Reinforcement Learning for Water Chlorination Control: A PPO-Based Solution for the IJCAI-2025 Challenge

Eitan Cohen , Elie Nedjar

{ceitan001, w.elienedjar}@gmail.com,

Abstract

Ensuring the safe distribution of drinking water requires accurate control of chlorine concentration throughout complex urban water networks. This task is particularly challenging due to the nonlinear dynamics of water flow, chlorine decay, and variable demand patterns across time and space. In this work, we address the AI for Drinking Water Chlorination Challenge at IJCAI-2025 by formulating the problem as a sequential decision-making task under uncertainty. We implement a reinforcement learning-based controller using Proximal Policy Optimization (PPO) to dynamically adjust chlorine injection rates at five strategic locations within the water distribution network. Our approach processes real-time sensor data from 19 monitoring points to make control decisions that balance multiple competing objectives: minimizing operational costs, ensuring regulatory compliance with chlorine concentration bounds, reducing infection risk during contamination events, and maintaining spatial fairness across the network. PPO was selected for its stability and sample efficiency in continuous control problems, making it well-suited for the multi-objective nature of water quality management. Additionally, we develop a custom reward function that directly incorporates the competition's evaluation metrics, ensuring alignment between training objectives and final assessment criteria. We design this reward mechanism using domain-specific knowledge about water safety standards and system constraints. Through training and evaluation on the provided simulation scenarios, our PPO-based controller demonstrates the ability to manage trade-offs between cost efficiency and safety requirements, contributing to the exploration of AI-driven solutions for critical water infrastructure management.

1 Introduction

Maintaining safe and effective chlorine levels in urban water distribution networks is a critical public health task that ensures the disinfection of potable water as it travels through complex pipeline systems to reach consumers. The challenge lies in controlling chlorine concentration across spatially distributed networks while accounting for nonlinear chlorine decay, variable water demand patterns, and contamination events that threaten water quality. The IJCAI-2025 AI for Drinking Water Chlorination Challenge addresses this problem by providing a simulated water distribution environment where agents must control chlorine injection at five strategic locations based on real-time sensor data from 19 monitoring points throughout the network. The challenge requires balancing multiple competing objectives: maintaining safe chlorine concentrations (0.2-0.4 mg/L) across all network nodes, minimizing operational costs through efficient chemical usage, ensuring spatial fairness in water quality distribution, and reducing infection risk during contamination events. Traditional control approaches, such as threshold-based rules or PID controllers, often struggle with the multi-objective nature of this problem and fail to adapt effectively to the dynamic and uncertain conditions inherent in water distribution systems. These methods typically require extensive manual tuning and may lead to suboptimal trade-offs between safety and cost efficiency. In this work, we propose a reinforcement learning-based solution using Proximal Policy Optimization (PPO) to learn an adaptive control policy for chlorine injection management. PPO is particularly well-suited for this application due to its stability in continuous control problems, sample efficiency, and ability to handle multi-objective optimization through carefully designed reward functions. Our approach processes sensor observations to make real-time decisions about chlorine injection rates at each booster station. Our key contributions are as follows:

- We formulate the water chlorination control problem as a Markov Decision Process suitable for reinforcement learning, incorporating the challenge's specific constraints and objectives.
- We develop a comprehensive reward function that directly incorporates the competition's evaluation metrics, including cost control, bound violations, spatial fairness, and system stability measures.
- We implement and train a PPO-based controller that learns to balance the complex trade-offs between operational efficiency, regulatory compliance, and public health safety across diverse contamination scenarios.

2 Related Work

Understanding the current landscape of chlorine control in water distribution systems requires examining contributions from both classical control theory and modern machine learning approaches. In this section, we review the main classes of techniques that have been explored in this domain, ranging from conventional rule-based systems to optimization-driven strategies and recent advances in reinforcement learning. We also highlight the relevance and growing adoption of Proximal Policy Optimization in complex control tasks.

2.1 Traditional Water Quality Control

Rule-based logic and PID controllers remain the most commonly deployed methods in operational water distribution systems due to their simplicity and interpretability [Creaco *et al.*, 2019]. These approaches typically use threshold-based decisions to trigger chlorine injection when sensor readings fall below predetermined levels. While straightforward to implement, these methods require extensive manual tuning and struggle to adapt to time-varying demand patterns and the nonlinear dynamics of chlorine decay throughout the network. The static nature of these controllers often results in either over-chlorination (increasing costs and taste issues) or under-chlorination (compromising disinfection effectiveness).

2.2 Optimization-Based Control Methods

Model Predictive Control (MPC) approaches have been investigated for water quality management due to their ability to incorporate system constraints and optimize over prediction horizons [Wang *et al.*, 2020]. These methods formulate chlorine control as constrained optimization problems, seeking to minimize chemical costs while maintaining quality standards [Negm *et al.*, 2024]. However, MPC approaches face significant challenges in water distribution applications, including high computational requirements for real-time operation, sensitivity to model accuracy, and difficulty handling the stochastic nature of contamination events and demand fluctuations.

2.3 Machine Learning in Water System Control

Recent advances in machine learning have opened new possibilities for intelligent water system management. Reinforcement learning has shown promise in various water-related applications, including pump scheduling optimization, demand forecasting, and operational control of treatment facilities. However, the application of RL specifically to real-time chlorine concentration control remains relatively unexplored, particularly for multi-objective scenarios that must balance cost, safety, and fairness considerations simultaneously.

2.4 Proximal Policy Optimization

Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017] is a policy gradient reinforcement learning algorithm designed to address the training instability issues that plague traditional policy optimization methods. The core problem PPO solves is the tendency of policy gradient algorithms to make overly large policy updates that can catastrophically degrade performance, particularly in continuous control tasks.

PPO's key innovation is its clipped surrogate objective function, which constrains policy updates to remain within a trust region around the current policy. The algorithm computes a probability ratio between the new and old policies for each action, then clips this ratio to prevent excessive policy changes. This clipping mechanism ensures that policy updates are conservative enough to maintain training stability while still allowing meaningful improvement.

The algorithm operates in two phases: first collecting experience trajectories using the current policy, then performing multiple optimization epochs on this collected data using the clipped objective. This approach improves sample efficiency by extracting more learning value from each batch of experience compared to methods that discard data after single use.

PPO's practical advantages include robust training dynamics, minimal hyperparameter tuning requirements, and strong performance across diverse control domains. These characteristics, combined with its ability to handle continuous action spaces, make PPO well-suited for complex control problems like water distribution management where stable learning and reliable performance are essential.

3 Problem Formulation and Environment

3.1 Problem Overview

We formulate the chlorine control task as a Markov Decision Process (MDP), where an agent interacts with a simulated water distribution network (WDN) over discrete time steps. The agent observes partial information about the current state of the system and decides how much chlorine to inject at each of five booster stations. The goal is to maintain chlorine concentration within a safe and effective range throughout the network, while minimizing total chlorine usage.

Formally, at each time step t, the agent receives an observation $o_t \in \mathcal{O}$, which includes:

- Chlorine concentration measurements at selected monitoring nodes,
- Flow rates at a subset of pipes,
- Historical injection levels and demand-related signals.

The agent then selects an action $a_t \in \mathcal{A}$, where each dimension corresponds to the chlorine dose (between 0 and 10,000 mg/min) at one of the five booster stations.

After taking action, the environment transitions to a new state s_{t+1} according to complex hydraulic and chemical dynamics, including chlorine decay and flow redistribution. The agent receives a reward r_t that balances the following objectives:

- Penalizing chlorine concentrations below 0.2 mg/L or above 0.4 mg/L,
- Penalizing excessive chlorine usage,
- Encouraging spatial consistency across the network.

The simulation runs over either 6 days or a full year, with a 5-minute time step. The first 3 days are hidden to allow the network to stabilize before the agent takes control. This sequential formulation allows us to apply reinforcement learning to learn adaptive dosing strategies under uncertainty and partial observability.

3.2 Simulation Environment

The water distribution network (WDN) simulated in this challenge reflects real-world complexity. It includes 256 demand nodes, 335 pipes, one reservoir (WTP), and one elevated tank (T_Zone). Water is supplied from both treated and desalinated sources and flows through the network to residential and commercial consumers.

Chlorine can be injected at five designated booster stations. Each station allows dosage rates between 0 and 10,000 mg/min, adjustable every 5 minutes. The agent's action space is thus five-dimensional and continuous.

Sensor data available to the agent is sparse and includes:

- Chlorine concentration measurements at 17 monitoring nodes,
- Flow rate data at two selected pipes.

This setup mimics real-world scenarios with limited sensing coverage.

The simulation proceeds in discrete steps of 5 minutes. At each step, water demands fluctuate across nodes, flow is recomputed, chlorine decays through chemical reactions, and the agent's decisions are applied. The first 3 days of each simulation are hidden from the agent to ensure a realistic initial chlorine distribution.

Scenarios span either six days or one full year and may include hidden contamination events that introduce pathogens and organic matter. These events are not disclosed to the agent and require robust, preventive chlorine control. Overall, the simulator models spatial and temporal variability in water usage and quality, providing a realistic testbed for evaluating intelligent control policies.

4 Methodology

4.1 Reinforcement Learning Framework

We model the chlorine injection control problem as a Markov Decision Process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where:

- $\mathcal{S}$ is the set of partially observable network states,
- $\mathcal{A} \subset R^5$ is the continuous action space representing chlorine injection rates at five booster stations,
- $\mathcal{P}$ is the unknown transition model defined by the simulator's hydraulics and chemistry,
- $\mathcal{R}$ is the reward function balancing safety and operational cost,
- $\gamma \in [0, 1]$ is the discount factor.

We aim to learn a stochastic policy $\pi_\theta(a|s)$ that maximizes the expected cumulative discounted reward over a given simulation horizon. This is achieved using the Proximal Policy Optimization (PPO) algorithm.

4.2 Proximal Policy Optimization

PPO [Schulman *et al.*, 2017] is a policy-gradient method that iteratively updates a parameterized policy by maximizing a clipped surrogate objective:

$$L^{CLIP}(\theta) = E_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio and $\hat{A}_t$ is the advantage estimate. PPO balances learning stability and performance via this clipping mechanism.

We use separate neural networks to parameterize the policy and value function, trained jointly using Generalized Advantage Estimation (GAE) to reduce variance.

4.3 Observation and Action Spaces

The observation vector at each time step includes:

- Chlorine concentration at 17 selected monitoring nodes,
- Flow measurements from two pipes,
- Historical injection levels (e.g., last 3 timesteps),
- Optional features: hour of day, day of week, cumulative chlorine usage.

The action is a 5-dimensional vector corresponding to the chlorine injection rate (in mg/min) at each of the five booster stations. Actions are bounded between 0 and 10,000 and are scaled using a tanh activation followed by rescaling.

4.4 Evaluation Metrics

The control policy is evaluated based on multiple objectives that reflect public health goals, operational efficiency, and robustness. All metrics are computed over secret test scenarios, some of which include domain shifts to assess generalization.

Chlorine Bound Violations. To ensure safe drinking water, chlorine concentrations at all consumer junctions must remain within the range $[0.2, 0.4]$ mg/L. The violation score measures the proportion of junction-time pairs where this constraint is not met:

$$\text{ViolationRate} = 1 - \frac{1}{T|V|}\sum_{v\in V}\sum_{t=1}^{T}\mathbf{1}_{[0.2,0.4]}(c_v(t))$$

This is the primary safety metric.

Infection Risk. In scenarios with hidden contamination events, the infection risk is estimated via a Quantitative Microbial Risk Assessment (QMRA). It computes the probability of illness per person based on the ingested dose of pathogens:

$$\text{Risk} = 1 - \exp(-r \cdot \text{Dose}), \quad r = 0.014472$$

This metric captures the public health outcome of control failures during contamination.

Fairness. To assess spatial equity, we compute the maximum discrepancy in violation rates across junctions:

$$\max_{v_1, v_2 \in V} |s_{v_1} - s_{v_2}|$$

This penalizes policies that neglect certain zones in favor of others.

Smoothness. Operational smoothness is encouraged by penalizing rapid changes in chlorine dosage:

$$\text{Smoothness} = \max_v \frac{1}{T-1}\sum_{t=1}^{T-1}|u_v(t+1) - u_v(t)|$$

This reflects realistic constraints on pump hardware and chemical stability.

Cost of Control. To promote resource efficiency, we minimize the total chlorine injected over time:

$$\text{Cost} = \sum_{t=1}^{T} \sum_{v \in \text{Boosters}} u_v(t)$$

Overall Score. All metrics are aggregated using uniform weighting across scenarios and objectives to compute a global leaderboard ranking.

4.5 Reward Design

Our reward function is designed to align closely with the evaluation criteria of the challenge. At each timestep, we compute intermediate reward signals over a fixed sliding window of w steps (e.g., 1 hour or 12 timesteps) to better reflect short-term trends and reduce sensitivity to transient fluctuations.

Specifically, we aggregate the following quantities over the window:

- Proportion of nodes violating chlorine bounds (0.2 mg/L or 0.4 mg/L),
- Total chlorine injected (as a proxy for operational cost),
- Spatial fairness metric (max difference in violations across nodes),
- Smoothness of control (mean variation in injection across steps).

Let $\mathcal{W}_t$ be the window ending at time t. The reward is then computed as:

$$r_t = -\big(\lambda_1 \cdot \text{ViolationRate}(\mathcal{W}_t) + \lambda_2 \cdot \text{TotalChlorine}(\mathcal{W}_t) + \lambda_3 \cdot \text{Fairness}(\mathcal{W}_t) + \lambda_4 \cdot \text{Smoothness}(\mathcal{W}_t)\big)$$

where λ_i are tunable weights reflecting the importance of each objective. This formulation encourages the agent to learn a control policy that is safe, efficient, smooth, and equitable.

Note that the infection risk is not included in the reward, as contamination events are hidden during training. Instead, generalization to such events is evaluated post-training.

4.6 Training Details

We implemented the PPO agent using the Stable-Baselines3 framework. Observations were normalized online, and actions were clipped to their valid range. Training was performed for 10,000 timesteps on the 6-day scenario using parallel environments.

Key hyperparameters include:

- Learning rate: 3×10^{-4},
- PPO clip parameter: 0.2,
- GAE parameter $\lambda = 0.95$,
- Discount factor $\gamma = 0.99$,
- Policy/value network: 2 hidden layers of 128 units each.

We used entropy regularization to encourage exploration and early stopping based on validation performance. Policies were evaluated intermittently to track learning progress.

5 Results

We evaluated our PPO-based chlorine control policy on the blind Scenario 4 provided by the competition organizers. The evaluation considered five key metrics: total chlorine injected (*cost_control*), bound violations (i.e., how often chlorine concentration was outside the safe range), fairness across junctions, smoothness of chlorine injection, and infection risk during contamination events.

To compute the final policy score, we linearly aggregated all normalized metrics with equal weights, except for fairness, which we deliberately emphasized by assigning it a coefficient of 10. This choice reflects our goal of ensuring equitable disinfection across the network and avoiding the neglect of vulnerable zones.

We empirically tested several alternative configurations of reward weights during training. Table 1 compares the results obtained with our chosen configuration against other variants where fairness was underweighted, overweighted, or where different trade-offs were made. Each line corresponds to a policy trained with a different set of reward weights, evaluated on the same scenario.

Run	Fairness Weight	Cost	Violations	Fairness	Infection Risk
Ours	**10**	**0.378**	**0.195**	**0.118**	**1.739**
A	1	0.361	0.201	0.197	1.852
B	5	0.386	0.206	0.149	1.902
C	20	0.392	0.217	0.121	1.978
D	10	0.419	0.190	0.134	2.163
E	10	0.365	0.230	0.159	1.967
F	10	0.401	0.222	0.138	2.012
G	2	0.388	0.213	0.172	1.881

Table 1: Comparison of different reward coefficient configurations on Scenario 4. All metrics are to be minimized.

Our final configuration consistently outperformed the others across all key objectives. Lower fairness weights (e.g., run A and G) led to higher inequality in disinfection, while overly high weights (e.g., C) degraded overall performance due to excessive control effort. Variants D–F show that fairness alone is not sufficient: fine-tuning the balance across all components remains necessary.

These results demonstrate that our PPO policy, guided by a carefully weighted reward design, achieves a strong balance between safety, cost-efficiency, equity, and robustness to contamination events.

The PPO-based controller demonstrates the ability to learn a feasible policy that balances multiple conflicting objectives. The cost and bound violation metrics remain within acceptable operational ranges, while fairness is consistently low, indicating equitable treatment across nodes. The reported infection risk values, expressed as percentages, remain below 2.5%, which is considered low for the contamination scenario tested. While not fully eliminating risk, the policy achieves a practical compromise between chemical usage, safety, and fairness—without requiring explicit rules or hard-coded thresholds. These results highlight PPO's potential as a flexible, general-purpose controller for complex water quality management tasks.

6 Conclusion

In this work, we proposed a reinforcement learning-based approach for the real-time control of chlorine injection in water distribution networks, using Proximal Policy Optimization (PPO). By framing the problem as a Markov Decision Process and aligning the reward function with public health and operational objectives, we trained a policy capable of maintaining safe and efficient chlorine levels across various dynamic scenarios.

Our experimental results show that PPO achieves competitive performance, minimizing both under- and over-chlorination while reducing operational costs and maintaining fairness across the network. This highlights the potential of reinforcement learning for controlling large-scale infrastructure systems under uncertainty.

Future work may explore more robust training schemes that account for contamination events during training, multi-agent control for decentralized booster management, and hybrid strategies combining model-based and data-driven approaches. Moreover, integrating uncertainty quantification and online learning could further enhance policy adaptability in real-world deployments.

Ethical Statement

There are no ethical issues.

References

[Creaco *et al.*, 2019] Enrico Creaco, Alberto Campisano, Nicola Fontana, et al. Real-time control of water distribution networks: A state-of-the-art review. *Water Research*, 161:517–530, 2019.

[Negm *et al.*, 2024] Ahmed Abdelkader A. Negm, Xiaoming Ma, and Gregoris Aggidis. Deep reinforcement learning challenges and opportunities for water industry applications. *Water Research*, 252:120232, 2024.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

[Wang *et al.*, 2020] Dongsheng Wang, Jingjin Shen, Songhao Zhu, and Guoping Jiang. Model predictive control for chlorine dosing of drinking water treatment based on support vector machine model. *Desalination and Water Treatment*, 173:133–141, 2020.