ISSN 2723-6471

Deep Deterministic Policy Gradient for Simulation-Based Control of Water Quality in Nano Aquatic Systems

Iwan Fitrianto Rahmad[1], Syahril Efendi[2,*], Poltak Sihombing[3], Henny Febriana Harumy[4]

[1]*Doctoral program of Computer Science, Universitas Sumatera Utara, Medan 20155, Indonesia*

[2,3,4]*Department of Computer Science, Universitas Sumatera Utara, Medan 20155, Indonesia*

(Received: February 10, 2026; Revised: April 5, 2026; Accepted: June 25, 2026; Available online: July 20, 2026)

Abstract

This study proposes a simulation-based Deep Deterministic Policy Gradient (DDPG) framework for water-quality control in nano aquatic systems. Nano tanks are highly sensitive to small disturbances because their limited water volume reduces buffering capacity and causes rapid changes in dissolved oxygen, ammonia, pH, temperature, biological oxygen demand, and chemical oxygen demand. To represent these coupled dynamics, a nonlinear simulation model adapted from the Continuously Stirred Tank Reactor concept is developed and implemented as an OpenAI Gym-compatible reinforcement learning environment. The DDPG agent learns continuous control actions related to aeration, feeding, and filtration through repeated interaction with the simulated nano-tank environment. The proposed nonlinear CSTR-DDPG framework is evaluated against a linear-model DDPG baseline using RMSE, cumulative reward, and closed-loop control performance. Simulation results show that the nonlinear model reduced RMSE by 45.2% for dissolved oxygen, 64.0% for ammonia, and 61.3% for pH compared with the linear baseline. The DDPG agent also achieved a 41.7% higher cumulative reward under the same reward structure. These findings indicate that nonlinear simulation can provide a more informative training environment for DDPG-based water-quality control. However, the present study remains limited to simulation-based evaluation, and physical validation using calibrated sensors, actuators, communication-delay analysis, and real nano-tank experiments is required in future work.

Keywords: CSTR Model, Deep Deterministic Policy Gradient, Nano Aquatic System, Nonlinear Simulation, Reinforcement Learning, Water-Quality Control

1. Introduction

Aquatic micro-ecosystems such as nano tanks are small volume aquatic environments that are increasingly used for educational, experimental, and small-scale aquaculture purposes [1], [2]. A nano tank generally refers to a small aquatic system with a water volume below 30 L, where limited water volume makes the ecosystem highly sensitive to small environmental and operational disturbances. In such systems, minor changes in feeding intensity, aeration, filtration, or ambient temperature may rapidly affect dissolved oxygen (DO), ammonia concentration, pH, temperature, biological oxygen demand (BOD), and chemical oxygen demand (COD) [3], [4], [5]. Because the buffering capacity of a nano tank is much lower than that of larger aquaculture systems, water-quality instability may occur within a relatively short period and may negatively affect aquatic organisms.

The main challenge in controlling water quality in nano aquatic systems lies in the nonlinear and interdependent relationship among physicochemical and biochemical parameters [6]. Excessive feeding, for example, may increase organic load and ammonia concentration. Increased ammonia may then stimulate nitrification processes, consume dissolved oxygen, and influence pH stability [7]. Temperature variation can further affect oxygen solubility and biochemical reaction rates, thereby modifying the interaction among DO, ammonia, and pH [8]. These interactions show that water quality variables in nano tanks cannot be treated as isolated parameters. Instead, they form a coupled dynamic system in which a disturbance in one variable may propagate to other variables and affect the overall stability of the aquatic environment [9], [10]. Conventional control approaches, including proportional-integral-derivative (PID), programmable logic controller (PLC)-based systems, and fuzzy rule-based control, have been widely used in

***Corresponding author: Syahril Efendi (syahril1@usu.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i3.1447

the water management and aquaculture systems because of their simplicity, interpretability, and relatively low computational requirements [11]. However, these methods are inadequate for nano-scale systems, where physicochemical and biological processes exhibit strong nonlinearity, rapid time-scale interactions, and tight interdependence among parameters [12]. Since this study does not experimentally compare the proposed method with tuned PID or fuzzy controllers, it does not claim direct superiority over these methods. Instead, the present work focuses on developing and evaluating a simulation-based reinforcement learning framework for adaptive water quality control under nonlinear nano tank dynamics.

Reinforcement learning (RL) is suitable for sequential decision-making problems because an agent learns to select actions through repeated interaction with an environment and receives feedback in the form of reward or penalty [13]. In water-quality control, RL can be used to determine adaptive operational actions, such as adjusting aeration, limiting feeding input, or modifying filtration intensity based on the current condition of the aquatic system [14]. Among RL algorithms, Deep Deterministic Policy Gradient (DDPG) is particularly relevant for continuous-control problems because it combines an actor network for generating deterministic continuous actions and a critic network for evaluating the long-term value of state-action pairs [15], [16]. This characteristic makes DDPG appropriate for nano aquatic systems, where actuator decisions may require gradual adjustment rather than only discrete on/off actions.

Nevertheless, the performance of an RL controller depends strongly on the quality of the simulation environment used during training. If the environment is too simplified, linearized, or weakly specified, the learned policy may become biased toward assumptions that do not sufficiently represent the actual nonlinear water-quality dynamics [17], [18]. This issue is particularly important in nano aquatic systems because the interactions among DO, ammonia, pH, temperature, BOD, and COD are sensitive to small changes in feeding, aeration, filtration, and ambient conditions. Therefore, a nonlinear simulation environment is needed to train and evaluate DDPG policies before any physical implementation is attempted.

To address this issue, this study proposes a simulation-based DDPG control framework for water-quality management in nano aquatic systems. The framework uses a nonlinear simulation model adapted from the Continuously Stirred Tank Reactor (CSTR) concept to represent the coupled dynamics of key water-quality variables. The state space consists of DO, ammonia, pH, temperature, BOD, and COD, while the action space includes controllable decisions related to aeration, feeding, and filtration. The simulation environment is implemented using an OpenAI Gym-compatible structure, allowing the DDPG agent to interact with the environment through the standard state-action-reward-next-state mechanism.

In this framework, DO, ammonia, and pH are treated as the primary controlled indicators because they directly reflect short-term water-quality safety. Temperature, BOD, and COD are included as supporting state variables because they influence the transition dynamics of the primary indicators. Temperature affects oxygen solubility and reaction rates, BOD reflects organic oxygen demand, and COD represents the broader oxidizable load in the system. Including these variables allows the DDPG agent to evaluate water-quality conditions more comprehensively and to learn control actions based on the multivariable state of the nano aquatic environment.

The research gap addressed in this study is the lack of a clearly formulated nonlinear simulation-based DDPG framework for water-quality control in nano aquatic systems. Previous studies have explored machine learning, reinforcement learning, water-quality monitoring, and aquaculture control in broader contexts [17], [19], [20], [21]. However, fewer studies have focused specifically on nano aquatic systems, where small water volume creates rapid fluctuations and strong coupling among water-quality variables. Moreover, many existing approaches emphasize monitoring or prediction, while fewer provide a structured simulation environment in which nonlinear water-quality dynamics, continuous control actions, reward formulation, and DDPG learning are evaluated together.

Therefore, the objectives of this study are fourfold: first, to formulate a nonlinear CSTR-based simulation model for representing nano aquatic water-quality dynamics; second, to construct an OpenAI Gym-compatible reinforcement learning environment using the state-action-reward-next-state structure; third, to train a DDPG agent for simulation-based control of aeration, feeding, and filtration actions; and fourth, to evaluate the proposed framework using prediction error, cumulative reward, and safety-oriented control behavior.

 ISSN 2723-6471

The main contributions of this study are threefold. First, it presents a nonlinear simulation model for representing the coupled dynamics of water-quality parameters in nano aquatic systems. Second, it integrates the model with a DDPG-compatible reinforcement learning environment for continuous control learning. Third, it evaluates the simulation-based control performance of the DDPG agent against a linear model-based baseline. The scope of this study is limited to simulation-based evaluation; physical implementation using calibrated sensors, actuators, communication-delay analysis, and closed-loop hardware testing remains a direction for future work.

2. Methodology

This study develops a simulation-based control framework for maintaining water-quality stability in nano aquatic systems using DDPG. The proposed framework consists of three main components: (i) a nonlinear CSTR-based simulation model representing the coupled water-quality dynamics of the nano tank, (ii) an OpenAI Gym-compatible reinforcement learning environment based on the state–action–reward–next-state structure, and (iii) a DDPG agent for learning continuous control actions. The overall research workflow is illustrated in figure 1.

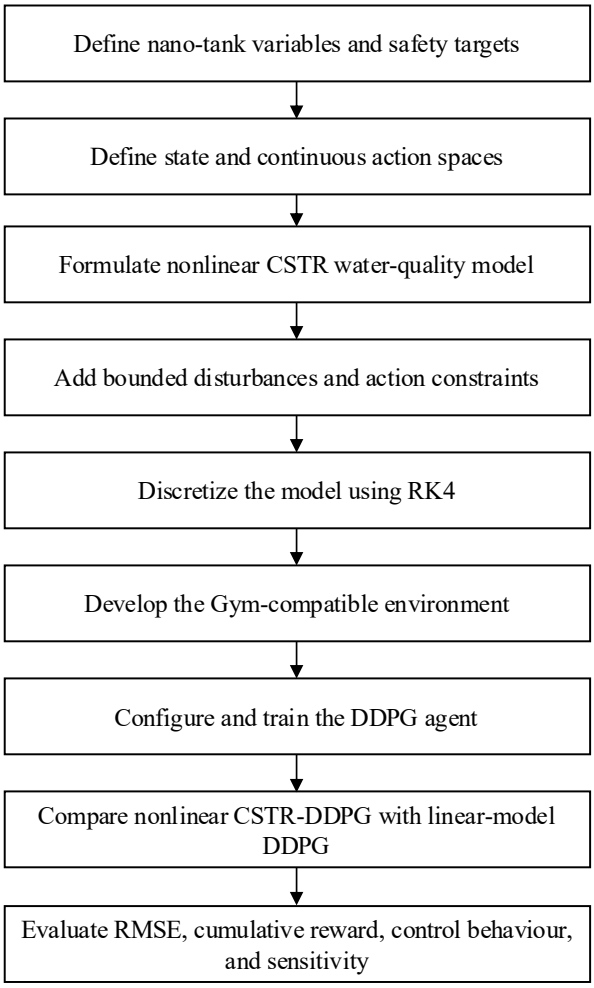

Figure 1. Research workflow of the proposed nonlinear CSTR-DDPG framework.

The methodological workflow consists of five stages. First, water-quality variables and control variables are defined based on the nano aquatic system. Second, a nonlinear simulation model is formulated to describe the interaction among dissolved oxygen, ammonia, pH, temperature, BOD, and COD. Third, the simulation model is implemented as a reinforcement learning environment. Fourth, the DDPG agent is trained through repeated interaction with the environment. Fifth, the control performance is evaluated using RMSE, cumulative reward, and safety-oriented control behavior. The evaluation in this study is limited to simulation-based experiments. Physical closed-loop implementation using calibrated sensors, actuators, communication-delay analysis, and real nano-tank hardware testing is not included in the present scope.

The process begins with defining the nano-tank state variables, control actions, safety limits, and reference values. A nonlinear CSTR-based model is then formulated and discretized using the fourth-order Runge–Kutta method. The resulting model is implemented as a gym-compatible reinforcement learning environment. Subsequently, the DDPG agent is configured and trained to determine continuous aeration, feeding, and filtration actions. Finally, the trained policy is compared with the linear-model DDPG baseline using RMSE, cumulative reward, closed-loop control behaviour, and hyperparameter sensitivity.

2.1. State and Action Representation

The system state is defined as a continuous vector representing the key water-quality parameters:

$$s_t = [DO_t, NH_{3,t}, pH_t, T_t, BOD_t, COD_t] \quad (1)$$

DO_t denotes dissolved oxygen, $NH_{3,t}$ denotes ammonia concentration, pH_t denotes acidity level, T_t denotes water temperature, BOD_t denotes biological oxygen demand, and COD_t denotes chemical oxygen demand at time step t.

In this formulation, DO, ammonia, and pH are treated as the primary controlled indicators because they directly reflect short-term water-quality safety in the nano tank. Temperature, BOD, and COD are included as supporting state variables because they affect the transition dynamics of the primary indicators through oxygen solubility, organic oxygen demand, biochemical reaction rates, and oxidizable pollutant load.

The control action vector is defined as:

$$a_t = [u_{aer,t}, u_{feed,t}, u_{filter,t}] \quad (2)$$

$u_{aer,t}$ is the aeration control, $u_{feed,t}$ is the feeding control, $u_{filter,t}$ is the filtration control at time step t. All control variables are modeled as continuous actions to enable fine-grained adjustment by the RL agent.

The action bounds are explicitly defined to ensure that the DDPG agent operates within physically meaningful and safety-constrained limits. The control variables, allowable ranges, and physical interpretations used in the simulation are summarized in table 1.

Table 1. Control Action Space and Safety Bounds

Control action	Symbol	Range	Interpretation
Aeration intensity	$u_{aer,t}$	[0,1]	normalized aerator intensity
Feeding input	$u_{feed,t}$	[0,0.1]g	feeding amount per control interval
Filtration intensity	$u_{filter,t}$	[0,1]	normalized filtration intensity

All actions are clipped to their allowable ranges before being applied to the simulation model. Feeding is interpreted as a feeding-scheduling control variable rather than an unlimited stabilizing actuator. Therefore, the safety layer prevents feeding from increasing when ammonia or organic load approaches unsafe conditions.

$$\tilde{a}_t = \text{clip}(a_t, a_{\min}, a_{\max}) \quad (3)$$

Were

$$a_{\min} = [0,0,0], \quad a_{\max} = [1,0.1,1] \quad (4)$$

Thus,

$$\tilde{a}_t = [\tilde{u}_{aer,t}, \tilde{u}_{feed,t}, \tilde{u}_{filter,t}] \quad (5)$$

is the safety-bounded action vector applied to the nonlinear simulation model.

Action clipping was used as a simple safety layer to prevent the DDPG policy from applying physically infeasible or unsafe control actions. However, clipping may bias the learned policy when the actor frequently proposes actions outside the feasible range. In this study, clipping was retained to ensure safe simulation behaviour, while more formal constrained reinforcement learning methods are left for future work.

2.2. Nonlinear CSTR-Based Simulation Model

The nano aquatic system is modelled as a nonlinear dynamic system adapted from the CSTR concept. The CSTR assumption is used as a computational approximation in which the small water volume is treated as a well-mixed compartment, allowing the interaction among water-quality variables to be represented through coupled differential equations [22].

In this study, the CSTR formulation is used as an analytical and simulation-based approximation, not as a physical reactor equipped with a mechanical stirrer. The complete-mixing assumption is justified by the small tank volume and by the continuous circulation generated by aeration bubbles and filtration flow. These mechanisms promote water movement and reduce spatial concentration gradients in the nano tank. Therefore, each state variable is interpreted as the average tank condition at a given control interval. Nevertheless, this assumption may be less valid for tanks with stagnant regions, weak aeration, irregular layouts, or heterogeneous biological distribution. For this reason, the proposed model should be interpreted as a simplified mechanistic representation for simulation-based reinforcement learning, rather than a complete spatial model of all nano-tank configurations.

The nonlinear equations were formulated using simplified mass-balance and water-quality interaction principles. The oxygen dynamics include aeration-dependent oxygen transfer, respiration-related oxygen loss, nitrification-related oxygen consumption, and organic-load oxygen demand. The ammonia dynamics include feeding input, nitrification removal, and filtration-related reduction. The pH dynamics approximate the acidifying effect of nitrification and the stabilizing effect of buffering. Temperature, BOD, and COD are included as supporting state variables because they influence oxygen solubility, biochemical reaction rates, and pollutant-load dynamics.

The nonlinear formulation was developed to represent the sensitivity of small-volume aquatic systems, where minor changes in feeding, temperature, or aeration may produce disproportionate changes in DO, ammonia, and pH. In particular, the oxygen transfer coefficient depends on aeration and temperature, respiration is temperature-dependent, nitrification couples ammonia reduction with oxygen consumption, and BOD/COD terms represent the organic load generated by feeding and waste accumulation. Thus, the nonlinear model is intended to provide a more expressive training environment than a purely linear transition model.

The nonlinear model was evaluated using a 30-day nano-tank dataset from a 25 L experimental setting under controlled disturbances, including feeding variation and ambient-temperature variation. The dataset was divided using an 80:20 hold-out validation scheme, where 80% of the data were used for model fitting and 20% were used for validation. This validation was performed to assess whether the nonlinear simulation model could approximate observed water-quality trajectories before being integrated into the DDPG environment. However, the DDPG policy itself has not yet been deployed as a real-time physical closed-loop controller; therefore, the control results remain simulation-based. The general state transition is expressed as:

$$s_{t+1} = F(s_t, a_t, \xi_t) \tag{6}$$

$F(\cdot)$ represents the nonlinear transition function, s_t is the current state, a_t is the control action, and ξ_t represents external disturbance.

The external disturbance term ξ_t represents unmodelled fluctuations in the nano tank, such as small feeding irregularities, ambient-temperature variation, sensor perturbation, and biological variability. In the simulation, the disturbance is added to the state-transition equation as a bounded stochastic perturbation. Its purpose is to prevent the agent from learning only deterministic trajectories and to improve robustness against uncertain operating conditions. During evaluation, the same disturbance-generation rule is applied consistently to both the nonlinear CSTR-DDPG framework and the linear-model DDPG baseline. The DO dynamics are represented as:

$$\frac{dDO}{dt} = K_{La}\left(u_{aer,t}, T\right)(DO^* - DO) - k_{resp}(T)DO - k_{nit}(NH_3, DO, T)DO - k_{BOD}(BOD, T)DO + \xi_{DO} \tag{7}$$

$K_{La}\left(u_{aer,t}, T\right)$is the aeration-dependent oxygen transfer coefficient, DO^* is the saturated dissolved oxygen level, $k_{resp}(T)$ represents temperature-dependent respiration loss, $k_{nit}(NH_3, DO, T)$ represents oxygen consumption due to

 ISSN 2723-6471

nitrification, $k_{BOD}(BOD,T)$ represents oxygen demand due to organic load, and ξ_{DO} denotes disturbance in DO dynamics. The main nonlinear coefficient functions are specified as follows:

$$\begin{gathered} K_{La}\left(u_{aer,t},T_t\right) = K_{La,0} + K_{La,max}u_{aer,t}\theta_{aer}^{(T_t-25)} \\ k_{resp}(T_t) = k_{resp,0}\theta_{resp}^{(T_t-25)} \\ k_{nit}\left(NH_{3,\mathrm{t}},DO_t,T_t\right) = k_{nit,0}\left(\frac{NH_{3,\mathrm{t}}}{K_{NH_3}+NH_{3,\mathrm{t}}}\right)\left(\frac{DO_t}{K_{\mathrm{DO}}+DO_t}\right)\theta_{nit}^{(T_t-25)} \\ k_{BOD}(BOD_t,T_t) = k_{BOD,0}\left(\frac{BOD_t}{K_{\mathrm{BOD}}+BOD_t}\right)\theta_{BOD}^{(T_t-25)} \end{gathered} \tag{8}$$

These expressions allow the oxygen-transfer, respiration, nitrification, and organic-load effects to vary with the current state of the nano tank instead of remaining constant as in a linear transition model.

The ammonia dynamics are expressed as:

$$\frac{dNH_3}{dt} = g_{feed}\left(u_{feed}\right) - k_{nit}(NH_3,DO,T)NH_3 - k_{filter}\left(u_{filter}\right)NH_3 + \xi_{NH_3} \tag{8}$$

$g_{feed}\left(u_{feed}\right)$ represents ammonia contribution from feeding input, $k_{nit}(NH_3,DO,T)$ represents ammonia removal through nitrification, and $k_{filter}\left(u_{filter}\right)$ represents filtration-related ammonia reduction. The feeding and filtration terms are defined as:

$$\begin{gathered} g_{feed}\left(u_{feed,t}\right) = \eta_{feed}u_{feed,t} \\ k_{filter}\left(u_{filter,t}\right) = k_{filter,0}u_{filter,t} \end{gathered} \tag{9}$$

η_{feed} is the ammonia-load conversion coefficient from feeding input and $k_{filter,0}$ is the maximum filtration-related removal coefficient.

The pH dynamics are represented as:

$$\frac{dpH}{dt} = -\alpha_{nit}k_{nit}(NH_3,DO,T) + \beta_{pH}\left(pH_{ref} - pH\right) + \xi_{pH} \tag{10}$$

α_{nit} represents the effect of nitrification on pH reduction, pH_{ref} is the reference pH value, and β_{pH} is the buffering coefficient. In the simulation setting, the pH reference is set near the neutral operating condition used in the reward function.

The temperature dynamics are expressed as:

$$\frac{dT}{dt} = \lambda_T(T_{amb} - T) + \xi_T \tag{11}$$

T_{amb} denotes ambient temperature and λ_T represents the heat-exchange coefficient between the water and surrounding environment.

The BOD and COD dynamics are represented as:

$$\begin{gathered} \frac{dBOD}{dt} = g_{BOD}\left(u_{feed,t}\right) - k_{deg}(BOD_t,DO_t,T_t)BOD_t - k_{filter,BOD}\left(u_{filter,t}\right)BOD_t + \xi_{BOD} \\ \frac{dCOD}{dt} = g_{COD}\left(u_{feed,t}\right) - k_{COD}(COD_t,T_t)COD_t - k_{filter,COD}\left(u_{filter,t}\right)COD_t + \xi_{COD} \end{gathered} \tag{12}$$

$g_{BOD}\left(u_{feed}\right)$ and $g_{COD}\left(u_{feed}\right)$ represent organic-load contribution from feeding, while the degradation and filtration terms represent the reduction of organic and oxidizable loads.

ISSN 2723-6471

To improve reproducibility, the main nonlinear model parameters, units, values or ranges, and implementation assumptions are summarized in table 2. The values are reported as simulation settings or calibrated ranges used to define stable nano-tank dynamics.

Table 2. Main nonlinear model parameters and simulation settings

Parameter	Description	Unit	Value/range used	Source/assumption
V	Nano tank volume	L	25	Experimental nano-tank setting
Δt	Control interval	min	6	Sensor/control sampling interval
DO^*	Saturated dissolved oxygen level	mg/L	8.0	Safe operating reference
pH_{ref}	Reference pH value	pH unit	7.2	Reward-function target
$K_{La,0}$, $K_{La,max}$	Baseline and maximum aeration oxygen transfer	interval^{-1}	0.03; 0.25	Simulation baseline & Aeration-response assumption
θ_{aer}, θ_{resp}, θ_{nit}, θ_{BOD}	Temperature correction factors	dimensionless	1.02; 1.03; 1.04; 1.03	Temperature-response assumption
$k_{resp,0}$	Baseline respiration loss coefficient	interval^{-1}	0.005–0.030	Calibrated simulation range
$k_{nit,0}$	Maximum nitrification coefficient	interval^{-1}	0.010–0.080	Calibrated simulation range
K_{NH_3}, K_{DO}	Half-saturation constants for nitrification	mg/L	0.05; 2.0	Nitrification-response assumption
$k_{BOD,0}$	Baseline BOD-related oxygen demand coefficient	interval^{-1}	0.005–0.050	Organic-load assumption
K_{BOD}	Half-saturation constant for BOD effect	mg/L	5.0	Organic-load assumption
η_{feed}	Ammonia-load conversion from feeding input	mg/L per interval	0.00–0.08	Feeding-response assumption
$k_{filter,0}$	Maximum ammonia removal by filtration	interval^{-1}	0.12	Filtration-response assumption
α_{nit}, β_{pH}	pH nitrification and buffering coefficients	pH unit per interval; interval^{-1}	0.22; 0.07	Calibrated pH-response & buffering coefficient
λ_T	Ambient heat-exchange coefficient	interval^{-1}	0.01–0.05	Ambient-temperature response assumption
k_{deg}, k_{COD}	BOD and COD degradation coefficients	interval^{-1}	0.005–0.030; 0.003–0.020	Calibrated simulation range
ξ_{DO}	DO disturbance	mg/L	$\mathcal{N}(0, 0.05^2)$, clipped to ±0.10	Bounded stochastic disturbance
ξ_{NH_3}	Ammonia disturbance	mg/L	$\mathcal{N}(0, 0.002^2)$, clipped to ±0.005	Bounded stochastic disturbance
ξ_{pH}	pH disturbance	pH unit	$\mathcal{N}(0, 0.01^2)$, clipped to ±0.03	Bounded stochastic disturbance
ξ_t	Temperature disturbance	°C	$\mathcal{N}(0, 0.10^2)$, clipped to ±0.20	Ambient disturbance
ξ_{BOD}, ξ_{COD}	Organic-load disturbance	mg/L	±2% of current value	Bounded stochastic disturbance

Nitrite is not included as a direct state variable in the DDPG observation vector. Its effect is represented through the lumped nitrification term $k_{nit}(NH_3, DO, T)$, thereby avoiding an unobserved state variable in the reinforcement learning formulation. This simplification reduces the dimensionality of the control problem, but it also limits the

model's ability to explicitly describe intermediate nitrification dynamics. Therefore, future work should extend the state representation to include nitrite and nitrate when reliable sensor data are available.

2.3. Numerical Discretization and Environment Implementation

The continuous-time system is discretized using the fourth-order Runge-Kutta method. Let $f(s_t, a_t)$ denote the right-hand side of the nonlinear dynamic model. The RK4 update is defined as:

$$\begin{aligned} k_1 &= \Delta t f(s_t, a_t) \\ k_2 &= \Delta t f\left(s_t + \frac{1}{2}k_1, a_t\right) \\ k_3 &= \Delta t f\left(s_t + \frac{1}{2}k_2, a_t\right) \\ k_4 &= \Delta t f(s_t + k_3, a_t) \\ s_{t+1} &= s_t + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \end{aligned} \tag{13}$$

The simulation time step is set to $\Delta t = 6$ minutes, following the control interval used in the nano-tank simulation setting. The episode is terminated when the system reaches a critical unsafe condition, specifically when $DO < 2$ mg/L or $NH_3 > 1$ mg/L. The simulation model is implemented as a custom OpenAI Gym-compatible environment. The environment follows two main functions. The reset() function initializes the state at the beginning of each episode, while the step(action) function applies the action, updates the state using RK4 integration, computes the reward, checks safety termination, and returns the tuple:

$$(s_{t+1}, R_t, done, info) \tag{14}$$

Algorithm 1. Nonlinear CSTR-Gym Environment Step

Input: current state s_t, action a_t, time step Δt

Output: next state s_{t+1}, reward R_t, termination flag done, information info

1. Receive action a_t from the DDPG actor network.
2. Clip the action into the allowable range:
   a_t ← clip(a_t, a_min, a_max)
3. Extract control variables:
   u_aer,t ← aeration intensity
   u_feed,t ← feeding input
   u_filter,t ← filtration intensity
4. Apply the action to the nonlinear CSTR-based simulation model.
5. Update the state using RK4 integration.
6. Update the state:
   s_{t+1} = s_t + (1/6)(k_1 + 2k_2 + 2k_3 + k_4)
7. Compute reward R_t using the water-quality reward function.
8. Check unsafe termination:
   if DO_t < 2 or NH3_t > 1:
      done ← True
   else:
      done ← False
9. Return (s_{t+1}, R_t, done, info).

This structure allows the DDPG agent to interact with the nano aquatic simulation environment in a standard reinforcement learning workflow.

2.4. Reward function

The reward function is explicitly formulated to penalize unsafe water-quality deviations while encouraging stable and smooth control actions. The total reward at time step t is defined as:

$$R_t = w_{DO}R_{DO,t} + w_{NH_3}R_{NH_3,t} + w_{pH}R_{pH,t} + w_uR_{u,t} + w_sR_{safe,t} \quad (15)$$

$R_{DO,t}$, $R_{NH_3,t}$, and $R_{pH,t}$ represent the reward components for the primary controlled variables, $R_{u,t}$ penalizes excessive action variation, and $R_{safe,t}$ penalizes unsafe states.

The dissolved oxygen reward is defined as:

$$R_{DO,t} = \begin{cases} 1, & 5 \le DO_t \le 8 \\ -\dfrac{|DO_t - 6.5|}{6.5}, & \text{otherwise} \end{cases} \quad (16)$$

The ammonia reward is defined as:

$$R_{NH_3,t} = \begin{cases} 1, & NH_{3,t} < 0.02 \\ -1, & NH_{3,t} > 0.5 \\ -NH_{3,t}, & \text{otherwise} \end{cases} \quad (17)$$

The pH reward is defined as:

$$R_{pH,t} = \begin{cases} 1, & 6.5 \le pH_t \le 8.0 \\ -\dfrac{|pH_t - 7.2|}{7.2}, & \text{otherwise} \end{cases} \quad (18)$$

The smooth-control penalty is defined as:

$$R_{u,t} = -\|a_t - a_{t-1}\|_2^2 \quad (19)$$

and the safety penalty is defined as:

$$R_{safe,t} = \begin{cases} -1, & DO_t < 2 \text{ or } NH_{3,t} > 1 \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

The reward weights are set as:

$$w_{DO} = 0.25, w_{NH_3} = 0.30, w_{pH} = 0.20, w_u = 0.10, w_s = 0.15 \quad (21)$$

The reward weights were selected based on safety-oriented control priorities and preliminary simulation trials. Ammonia was assigned the highest weight because ammonia accumulation is a critical toxicity indicator in small-volume aquatic systems. Dissolved oxygen and pH were also given high weights because they directly reflect short-term water-quality safety. The action-smoothness term was assigned a lower weight to discourage abrupt actuator changes without dominating the main water-quality objectives. The safety penalty was included to discourage critical unsafe states. These weights were not obtained through automatic optimization; therefore, the effect of reward-weight selection is acknowledged as a limitation and should be further examined through systematic sensitivity analysis in future work.

2.5. DDPG-based training procedure

The control problem is formulated as a continuous-state and continuous-action Markov Decision Process (MDP) [23], [24]. At each step, the agent observes the state s_t, selects an action a_t, receives a reward R_t, and observes the next state s_{t+1}. DDPG is employed because it is suitable for continuous control problems [15], [25]. The algorithm uses an actor, critic structure. The actor network $\pi_\theta(s)$ maps the current state to a deterministic action, while the critic network $Q_\phi(s,a)$ estimates the value of the state-action pair. The target value for critic learning is computed as:

$$y_t = R_t + \gamma(1 - d_t)Q_{\phi'}\left(s_{t+1}, \pi_{\theta'}(s_{t+1})\right) \quad (22)$$

γ is the discount factor, d_t is the termination indicator, and $Q_{\phi'}$ and $\pi_{\theta'}$ are the target critic and target actor networks.

The critic network is updated by minimizing:

$$L(\phi) = \mathbb{E}\left[\left(Q_\phi(s_t, a_t) - y_t\right)^2\right] \tag{23}$$

The actor network is updated using the deterministic policy gradient:

$$\nabla_\theta J \approx \frac{1}{|B|}\sum_{s\in B} \nabla_a Q_\phi(s,a)\,\big|_{a=\pi_\theta(s)}\, \nabla_\theta \pi_\theta(s) \tag{24}$$

Target networks are updated using soft update:

$$\begin{aligned} \phi' &\leftarrow (1-\tau)\phi' + \tau\phi \\ \theta' &\leftarrow (1-\tau)\theta' + \tau\theta \end{aligned} \tag{25}$$

The DDPG training procedure followed the standard actor–critic workflow. At each time step, the actor selected a continuous action with Gaussian exploration noise, the action was clipped to the allowable bounds, and the environment returned the next state, reward, and termination flag. The transition $(s_t, a_t, R_t, s_{t+1}, d_t)$ was stored in the replay buffer. Mini batches were then sampled to update the critic using the Bellman target and to update the actor using the deterministic policy gradient. Target networks were updated using the soft-update coefficient τ.

To improve reproducibility, the DDPG architecture, optimization settings, exploration strategy, and computational configuration are summarized in table 3. The same training configuration was used for both the nonlinear CSTR-DDPG framework and the linear-model DDPG baseline.

Table 3. DDPG architecture and training configuration

Component	Setting
Algorithm	DDPG
State dimension	6
Action dimension	3
Actor network & Critic network	MLP, 2 hidden layers, 256 neurons each, ReLU activation
Optimizer	Adam
Actor learning rate	1×10^{-4}
Critic learning rate	5×10^{-4}
Replay buffer size	100,000 transitions
Learning starts	1,000 transitions
Batch size	128
Discount factor (γ)	0.99
Soft update coefficient (τ)	0.005
Exploration strategy	Gaussian action noise, ($\sigma = 0.05$)
State normalization	Min-max normalization
Action constraint	Clipping to predefined safety bounds
Training length	200,000 timesteps
Evaluation policy	Deterministic actor without exploration noise
Software/hardware	Python-based Gym environment, 16 GB RAM/cloud environment

The actor and critic networks were implemented as multilayer perceptrons with two hidden layers of 256 neurons each and ReLU activation functions. Adam optimization was used for both networks. Gaussian action noise with $\sigma = 0.05$ was used during training, while deterministic actor output was used during evaluation. No adaptive noise decay was applied in this study.

2.6. Baseline and performance evaluation

To avoid ambiguity in the comparison, the proposed method is evaluated against a linear model-based DDPG baseline. The proposed method uses the nonlinear CSTR-based simulation environment, while the baseline uses a linearized transition model with the same state variables, action variables, reward function, and DDPG training procedure. The primary evaluation metric is Root Mean Square Error (RMSE), defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{26}$$

y_i is the target or observed reference value and $\hat{y}_i$ is the simulated controlled value. RMSE is computed for DO, ammonia, and pH. Cumulative reward is also used to evaluate the learning effectiveness and stability of the DDPG policy.

3. Results

3.1. Simulation-Based RMSE Evaluation

The first evaluation examined the closed-loop tracking performance of the nonlinear CSTR-DDPG framework and the linear-model DDPG baseline. Both methods used the same state variables, action variables, reward function, training length, and evaluation procedure. The only difference between the two approaches was the transition model used by the reinforcement learning environment. Therefore, the comparison was designed to evaluate the influence of the nonlinear simulation environment on policy learning, rather than to validate the model against physical water-quality measurements.

RMSE was computed between the controlled simulated trajectories and the predefined safe reference values for the three primary water-quality indicators: dissolved oxygen, ammonia, and pH. The reference values were selected from the target operating ranges used in the reward function, namely DO near 6.5 mg/L, ammonia close to 0 mg/L, and pH near 7.2. The RMSE comparison between the linear-model DDPG baseline and the nonlinear CSTR-DDPG framework is presented in table 4. Thus, the reported RMSE values represent simulation-based reference tracking errors and should not be interpreted as measurement errors against real nano-tank observations.

Table 4. Simulation-based RMSE comparison between linear-model DDPG and nonlinear CSTR-DDPG

Parameter	Linear Model DDPG RMSE	Nonlinear Model CSTR-DDPG RMSE	RMSE reduction
DO	1.24 mg/L	0.68 mg/L	45.2%
NH_3	0.25 mg/L	0.09 mg/L	64.0%
pH	0.31	0.12	61.3%

As shown in table 4, the nonlinear CSTR-DDPG framework produced lower RMSE values than the linear-model DDPG baseline for all three controlled variables. The largest improvement was observed for ammonia, followed by pH and dissolved oxygen. These results suggest that the nonlinear simulation environment provides a more expressive representation of coupled water-quality dynamics than the linear baseline. However, because the evaluation was conducted entirely in simulation, the result should be interpreted as simulation-based performance improvement rather than physical validation of the proposed model. Figure 2 presents the RMSE comparison between the linear-model DDPG baseline and the nonlinear CSTR-DDPG framework.

ISSN 2723-6471

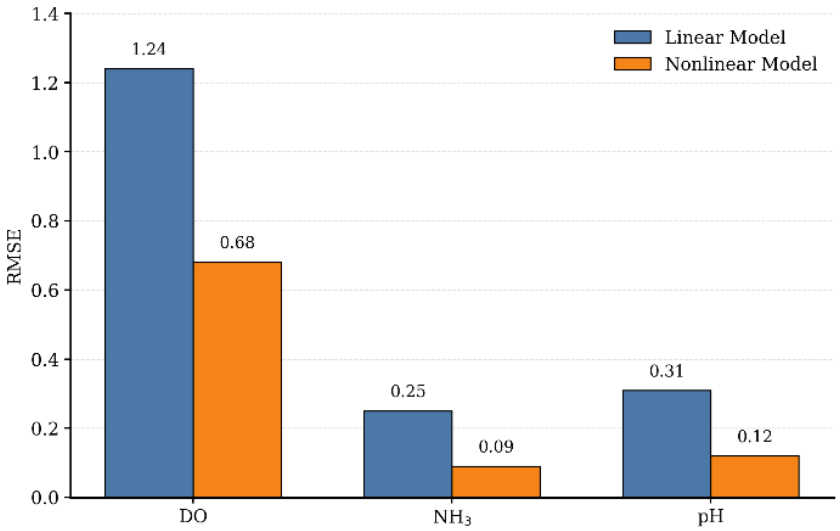

Figure 2. RMSE comparison between the linear model and nonlinear CSTR-based simulation model.

3.2. DDPG learning convergence

The second evaluation examined the learning convergence of the DDPG agent in the nonlinear simulation environment. The reward curve was used to observe whether the agent gradually learned a more stable control policy during training. Because the convergence analysis is based on a single training curve, no claim is made regarding variance across random seeds or statistical robustness of episode-level performance. The reward trajectory shows three learning phases. During the early stage, the cumulative reward was highly unstable and remained mostly in the negative region. This indicates that the agent was still exploring the continuous action space and had not yet learned effective control behavior. In this phase, unsafe or inefficient actions may occur more frequently, such as excessive feeding or insufficient aeration.

During the middle stage, the moving average reward increased gradually, indicating that the agent started to learn the relationship between control actions and water-quality response. The agent began to avoid actions that produced unsafe ammonia accumulation or dissolved oxygen reduction. After the later training stage, the reward curve became more stable and approached a positive region. This suggests that the DDPG agent reached a more consistent control policy in the simulation environment. Nevertheless, the convergence result only demonstrates learning stability within the simulated setting. Figure 3 presents the cumulative reward obtained during DDPG training in the nonlinear simulation environment.

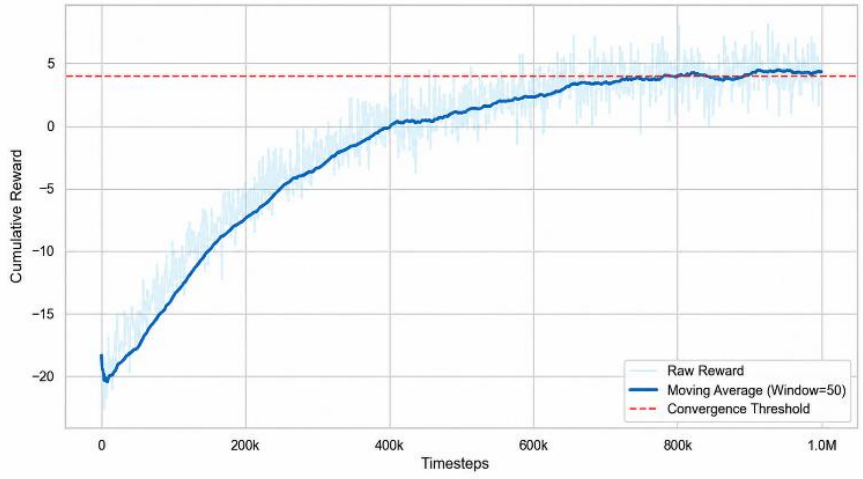

Figure 3. Cumulative reward convergence of DDPG training in the nonlinear simulation environment.

3.3. Closed-Loop control behavior

The closed-loop behavior of the trained DDPG policy was further examined to determine whether the learned actions were consistent with water-quality control objectives. During evaluation, the deterministic actor output was used without exploration noise. The agent adjusted aeration, feeding, and filtration actions based on the observed state of dissolved oxygen, ammonia, pH, temperature, BOD, and COD. When ammonia increased, the learned policy tended to reduce feeding input and increase filtration support. When dissolved oxygen decreased, the policy increased aeration intensity to restore the system toward the safe operating range. These responses indicate that the agent learned coordinated control behavior from the multivariable state representation. Nevertheless, this behavior was observed only within the simulation environment. Therefore, further testing with real actuators, sensor noise, response delay, and biological variability is required before practical implementation.

3.4. Hyperparameter sensitivity

The sensitivity analysis was conducted as an exploratory assessment to examine how selected DDPG hyperparameters affected reward convergence and policy stability. The tested hyperparameters included critic learning rate, replay buffer size, discount factor, soft update coefficient, and exploration noise. The results are summarized in table 5.

Table 5. Sensitivity of selected DDPG hyperparameters

Hyperparameter	Test Range	Selected value	Observed effect
Critic learning rate η_c	$10^{-6} - 10^{-2}$	5×10^{-4}	Influenced reward convergence
Replay buffer size B	$10^4 - 10^6$	10^5	Influenced reward stability
Discount factor γ	$0.90 - 0.99$	0.95	Influenced long-term reward
Soft update coefficient τ	$0.001 - 0.05$	0.005	Larger values produced policy oscillation
Exploration noise σ	$0.01 - 0.10$	0.05	Influenced exploration and action stability

The sensitivity analysis indicates that DDPG performance is affected by learning rate, replay buffer size, discount factor, target-network update rate, and exploration noise. However, this analysis was exploratory and was not intended as a full statistical ablation study. Repeated training with multiple random seeds, confidence intervals, and statistical significance testing should be included in future work to provide a more rigorous evaluation of hyperparameter effects.

3.5. Cumulative reward comparison

The final evaluation compared the cumulative reward obtained by the nonlinear CSTR-DDPG framework and the linear-model DDPG baseline. To avoid ambiguity, RMSE values are reported only in table 4, while this subsection focuses on reward-based control performance. The cumulative reward comparison is presented in table 6.

Table 6. Cumulative reward comparison between nonlinear CSTR-DDPG and linear-model DDPG

Metric	Nonlinear CSTR-DDPG	Linear-model DDPG	Difference
Cumulative reward per episode	850	600	+41.7%

As shown in table 6, the nonlinear CSTR-DDPG framework achieved a higher cumulative reward than the linear-model DDPG baseline. This result suggests that the nonlinear simulation environment helped the agent learn more stable control behavior under the defined reward structure. However, because the result is reported as a single-value comparison, it should be interpreted cautiously. Future evaluation should include repeated runs, standard deviation, confidence intervals, and comparisons with alternative RL algorithms such as TD3, SAC, PPO, and other continuous-control methods.

4. Discussion

The results of this study indicate that a nonlinear CSTR-based simulation environment can provide a more informative training condition for DDPG-based water-quality control than a linear model-based baseline. The lower RMSE values obtained for DO, NH_3, and pH suggest that the nonlinear formulation is better able to represent the coupled response of key water-quality variables under simulated nano-tank dynamics. This improvement can be explained by the

inclusion of oxygen transfer, ammonia-related nitrification, organic-load effects, temperature influence, and pH response, whereas the linear baseline assumes simpler proportional state transitions [6], [22].

The improvement in ammonia prediction and control is particularly important because ammonia is strongly affected by feeding input, organic decomposition, filtration, and nitrification [7], [10]. In a nano aquatic system, a small increase in feeding or organic load can lead to a relatively large change in ammonia concentration. Therefore, the inclusion of temperature, BOD, and COD as supporting state variables helps represent the broader transition dynamics that influence short-term water-quality stability. The reward convergence results shows that the DDPG agent gradually learned a more stable control policy within the simulation environment. During early training, the reward curve was unstable because the agent explored different combinations of aeration, feeding, and filtration. As training progressed, the moving average reward increased, indicating that the agent learned to avoid actions that produced unsafe deviations in DO, ammonia, or pH. However, this result demonstrates learning stability only within the simulated environment and should not be interpreted as evidence of fully validated physical autonomous control.

The learned policy also showed behaviour consistent with water-quality management objectives. When ammonia increased, the agent tended to reduce feeding input and increase filtration or aeration support. When DO decrease, the agent increased aeration and avoided actions that could further increase oxygen demand. This coordinated response is useful because nano aquatic systems are multivariable environments, where feeding decisions may affect ammonia accumulation, oxygen consumption, and pH stability [7], [10].

The comparison between the nonlinear CSTR-DDPG framework and the linear-model DDPG baseline suggests that the quality of the simulation environment strongly affects the learned control policy. Both approaches use the same DDPG algorithm and reward structure, but they differ in the model used to generate state transitions. The higher cumulative reward and lower RMSE values obtained by the nonlinear CSTR-DDPG framework indicate that the agent benefits from a more expressive simulation environment. The hyperparameter sensitivity results show that DDPG performance is affected by learning rate, replay buffer size, discount factor, soft update coefficient, and exploration noise. In particular, the soft update coefficient must remain small to avoid unstable target-network updates. The selected value of (τ=0.005) provided stable updates within the tested range, while larger values produced policy oscillation.

It is important to clarify that this study does not claim direct experimental superiority over tuned PID, fuzzy logic, or other conventional controllers because these methods were not implemented as direct baselines. Previous studies have demonstrated the importance of comparing reinforcement-learning controllers with conventional control approaches under consistent operating conditions [14]. Future studies should therefore include tuned PID, fuzzy logic, model predictive control, and other reinforcement learning algorithms to provide a broader comparative evaluation.

The findings have practical implications for simulation-based control design in nano aquatic systems. A small-volume tank is sensitive to disturbance, and direct trial-and-error learning on a physical tank may be risky for aquatic organisms. Training the DDPG agent in a nonlinear simulation environment provides a safer preliminary step before physical testing. Safety constraints are particularly important when reinforcement learning is applied to systems in which exploratory actions may produce undesirable operating conditions [24]. However, simulation cannot fully represent all real-world uncertainties. Real sensors may introduce noise and drift, actuators may have delays or limited precision [5], [21]. Biological responses may vary across species, feeding regimes, and tank conditions. These factors may affect the performance of a policy that appears stable in simulation.

The main limitation of this study is that the evaluation remains simulation-based. Physical closed-loop validation using calibrated sensors, real actuators, communication-delay analysis, and actual nano-tank experiments have not yet been conducted. In addition, the model does not explicitly represent individual organism behaviour, microbial population dynamics, long-term ecological adaptation, or sensor and actuator failure. The study also focuses on a single nano aquatic configuration, so generalization to different tank sizes, species, stocking densities, and multi-tank systems remains to be evaluated.

In addition, the present study does not yet evaluate sample efficiency across different reinforcement learning algorithms. The DDPG agent was trained for 200,000 timesteps, but this single training budget does not show whether the proposed framework learns faster or slower than alternative algorithms such as TD3, SAC, or PPO. Therefore, the

reported convergence curve should be interpreted only as evidence that the DDPG policy can learn a stable policy within the proposed simulation environment, not as proof of superior sample efficiency. Future studies should compare learning curves, final rewards, training stability, and computational time across multiple RL algorithms under the same nonlinear simulation environment.

A practical sim-to-real transfer strategy is also required before physical deployment. A possible transfer procedure includes calibrating the nonlinear simulation model with real nano-tank sensor data, adding sensor noise and actuator delay during simulation training, validating the trained policy in a shadow mode without directly controlling the actuators, and gradually moving from supervised semi-autonomous operation to full closed-loop control. This staged deployment is necessary because real sensors may drift, actuators may respond with delay, and biological conditions may vary over time. Without this step, the learned policy should not be considered ready for autonomous physical operation. Future research should validate the proposed framework in a real nano tank with calibrated sensors and controllable actuators. Future studies should also incorporate sensor noise, actuator delay, and hardware failure scenarios, and compare DDPG with PID, fuzzy logic, model predictive control, and alternative deep reinforcement learning algorithms under the same evaluation setting.

5. Conclusion

This study developed a simulation-based DDPG framework for water-quality control in nano aquatic systems. A nonlinear CSTR-based simulation model was formulated to represent the coupled dynamics of dissolved oxygen, ammonia, pH, temperature, BOD, and COD, and was integrated into an OpenAI Gym-compatible reinforcement learning environment. The DDPG agent learned continuous control actions for aeration, feeding, and filtration within the simulated nano-tank setting. The nonlinear CSTR-DDPG framework produced lower RMSE values than the linear-model DDPG baseline, with reductions of 45.2% for dissolved oxygen, 64.0% for ammonia, and 61.3% for pH. It also achieved a higher cumulative reward under the same reward structure. These results suggest that the nonlinear simulation environment provided a more informative training setting for DDPG-based control than the linear transition model.

However, the findings should be interpreted within the scope of simulation-based evaluation. The trained policy has not yet been validated in a physical closed-loop nano tank with calibrated sensors, actuator dynamics, communication delays, and biological variability. Future work should validate the framework in a real nano-tank system, incorporate sensor noise and actuator-delay models, compare DDPG with TD3, SAC, PPO, PID, fuzzy logic, and model predictive control, and evaluate robustness across multiple random seeds.

6. Declarations

6.1. Author Contributions

Conceptualization: I.F.R., S.E., and P.S.; Methodology: I.F.R., S.E., and P.S.; Software: I.F.R.; Validation: I.F.R., S.E., P.S., and H.F.H.; Formal Analysis: I.F.R., S.E., and P.S.; Investigation: I.F.R.; Resources: S.E., P.S., and H.F.H.; Data Curation: I.F.R. and H.F.H.; Writing Original Draft Preparation: I.F.R.; Writing Review and Editing: S.E., P.S., H.F.H., and I.F.R.; Visualization: I.F.R.; Supervision: S.E. and P.S.; Project Administration: I.F.R. and H.F.H.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data and simulation results supporting the findings of this study are available from the corresponding author upon reasonable request.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Proia, F. Cassió, C. Pascoal, A. Tlili, and A. M. Romaní, "The use of attached microbial communities to assess ecological risks of pollutants in river ecosystems: The role of heterotrophs," in *Emerging and Priority Pollutants in Rivers*, vol. 2012, no. January, pp. 55–83, 2012, doi: 10.1007/978-3-642-25722-3_3.

[2] B. Wong, K. T. Cheng, K. C. Y. E. Ngai, K. K. Yan, K. W. G. Lam, and C. K. D. Lo, "Solar-Aqua Sphere: A multifunctional solar-powered device for water generation, purification, and aquatic ecosystem support," *Int. J. Adv. Sci.*, vol. 1, no. 2, pp. 1-7, 2025, doi: 10.70731/6r9rnk47.

[3] A. K. Aufdenkampe, E. Mayorga, P. A. Raymond, J. M. Melack, S. C. Doney, S. R. Alin, R. E. Aalto, and K. Yoo, "Riverine coupling of biogeochemical cycles between land, oceans, and atmosphere," *Front. Ecol. Environ.*, vol. 9, no. 1, pp. 53–60, Feb. 2011, doi: 10.1890/100014.

[4] M. Łopata, J. K. Grochowska, R. Augustyniak-Tunowska, and R. Tandyrak, "Possibilities of improving water quality of degraded lake affected by nutrient overloading from agricultural sources by the multi-point aeration technique," *Appl. Sci.*, vol. 13, no. 5, p. 2861, pp. 1-16, Feb. 2023, doi: 10.3390/app13052861.

[5] C.-H. Chen, Y.-C. Wu, J.-X. Zhang, and Y.-H. Chen, "IoT-based fish farm water quality monitoring system," *Sensors*, vol. 22, no. 17, p. 6700, pp. 1-12, Sep. 2022, doi: 10.3390/s22176700.

[6] S. M. Elsherif, A. F. Taha, A. A. Abokifa, and L. Sela, "Comprehensive framework for controlling nonlinear multispecies water quality dynamics," *J. Water Resour. Plan. Manag.*, vol. 150, no. 2, pp. 1-13, Feb. 2024, doi: 10.1061/JWRMD5.WRENG-6179.

[7] S. Godoy-Olmos, I. Jauralde, R. Monge-Ortiz, M. C. Milián-Sorribes, M. Jover-Cerdá, A. Tomás-Vidal, and S. Martínez-Llorens, "Influence of diet and feeding strategy on the performance of nitrifying trickling filter, oxygen consumption and ammonia excretion of gilthead sea bream (*Sparus aurata*) raised in recirculating aquaculture systems," *Aquac. Int.*, vol. 30, no. 2, pp. 581–606, Apr. 2022, doi: 10.1007/s10499-021-00821-3.

[8] B. J. Eck, H. Saito, and S. A. McKenna, "Temperature dynamics and water quality in distribution systems," *IBM J. Res. Dev.*, vol. 60, no. 5–6, pp. 7:1–7:8, Sep. 2016, doi: 10.1147/JRD.2016.2594128.

[9] M. Alghamdi and Y. G. Haraz, "Smart biofloc systems: Leveraging artificial intelligence (AI) and Internet of Things (IoT) for sustainable aquaculture practices," *Processes*, vol. 13, no. 7, p. 2204, pp. 1-26, Jul. 2025, doi: 10.3390/pr13072204.

[10] F. M. Yusoff, W. A. D. Umi, N. M. Ramli, and R. Harun, "Water quality management in aquaculture," *Cambridge Prisms: Water*, vol. 2, no. May, p. e8, pp. 1-22, May 2024, doi: 10.1017/wat.2024.6.

[11] R. Harahap, A. Adyatma, and F. Fahmi, "Automatic control model of water filling system with Allen Bradley Micrologix 1400 PLC," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 309, no. February, p. 012082, pp. 1-8, Feb. 2018, doi: 10.1088/1757-899X/309/1/012082.

[12] M. A. Ali, A. H. Miry, and T. M. Salman, "IoT based water tank level control system using PLC," in *2020 Int. Conf. Comput. Sci. Softw. Eng. (CSASE)*, vol. 2020, no. Apr., pp. 7–12, 2020, doi: 10.1109/CSASE48920.2020.9142067.

[13] K. Shakya, G. Pillai, and S. Chakrabarty, "Reinforcement learning algorithms: A brief survey," *Expert Syst. Appl.*, vol. 231, no. November, pp. 1-18, Nov. 2023, doi: 10.1016/j.eswa.2023.120495.

[14] F. Hernández-del-Olmo, E. Gaudioso, N. Duro, R. Dormido, and M. Gorrotxategi, "Advanced control by reinforcement learning for wastewater treatment plants: A comparison with traditional approaches," *Appl. Sci.*, vol. 13, no. 8, pp. 1-19, Apr. 2023, doi: 10.3390/app13084752.

[15] H. Tan, "Reinforcement learning with deep deterministic policy gradient," in *2021 Int. Conf. Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, vol. 2021, no. May, pp. 82–85, 2021, doi: 10.1109/CAIBDA53561.2021.00025.

[16] E. H. Sumiea, S. J. Abdulkadir, H. S. Alhussian, S. M. Al-Selwi, A. Alqushaibi, M. G. Ragab, and S. M. Fati, "Deep deterministic policy gradient algorithm: A systematic review," *Heliyon*, vol. 10, no. 9, pp. 1-26, May 2024, doi: 10.1016/j.heliyon.2024.e30697.

[17] M. El-Harbawi, "Air quality modelling, simulation, and computational methods: A review," *Environ. Rev.*, vol. 21, no. 3, pp. 149–179, Sep. 2013, doi: 10.1139/er-2012-0056.

[18] B. Feng, J. Ma, Y. Liu, L. Wang, X. Zhang, Y. Zhang, J. Zhao, W. He, Y. Chen, and L. Weng, "Application of machine learning approaches to predict ammonium nitrogen transport in different soil types and evaluate the contribution of control factors," *Ecotoxicol. Environ. Saf.*, vol. 284, no. October, pp. 1-11, Oct. 2024, doi: 10.1016/j.ecoenv.2024.116867.

[19] Q. Jiang, J. Li, Y. Sun, J. Huang, R. Zou, W. Ma, H. Guo, Z. Wang, and Y. Liu, "Deep-reinforcement-learning-based water diversion strategy," *Environ. Sci. Ecotechnology*, vol. 17, no. Jan., pp. 1-12, Jan. 2024, doi: 10.1016/j.ese.2023.100298.

[20] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and P. Joelle, "An introduction to deep reinforcement learning," *Found. Trends Mach. Learn.*, vol. 11, no. 3–4, pp. 219–354, Dec. 2018, doi: 10.1561/2200000071.

[21] Essamlali, H. Nhaila, and M. El Khaili, "Advances in machine learning and IoT for water quality monitoring: A comprehensive review," *Heliyon*, vol. 10, no. 6, pp. 1-19, Mar. 2024, doi: 10.1016/j.heliyon.2024.e27920.

[22] J. Vojtesek and P. Dostal, "Simulation analyses of continuous stirred tank reactor," in *ECMS 2008 Proceedings*, vol. 2008, no. Jun., pp. 506–511, 2008, doi: 10.7148/2008-0506.

[23] Z. Wei, J. Xu, Y. Lan, J. Guo, and X. Cheng, "Reinforcement learning to rank with Markov decision process," in *Proc. 40th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, New York, NY, USA, vol. 2017, no. Aug., pp. 945–948, 2017, doi: 10.1145/3077136.3080685.

[24] Wachi and Y. Sui, "Safe reinforcement learning in constrained Markov decision processes," in *Proc. 37th Int. Conf. Machine Learning*, H. Daumé III and A. Singh, Eds., *Proc. Mach. Learn. Res.*, vol. 119, no. 2020, pp. 9797–9806, 2020.

[25] P. Wang, H. Li, and C.-Y. Chan, "Continuous control for automated lane change behavior based on deep deterministic policy gradient algorithm," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, vol. 2019, no. Jun., pp. 1454–1460, 2019, doi: 10.1109/IVS.2019.8813903.