using barrier certificates. *Connection Science*, *34*(1), 2822–2844. https://doi.org/10.1080/09540091.2022.2151567

Zhou, P. (2019). Development of data-driven models for influent prediction at wastewater treatment plants. Master's thesis. McMaster University. http://hdl.handle.net/11375/24368.