Final Reinforcement Learning-driven advice provision module for advice provision based on Soft Sensors

Deliverable 4.4

WP4 Soft sensors for water quality monitoring and improved water system performance awareness

Funded by the European Union

GRANT AGREEMENT NUMBER	101082035		
FULL TITLE / ACRONYM	ToDrinQ		
START DATE	01-12-2022	DURATION	48 months
END DATE	30-11-2026		
PROJECT URL	www.todrinq.eu		
WORK PACKAGE No and title	WP4 Soft sensors for water quality monitoring and improved water system performance awareness		
DELIVERABLE TITLE	Final Reinforcement Learning-driven advice provision module for advice provision based on Soft Sensors		
ACTUAL DATE OF DELIVERY	12-12-2025 (1) & 20-04-2026 (2)		
NATURE	R	DISSEMINATION LEVEL	Public
LEAD BENEFICIARY	National Technical University of Athens		
RESPONSIBLE AUTHOR	Prof. Christos Makropoulos, NTUA		
CONTRIBUTIONS FROM	Iosif Spartalis, NTUA Greg Kyritsakas, TUD Siddhart Seshan, KWR Panagiotis Kossieris, NTUA George Bariamis, NTUA Professor Luuk Rietveld, TUD		

Document history

Version	Description (Section, page number)	Author	Organisation short name
V0.1	First draft	Iosif Spartalis, NTUA Greg Kyritsakas, TUD Jasper Imnik, KWR Panagiotis Kossieris, NTUA George Bariamis, NTUA	National Technical University of Athens
V0.2	Draft final 1 Processed of feedback received from reviewers / partners	Iosif Spartalis, NTUA Greg Kyritsakas, TUD Jasper Imnik, KWR Panagiotis Kossieris, NTUA George Bariamis, NTUA	National Technical University of Athens
V0.3	Review	Lydia Vamvakeridou-Lyroudia	KWR Water Research Institute
V0.4	Review	Christos Makropoulos	National Technical University of Athens
V1.0	Final version	Luuk Rietveld	Delft University of Technology
V2.0	Revised version	Christos Makropoulos	National Technical University of Athens

Quality control

Author	Organisation short name	Role	Date
Christos Makropoulos	National Technical University of Athens	Deliverable Leader	10-11-2025
George Bariamis	National Technical University of Athens	Work Package Leader	19-11-2025
Lydia Vamvakeridou-Lyroudia	KWR Water Research Institute	Reviewer 1	21-11-2025
Luuk Rietveld	Delft University of Technology	Scientific project coordinator	09-11-2025
Danitsja van Heusden-van Winden	Delft University of Technology	Project coordinator	12-11-2025
Panagiotis Kossieris	National Technical University of Athens	Deliverable Leader	14-4-2026
Christos Makropoulos	National Technical University of Athens	Work Package Leader	14-4-2026
Danitsja van Heusden	Delft University of Technology	Project coordinator	20-04-2026

Abbreviations

AI	-	Artificial Intelligence
AM	-	Approximation Model
AWD	-	Amsterdam Water supply Dunes
BC	-	Behavioural Cloning
Cl	-	Chlorine
DC	-	Demo Case
DCL	-	Demo Case Leader
DDPG	-	Deep Deterministic Policy Gradient
DOC	-	Dissolved Organic Carbon
DPG	-	Deterministic Policy Gradient
DQN		Deep Q-learning network
DWTP	-	Drinking Water Treatment Plant
$FeCl_3$	-	Ferric Chloride
FCl	-	Free Chlorine
FTU	-	Formazin Turbidity Unit
HAA	-	Haloacetic Acids
KPI	-	Key Performance Indicator
LLM		Large Language Model
MAE	-	Mean absolute error
MDP	-	Markov Decision Process
ML		Machine Learning
NOM	-	Natural Organic Matter
NTU	-	Nephelometric Turbidity Unit
PC	-	Project Coordinator
Ppm	-	Parts per Million
RL	-	Reinforcement learning
CADA	-	Supervisory Control and Data Acquisition
TD		Temporal Difference
TD3-BC	-	Twin Delayed Deep Deterministic Policy Gradient
THMs	-	Trihalomethanes
TOC	-	Total Organic Carbon

TTHM - Total Trihalomethanes

UV254 - Ultraviolet Absorbance at 254 nm

Short name	Legal Name
TUD	TECHNISCHE UNIVERSITEIT DELFT
RWTH	RHEINISCH-WESTFAELISCHE TECHNISCHE HOCHSCHULE AACHEN
CEB	CENTRE BELGE D'ETUDE ET DE DOCUMENTATION DE L'EAU
NTUA	ETHNICON METSOVION POLYTECHNION
KWR	KWR WATER BV
WTNT	STICHTING WATERNET
HWL	HET WATERLABORATORIUM NV
VEF	VEOLIA EAU - COMPAGNIE GENERALE DES EAUX SOCIETE EN COMMANDITE PAR ACTIONS
VEOCZ	VEOLIA CESKA REPUBLIKA, A.S.
EYDAP	ETAIREIA YDREYSEOS KAI APOCHETEFSEOS PROTEYOYSIS ANONIMI ETAIREIA
OLI	OLISENS TECH
OXY	OXYMEM LIMITED
ORV	ORVION B.V.
CHEM	CHIMIKI TECHNOLOGIA P. DIMOPOULOU -P.TAZES & SIA OE
WE	WATER EUROPE
ALTIS	ALTIS Groupe SA
BNV	BNOVATE TECHNOLOGIES SA
VEHO	Veolia Holding Ceska Republika AS

Table of contents

List of figures

List of tables

Executive summary

Drinking water treatment plants (DWTPs) are critical components of the water supply chain, ensuring the uninterrupted delivery of high-quality water. However, the complexity of large-scale operations, combined with the variability of raw water quality, imposes significant challenges. To address these issues, the ToDrinQ project has developed advanced tools to support plant operators by delivering real-time, stage-specific treatment advice. This document constitutes the final deliverable of Task 4.4, presenting the finalized Reinforcement Learning (RL) algorithms designed to provide evidence-based advice to operators. The project leveraged Reinforcement Learning, a machine learning approach where agents learn sequential decision-making to maximize cumulative rewards. While RL has been explored in wastewater treatment, its application in drinking water treatment remains nascent. Consequently, the ToDrinQ developments represent a significant advancement beyond the current state of the art in the automated management and control of DWTP processes.

Demo Case #1: Leiduin Plant (Amsterdam) For the Leiduin plant, which supplies approximately 70% of Amsterdam's drinking water using river water from the Lek Canal, the focus was on optimizing the coagulation–flocculation process. An RL agent was employed (a Deep Q-Learning Network (DQN) to suggest optimal dosages of $FeCl_3$ that ensure particle destabilization and floc formation in six-hour intervals.

Demo Case #2: Polydendri Plant (Athens) For the Polydendri WTP, which serves the Athens metropolitan area with a capacity of 200,000 m^3/day, three distinct RL models were developed to optimize critical stages of the multi-stage treatment process, pre-chlorination, coagulation and post-disinfection. All three Athens models utilize the Twin Delayed Deep Deterministic Policy Gradient with Behaviour Cloning (TD3-BC) algorithm, suitable for continuous control tasks in offline environments.

To address the scarcity of open-source simulation tools compatible with modern AI frameworks, the project team successfully re-implemented the core functionalities of the Stimela modelling suite into Python, resulting in the development of **pyStimela package**. This development overcomes the limitations of commercial, license-restricted software, enabling the seamless coupling of physics-based water treatment simulations with Reinforcement Learning agents. Used specifically for the pre-chlorination RL environment (RL Model #4), pyStimela served as a robust proof-of-concept for online RL training, demonstrating the technical feasibility of real-time interaction between AI agents and dynamic process simulators.

This Deliverable (D4.4) presents the finalized Reinforcement Learning algorithms for the four advice provision modules. These models have successfully demonstrated stability and the capacity to optimize chemical dosing in offline environments. Having achieved the objectives of Task 4.4, these modules are now ready for transition to Work Package 7 (Task 7.4). In the next phase, they will be integrated into the online modular platform to undergo real-world operator validation and continuous learning loops within the Demonstration Cases (WP2). This document is the final Deliverable for WP4. Highlighting the project's integrated approach, pyStimela has also been employed in Work Package 6 to power the optimization module of the design support tool, with progress detailed in Deliverable 6.2.

Even though the RL models have been developed around the needs of Demo Case #1 and Demo Case #2, they have strong potential for replication, upscaling, and adoption in diverse settings, since the treatment processes examined typically appear in almost all DWTPs in Europe and beyond. Particularly, the algorithms can be generalized to other DWTPs by adapting input parameters, algorithms, and reward functions to suit specific plant configurations. These RL agents are advisory systems designed to support, not replace, human operators, ensuring compliance with safety protocols.

1. Introduction

1.1 Context and objectives

Drinking water treatment plants (DWTPs) are critical infrastructures within the water supply chain, tasked with ensuring the uninterrupted delivery of high-quality water to end users. However, the complexity of large-scale operations, combined with the inherent variability of raw water quality, imposes significant operational challenges. To address these, the ToDrinQ project is developing advanced tools to support plant operators in the optimal management of DWTPs by delivering real-time, stage-specific treatment advice.

The operation of Drinking Water Treatment Plants (DWTPs) is a complex, dynamic task requiring the continuous balancing of water quality compliance, operational safety, and resource efficiency. Traditional control strategies often rely on reactive adjustments based on operator experience or static setpoints. **ToDrinQ aims to transition these systems toward proactive, evidence-based management by developing an advice provision module driven by Artificial Intelligence (AI).**

This deliverable presents the final outcomes of Task 4.4 and of WP4, specifically the development of Reinforcement Learning (RL) algorithms designed to optimize key treatment processes. Unlike supervised learning, which predicts outcomes based on historical labels, RL trains an "agent" to make sequential decisions (e.g., dosing adjustments) to maximize a cumulative reward (e.g., minimized chemical use while maintaining turbidity targets).

1.2 Reinforcement learning in process control

Reinforcement Learning (RL) is a branch of machine learning where agents learn to make sequential decisions by interacting with an environment to maximize cumulative rewards (Sutton & Barto, 2018). Unlike traditional optimization methods, RL is particularly well-suited for dynamic systems with complex, nonlinear behaviours. In water treatment processes, these characteristics are prevalent due to variability in water quality, seasonal changes, and operational constraints (Zheng et al., 2020). The flexibility of RL in handling real-time data makes it highly suitable for predictive control applications in water treatment. Data-driven RL models have demonstrated improved process control by adapting to changing water quality data, which is critical for effective water treatment optimization (Wang et al., 2018). Through applications of model-free RL, water treatment facilities can enhance water quality, improve operational efficiency, and minimize costs, contributing to more sustainable water management practices (Singh & Yadav, 2021).

Within this task, three RL use cases are being developed by NTUA to optimize key water treatment processes: pre-chlorination, coagulation, and post-chlorination. Each RL application follows a model-free framework, relying primarily on data collected from the monitoring systems of the Polydendri Drinking Water Treatment Plant (Demo Case #2). This data provides essential water quality parameters, enabling the optimization of chemical dosing strategies and process control actions without requiring an explicit mathematical model of the treatment plant. The RL agents are designed to achieve specific performance objectives established by plant operators, such as minimizing chemical consumption while maintaining compliance with target water quality standards. By continuously learning from real-time and historical operational data, each RL implementation adapts its decision-making policies to enhance treatment efficiency, process robustness, and overall operational stability under varying environmental and inflow conditions.

In parallel, TUD has developed a RL agent to optimize ferric chloride dosing ($FeCl_3$), specifically for the coagulation stage. This agent leverages real-time data from the plant's Supervisory Control and Data Acquisition (SCADA) system, combined with turbidity predictions from a supervised learning model, to recommend the optimal $FeCl_3$ dosage for 6-h intervals. The goal is to enhance the effectiveness of coagulation by adjusting dosage levels in response to changing water conditions, ultimately improving treatment efficiency and ensuring consistent water quality. This multi-objective approach enables DWTP to balance efficiency and performance, minimizing chemical usage while maintaining effective water treatment.

1.3 State of the Art and the ToDrinQ Innovation

RL is a subfield of machine learning, alongside supervised and unsupervised learning. While supervised learning focuses on predicting specific outputs, such as numerical values or categorical labels, based on labelled input data, and unsupervised learning aims to discover patterns or clusters within unlabelled data, RL takes a different approach. In RL, the objective is not merely to predict or classify but to train an agent to interact with an environment in a way that maximizes a given reward. The agent learns by trial and error, selecting actions in response to environmental states and receiving feedback in the form of rewards or penalties, which guides its decision-making process over time (Sutton & Barto, 2018).

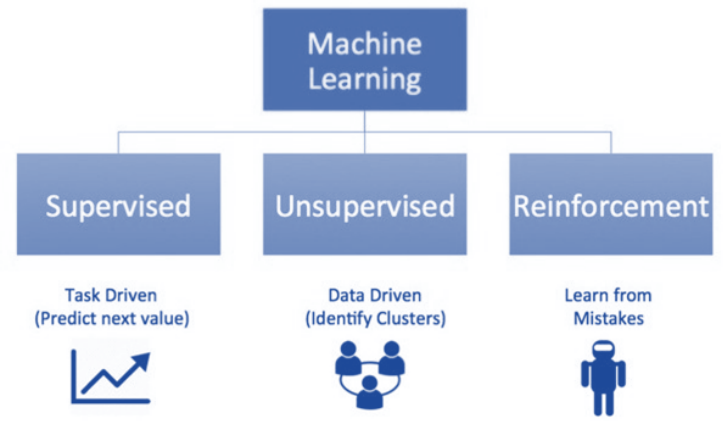

Figure 1: Types of Machine Learning

The RL framework consists of several key components: the agent, the environment, the actions, the state, and the reward signal and it is considered as a Markov Decision Process (MDP) as shown in Figure 2 below. The agent represents the decision-maker, the environment refers to the context or system with which the agent interacts, and the actions are the choices the agent can make at any given state of the environment. The reward signal is a numerical value that indicates the success or failure of an action in reaching the desired outcome. Through repeated interactions, the agent learns to associate actions with rewards and develops strategies, known as policies, that optimize its performance. This makes RL particularly effective for complex decision-making tasks where explicit programming of rules and outcomes is impractical or infeasible.

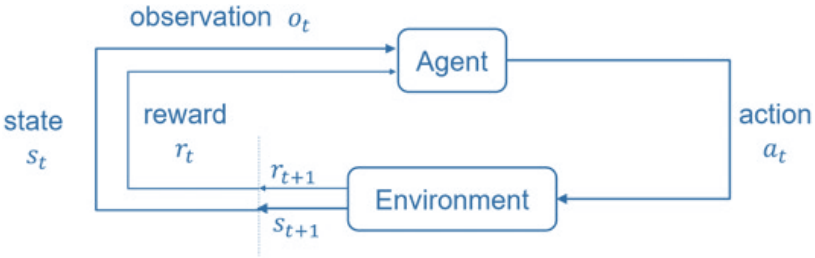

Figure 2: RL algorithm main components

RL has shown remarkable success in various simulated environments, from video games (Mnih et al., 2015a) to board games like Go (Silver, et al., 2016). However, there is increasing evidence that applying RL to real-world problems presents unique challenges, particularly when it comes to the setup of the RL environment. Currently, there is an increasing trend in the employment of RL approaches in real-world problems in various fields such as, self-driving cars (Kiran et al., 2021) and automotive manufacturing systems (Leng et al., 2022; Koch et al., 2023). The RL environment encapsulates everything the agent interacts with, including the dynamics, states, and reward structures. In real-world problems, creating a trustworthy and representative environment is non-trivial (Dulac-Arnold et al., 2019). It requires a deep understanding of the problem domain and how the agent's actions translate into outcomes.

Identifying relevant state representation, in other words determining what parameters or features should constitute the agent's state us is one of the central challenges in applying RL in practice (Gu, et al., 2024). The state must capture all the relevant information needed for decision-making, without being overwhelmed by unnecessary or redundant details. In real-world problems, the state space can be very large, continuous, noisy, or partially observable, complicating the task. In addition, designing an appropriate reward function is crucial and difficult. A poorly defined reward function can lead to undesirable or unintended behaviours, such as agents optimizing for speed at the expense of safety or quality.

In general, real-world applications of RL face several critical challenges related to data efficiency, safety, uncertainty, transferability, and interpretability. RL algorithms typically require large amounts of data, but gathering this data in practical environments can be costly or infeasible (Mnih et al., 2015b). Ensuring the safety of RL agents during exploration, particularly in critical environments like healthcare or autonomous systems, adds complexity (Amodei et al., 2016). Additionally, real-world environments often exhibit uncertainty and randomness, making it difficult for RL models to handle unpredictable variables, which can reduce their robustness (Plappert et al., 2018).

Another major issue is that RL algorithms struggle to generalize, or transfer learned policies across different tasks or environments, limiting their practical applicability (Zhu et al., 2020). Furthermore, RL models, especially those based on deep learning, lack interpretability, making it hard to explain their decisions in critical applications (Puiutta & Veith, 2020). This lack of transparency can hinder trust and wider adoption in fields requiring accountability, such as finance or defence. Addressing these challenges is essential for the successful integration of RL into real-world systems.

In the field of drinking water treatment, the application of RL remains in an early stage of development. While there exists a substantial body of scientific literature focused on the application of various supervised learning algorithms, particularly in the prediction and optimization of key process parameters, such as the formation of disinfection by-products (Peleato et al., 2018; Singh & Gupta, 2012) or the determination of coagulant dosage (Gagnon et al., 1997; Gomes et al., 2015; Griffiths & Andrews, 2011; Haghiri et al., 2018; K. Zhang et al., 2013),the research on the application of RL is comparatively limited.

To date, there are only a few published studies that explore the use of RL within drinking water treatment systems. A recent publication presents a methodology for developing a decision support system (DSS) that recommends optimal coagulant and chlorine dosages. This system models the key stages of drinking water treatment (coagulation, sedimentation, filtration, and disinfection) by employing reinforcement learning techniques (Álvarez Díez, et al., 2024). In contrast, the majority of RL-related research has been directed towards wastewater treatment, where RL has been more extensively studied for its potential to optimize complex, dynamic treatment processes (Croll et al., 2023; Mohammadi et al., 2024; Syafiie et al., 2011). The limited scope of RL applications in drinking water treatment can be attributed to several factors, including the high safety standards, less operational variability, regulatory constraints, and the need for robust, real-time optimization, which collectively make the adoption of experimental and trial-based learning methods such as RL more challenging.

The revised EU Drinking Water Directive necessitates a transition from reactive compliance to a proactive risk-based approach, making Reinforcement Learning essential for predicting and mitigating hazards like turbidity spikes before they impact water safety. RL algorithms are capable of handling the Directive's stricter limits on emerging pollutants and disinfection by-products by optimizing complex chemical dosing trade-offs in real-time, a task that is difficult for human operators to balance simultaneously. Additionally, these data-driven models significantly enhance operational resilience against climate change by processing vast amounts of sensor data to adapt treatment processes faster and more accurately during extreme weather events.

To bridge this gap while addressing the safety and data efficiency challenges outlined above, ToDrinQ has adopted an Offline (Batch) Reinforcement Learning approach. Rather than training agents on the live plant (which is not feasible for safety reasons) or relying solely on simulations, the project utilizes historical SCADA and logbook data to reconstruct the environment.

By employing advanced algorithms such as Twin Delayed Deep Deterministic Policy Gradient with Behaviour Cloning (TD3-BC), the project ensures that the agents learn optimal policies that remain within the safe operational bounds defined by historical expert operator behaviour. This approach allows ToDrinQ to deliver the benefits of AI-driven optimization (such as reduced chemical consumption and improved resilience) without compromising the safety of the drinking water supply. Each proposed solution is carefully adapted to the unique circumstances and requirements of individual demo cases, ensuring that the RL approaches are both contextually appropriate and practically effective. This tailored methodology allows for a more precise application of RL techniques to optimize various aspects of water treatment processes.

2. Reinforcement Learning models for Athens Demo Case #2

2.1 Demo case description, Polydendri DWTP

As a Demo Case (DC) for developing the RL model the Polydendri DWTP was selected (DC#2). The plant situated 237 meters above sea level, is located north of Athens and was constructed in 1986. The Polydendri Water Treatment Plant, located near the Athens-Lamia National Road by the Afidnes toll station, consist of two identical units, each with a nominal capacity of 100,000 cubic meters per day. Water is drawn from the Yliki-Marathon aqueduct, allowing for the treatment of water from various alternative raw water sources, including Mornos, Yliki, and the groundwater wells of Mavrosouvala, Viliza, and Yliki.

The Polydendri DWTP employs a multi-stage treatment process to ensure water quality and safety. The primary treatment begins with chemical oxidation and pre-disinfection using chlorine, followed by coagulation and sedimentation in two pulsator tanks, one dedicated to each processing line. This is succeeded by filtration through eight Aguazur sand filters, which further remove particulate matter. The final stage involves post-disinfection to adjust and correct free chlorine levels before distribution. The treated water is then stored in two large reservoirs with capacities of 33,000 and 27,000 cubic meters, respectively, ensuring adequate storage for subsequent supply. An aerial overview is provided in the Figure 3 below.

Figure 3: Polydendri DWTP

2.2 Challenges

2.2.1 Mapping of the DWTP and Data retrieval

As preliminary but very critical step in designing and developing the RL models for Polydendri DWTP demo case was the comprehensive mapping of the primary processes and the identification of critical monitoring parameters at the processing plant. This task was undertaken in close collaboration with the plant's personnel to ensure alignment with operational realities. The treatment process at the Polydendri DWTP, as shown in Figure 4, begins with a common pre-disinfection/oxidation stage, utilizing chlorine as the primary disinfectant, with an option to substitute sodium hypochlorite. This initial disinfectant

dosage, along with the flow rate, is controlled and recorded through the SCADA system. Meanwhile, turbidity and temperature of the raw, untreated water are monitored manually with handheld instruments every three hours. Following the pre-disinfection, the water undergoes a common screening stage and subsequently diverges into two separate treatment lines for further purification.

Each treatment line comprises several stages: a de-gritting stage, a coagulation/sedimentation stage, sand filtration, and post-disinfection. Key parameters in these stages, such as coagulant (aluminium sulfate), poly-electrolyte, and disinfectant dosages, are consistently logged in the SCADA system for monitoring and adjustment. After treatment, the purified water is stored in two distinct clear water tanks. An optional disinfection treatment is available at the outlet of these tanks for maintaining optimal free chlorine levels. Free chlorine concentrations are manually measured every three hours at the de-gritting stages, coagulation tank inlets, and filter inlets, while online sensors continuously monitor free chlorine levels at the inlet and outlet of the clear water tanks. Additionally, the turbidity of the treated water is measured to ensure quality control. Manual turbidity measurements are taken at the filter inlets, while online sensors continuously monitor this parameter at the outlets of the clear water tanks.

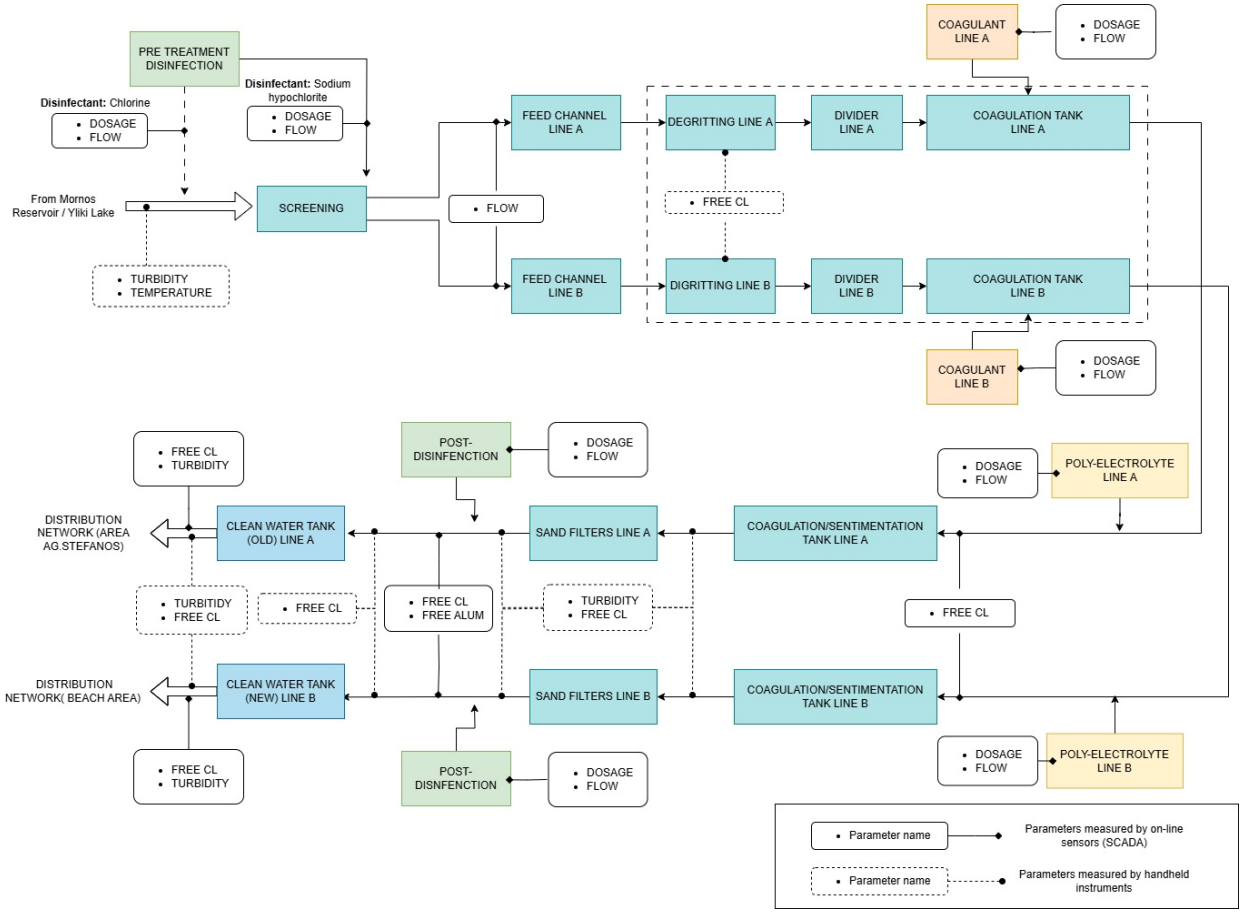

Figure 4: Polydendri DWTP main treatment processes and monitored parameters map

Three key processes were selected as appropriate use cases for the development of RL algorithms: **pre-disinfection**, **coagulation/flocculation**, and **post-disinfection**. Following a thorough transfer of domain knowledge from the DWTP operators, the essential parameters for each of these processes were identified. Further details about these parameters are provided in the dedicated sections for each respective process.

2.2.2 Data sources and data retrieval

The subsequent step focused on assessing the availability of data. As previously mentioned, some critical parameters were manually collected using handheld instruments and recorded in handwritten logbooks. Additionally, certain parameters that, according to scientific literature (Gagnon et al., 1997; Q. Zhang & Stanley, 1999), play a crucial role in the selected processes, such as pH and total trihalomethanes (THMs), are measured once daily by EYDAP's Chemistry Laboratory. In addition, the parameters of the untreated water are monitored along the water transportation network using online sensors, and the corresponding measurements are accessible through an online platform.

As an initial step, a one-month dataset corresponding to November 2018 was retrieved from the SCADA system of the DWTP. In order to improve data completeness and compensate for missing or non-recorded variables, the corresponding operational logbooks were manually inspected, scanned, and digitized. This combined effort resulted in a dataset consisting of approximately 240 data points.

Despite this initial effort, the dataset exhibited significant limitations. Drinking water treatment processes are characterized by high operational stability and low variability, as they are typically designed to operate under steady-state conditions with limited fluctuations. Consequently, the data collected over a one-month period were relatively sparse in terms of system dynamics and variability. This sparsity limited the ability to adequately capture transient behaviors or rare operational conditions. In addition, parameters measured by the EYDAP Chemistry Laboratory were not included in this initial dataset, primarily due to their low sampling frequency, which typically consisted of one measurement per day. Such temporal resolution was not compatible with the higher-frequency data required for dynamic modeling and reinforcement learning applications.

The initial dataset served as the foundation for the development of preliminary RL use cases, as presented in Deliverable D.4.2. Given the absence of real-time interaction capabilities and the limited dataset, the analysis focused on offline reinforcement learning approaches, where policies are derived from pre-collected historical data. Within this framework, considerable effort was devoted to the formulation of the MDP, including the selection and combination of measured variables to define the state space, as well as the design of reward functions aligned with operational objectives such as process efficiency and water quality.

Furthermore, different methodologies for evaluating RL performance in an offline setting were explored. However, the limited size and variability of the dataset imposed significant constraints on the learning process. In particular, the lack of sufficient state-action coverage hindered the development of robust and generalizable policies, and increased the uncertainty associated with policy evaluation.

In response to these limitations, a more extensive data acquisition campaign was undertaken. Multiple visits to the DWTP were conducted, during which approximately 900 pages of operational logbooks were photographed. These logbooks contained tabulated records comprising 28 variables and spanned a period of approximately 15 months of plant operation. The collected material was subsequently digitized through a manual process, substantially increasing both the temporal coverage and the dimensionality of the dataset.

In parallel with this effort, additional data sources were incorporated to further enhance the dataset. Extended historical data were retrieved from the SCADA system, while supplementary operational information was obtained from the EYDAP Conveyance and Transportation monitoring system. Periodic measurements from the EYDAP Chemistry Laboratory were also included despite their lower frequency, as they provided valuable information on water quality parameters. Moreover, meteorological data were

integrated using the ERA5 global climate reanalysis dataset, enabling the inclusion of environmental variables such as temperature and precipitation for the geographical area of the Polydendri DWTP.

The expansion and integration of multiple data sources resulted in a significantly more comprehensive dataset, which improved the representation of system dynamics and operational variability. This enriched dataset enabled a more realistic formulation of the MDP, allowing for a more accurate definition of state variables and reward functions. Additionally, the increased data volume and diversity enhanced the reliability of offline policy evaluation methods and reduced uncertainty in the learning outcomes. Overall, this process demonstrated that the successful application of reinforcement learning in drinking water treatment systems is critically dependent on the availability of sufficiently large and diverse datasets. The initial findings clearly indicated that limited datasets are inadequate for capturing the complexity of such systems, thereby necessitating extensive data collection and integration efforts. A detailed summary of all data sources and their characteristics is provided in Table 1. Moreover, due to technical issues, a significant portion of the historical SCADA data could not be retrieved. Consequently, for two out of the three RL use cases, the core datasets were derived primarily from the digitized handwritten logbooks. To address these limitations in data availability, we adopted not only an offline RL approach but also an online training methodology, as described in the following section.

Table 1: Available data sources for DC#2

Data source	Parameters of interest	Measurement acquisition mean	Rate of recording	Data points	Period coverage
DWTP SCADA	• Flow rate • Turbidity • Free Cl concentration • Free Alum concentration • Coagulant/Disinfectant dosages	Online sensors	Per second	Approx. 31.104.000	1/1/2024 – 5/4/2024 (Flow rate only) 1/7/2024 – 30/7/2025 (Disinfectant dosage for post-disinfection, Free Cl at Tank inlet)
DWTP Logbooks	• Flow rate • Turbidity • Free Cl concentration • Coagulant/Disinfectant dosages	Manual sampling/lab measurements	Approximately every 3 hours	3800	1/1/2024 – 9/4/2025 (15 months)
EYDAP Chemistry Lab	• pH • Free Alum concentration • THMs concentration	Manual sampling/lab measurements	daily	273	1/1/2024 – 30/9/2024 (9 months)
EYDAP Conveyance & Transportation monitoring system	• Total Organic Carbon concentration (TOC) • Water Temperature	Online sensors	Every 5 minutes	Approx. 130.000	1/1/2024 – 30/6/2025
ERA5	• Air temperature • Surface net solar radiation	climate reanalysis model	hourly	13.128	1/1/2024 – 30/6/2025

2.2.3 Selecting RL Algorithm

As outlined in the introduction, a key component of any RL algorithm is the environment in which the RL agent interacts and learns through trial and error. However, when applying RL to drinking water treatment processes, significant challenges arise. Real-time training of an RL agent within an operational DWTP is not a feasible option. Beyond the evident safety concerns, the inherently slow and stable nature of water treatment processes makes real-time training impractical and inefficient.

To address these challenges, employing simulation software as an RL environment represents a viable alternative. However, most available simulation tools for water treatment processes are commercially licensed and entail substantial costs, which significantly limit their accessibility. An exception is the open-source software **Stimela** (Van der Helm & Rietveld, 2002), developed in MATLAB/Simulink (Van der Helm & Rietveld, 2002). A comprehensive evaluation of Stimela's capabilities and functionalities was conducted, focusing on the required input parameters for the simulated water treatment processes, specifically chlorination, coagulation, and flocculation. Nevertheless, certain limitations were identified. Although chlorination is simulated in Stimela using chlorine as the disinfectant, the Polydendri DWTP is in the process of transitioning to sodium hypochlorite as its primary disinfectant. As a result, the available historical data related to chlorine usage are relatively limited. Furthermore, the Stimela chlorination module was deemed unsuitable for replicating the post-disinfection process at Polydendri DWTP, due to significant differences in system configuration. In the demo case, for instance, there is no open disinfection tank, and the disinfectant is injected directly into the feed pipe before the inlet of the clear water tank.

Nonetheless, since Stimela was also adopted in WP6 for the development of the design-support tool under Task 6.4 (reported in D6.2) and a Python reimplementation (**pyStimela**) was developed, it was decided to test the incorporation of pyStimela as a supplementary component to the main RL approach. Specifically, pyStimela was integrated into two RL environments, pre-disinfection and coagulation/flocculation, to facilitate online RL training and provide a limited yet meaningful correspondence with the actual treatment characteristics of the Polydendri DWTP. The development of these environments and the training algorithms will be presented in detail in a subsequent section.

To ensure an accurate representation of the Polydendri treatment processes and to account for the limitations of the previously discussed simulation-based approach, a purely data-driven methodology was adopted. In this framework, the RL agent is trained exclusively on historical data, a training paradigm commonly referred to as offline or batch RL. In this alternative approach, the RL environment is effectively represented by the real-world system itself, eliminating the need for simulation software. Historical datasets are used to reconstruct sequences of state transitions, reflecting the recorded operational parameters and actions taken by the DWTP operators. The design of the state representation and the reward function constitute a key component of this methodology representing one of the major challenges to be addressed.

Subsequently, the effectiveness of this methodology is inherently dependent on the availability and quality of historical data, as well as the comprehensiveness of its representation across a wide range of potential scenarios in system behaviour. This ensures that the RL agent not only replicates the decision-making patterns of the DWTP operators but also demonstrates the capacity to generalize to novel and previously unseen conditions. As the volume of historical data continues to grow over time, it is expected that the performance of the RL agent will improve incrementally. This improvement will enable the agent to make increasingly robust decisions and achieve higher levels of system optimization, thereby contributing to more efficient and resilient operational outcomes.

Considering the aforementioned details and the continuous nature of the parameters monitored in the DWTP, a subset of which will constitute the Markov Decision Process (MDP) states in the RL models, a suitable algorithm was selected. Given that dosage levels in each of the three use cases (coagulation, pro-disinfection, and post-disinfection) are also represented as continuous values, it was deemed appropriate to adopt an RL training algorithm from the Deep Deterministic Policy Gradient (DDPG) family (Sewak, 2019). Algorithms within this family are well-suited to environments characterized by continuous action and state spaces, enabling efficient policy learning in contexts where precise control over dosage levels is required.

Subsequently, the RL agent training algorithm that was implemented for the three use cases, was the TD3-BC(Fujimoto & Gu, 2021), which is an offline variation of the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm (Fujimoto et al., 2018). The initial TD3 algorithm is an advanced form of DDPG (Lillicrap et al., 2015), designed for continuous action spaces. TD3 builds on DDPG's actor-critic framework, where the actor learns to select actions that maximize cumulative rewards, and the critic estimates the value of those actions. However, DDPG can suffer from issues like overestimation bias and instability. TD3 addresses this with two main improvements: using twin critics and delayed policy updates. The twin critics help reduce overestimation by having two Q-networks and using the minimum value between them to update the policy, making learning more robust. The delayed updates, where the actor (policy) is updated less frequently than the critic, add stability to the learning process. Additionally, TD3 adds target smoothing by introducing a small noise to the action selection in target networks, which helps the agent explore more smoothly and avoid sharp changes in policy. These refinements make TD3 a powerful algorithm for stable and effective learning in complex environments with continuous control.

The TD3-BC algorithm modifies the standard TD3 framework to improve its performance in offline reinforcement learning settings, where the agent is trained solely on a fixed dataset of prior experiences rather than interacting with the environment. In TD3-BC, an additional behaviour cloning (BC) term is added to the objective function, which encourages the learned policy to stay closer to the actions in the dataset. This is particularly helpful in offline learning, as it prevents the policy from diverging towards out-of-distribution actions that the critic may inaccurately evaluate due to the lack of exploration. A summary of the TD3-BC architecture is presented in Figure 5. Each batch sample from the replay buffer contains the system state at time $t(s_t)$, the corresponding expert action (a_t), the reward received at time $t(r_t)$, and the subsequent state at time $t+1(s_{t+1})$. This collection of transitions forms the basis for training both the actor and critic networks within the TD3-BC framework.

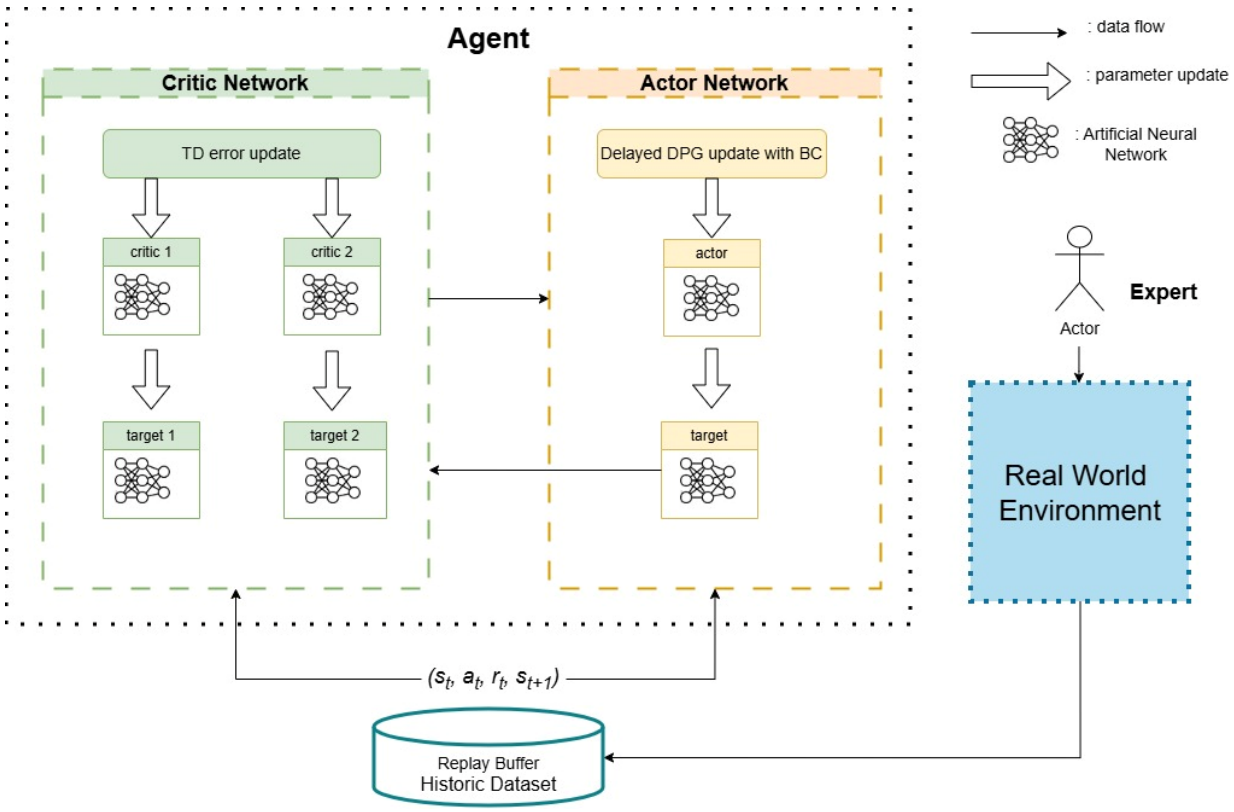

Figure 5: The TD3-BC Architecture

2.2.4 Development of the RL models workflow

Now, with the availability of a substantially larger and more diverse dataset in relation with the beta versions on D4.2, significantly more opportunities emerged for systematically exploring the core components and workflow of RL. The increased temporal coverage and improved representation of operational variability enabled a more rigorous investigation of alternative MDP formulations, including different configurations of state variables, more expressive reward function structures, and more reliable offline evaluation strategies. In contrast to the initial phase, where data sparsity constrained both model development and validation, the enriched dataset provided sufficient state–action coverage to support more robust and realistic RL training, reducing uncertainty and improving the generalizability of the learned policies.

Once the RL training algorithm has been established, the subsequent and often most demanding task involves the definition of the RL environment for each use case. This stage entails specifying the fundamental elements of the MDP: the state space, action space, and reward function. Among these components, the action space is typically the most straightforward to define. In the context of drinking water treatment processes, such as chlorination and coagulation/flocculation, the actions correspond directly to the operational dosage setpoints of disinfectants or coagulants.

Conversely, the definition of the state space presents a considerably greater challenge. It requires the careful selection of process variables that adequately represent the system's dynamic behaviour and provide the RL agent with sufficient information to make effective control decisions. This selection process is highly nontrivial, as it demands both extensive experimentation and substantial domain expertise. The inclusion of irrelevant or redundant measurements can degrade learning performance,

while an incomplete state representation may prevent the agent from capturing essential system dynamics.

Equally challenging is the design of an appropriate reward function, which governs the agent's learning objectives. Water treatment operations can be formulated as continuous control problems, in which specific process and water quality parameters must remain within regulatory and operational limits. Accordingly, the reward function should assign positive rewards when these constraints are satisfied and penalties when deviations occur. In addition to maintaining compliance and process stability, operational optimization objectives, such as minimizing chemical usage, must also be embedded in the reward design. Consequently, the overall reward function generally comprises two principal components: (i) a constraint satisfaction term, which ensures adherence to water quality and safety standards, and (ii) an efficiency term, which promotes reduced chemical consumption and improved process economy. These components are typically combined through a weighted sum, where the weighting factors determine the relative importance of safety versus efficiency objectives. The selection and tuning of these weights are crucial and usually require systematic experimentation to achieve an appropriate balance between stable control performance and process optimization.

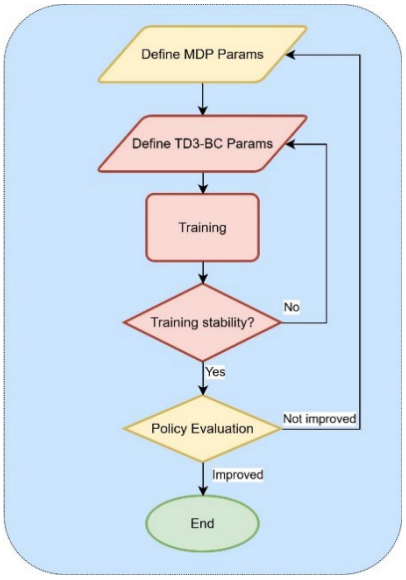

Figure 6: High level RL development process

The determination and tuning of the hyperparameters of the TD3-BC algorithm constitute another critical and nontrivial stage in developing an effective RL control framework for water treatment applications. TD3-BC extends the standard TD3 algorithm by incorporating a behaviour cloning term into the actor update to enhance stability and mitigate overestimation bias when learning from limited or noisy data. The key parameters that require definition include the learning rates of the actor and critic networks (which control the speed of policy and value function updates), the discount factor (γ) governing the weight assigned to future rewards, the target smoothing coefficient (τ) determining the rate of soft updates in target networks, the policy noise and noise clipping parameters (which stabilize exploration during training), the batch size for gradient updates, and the behaviour cloning coefficient (α) controlling the trade-off between imitation of prior data and autonomous policy improvement.

Tuning these parameters serves primarily to ensure training stability and policy convergence while preventing phenomena such as divergence, oscillatory learning, or overfitting to suboptimal policies. Monitoring diagnostic metrics such as the critic loss, actor loss, and average episodic reward throughout training is essential for assessing learning progress and detecting instabilities in value estimation or policy updates. Optimal parameter configuration typically requires systematic experimentation and sensitivity analysis, as the interactions between parameters are complex and context-dependent, particularly in continuous process control domains characterized by nonlinear and time-varying process dynamics, such as drinking water treatment operations.

In this study, TD3-BC parameter determination was performed through a combination of random experimentation and domain-informed reasoning (Park, et al., 2018) (Álvarez Díez, et al., 2024), using exclusively historical process data. Because no prior benchmarks existed for this specific application and the absence of a simulated environment limited the scope for controlled testing, a systematic grid search was not feasible. Instead, a series of exploratory experiments were conducted to evaluate the effect of

different parameter configurations on training stability, convergence behaviour, and overall policy performance. A brief sensitivity analysis was also performed to examine the influence of key TD3-BC hyperparameters on learning stability and performance. By systematically varying selected parameters and observing changes in the critic loss, actor loss, and episodic reward, this analysis provided insight into the robustness of the training process and guided the refinement of the final parameter configuration using historical process data.

After achieving training stability with the TD3-BC algorithm, the next phase focused on the afore mentioned selection and refinement of the MDP components. This stage involved comparing the learned policy with the existing operational policy used by the DWTP operators. Once stable learning behaviour was established, (see Figure 6) attention returned to the primary research challenge, namely, the specification of an appropriate state space and the determination of the weighting factors for both the constraint satisfaction and efficiency terms within the reward function. To evaluate the suitability of the selected MDP structure and reward formulation, a benchmarking procedure was required. Ideally, this would involve a direct comparison between the learned policy and the existing control strategy in the actual DWTP environment. However, such an approach was not feasible due to operational constraints and the absence of a fully accurate process simulation. Consequently, an alternative evaluation strategy was adopted.

Following the methodology proposed by (Voloshin, Le, Jiang, & Yue, 2021) Approximation Models (AMs) were developed for each reinforcement learning scenario. These models approximate the dynamics of the real DWTP processes, thereby enabling a comparative analysis between the learned and existing policies using portions of historical operational data. In this framework, both policies are applied to identical MDP states: the agent's selected action is input to the AM, which generates the subsequent state and associated outcomes, while the corresponding historical (operator) action is also applied to yield the next state and reward. By aggregating performance across multiple state transitions, the average effectiveness of the learned policy is quantitatively compared to that of the existing policy. This systematic evaluation provides insights into the influence of different state representations and reward-weighting configurations, following the principles outlined by (Tang & Wiens, 2021).

2.3 RL model #1: Optimization of pre-disinfection stage

2.3.1 Problem statement

The pre-disinfection and oxidation stage at the Polydendri DWTP is implemented immediately downstream of the main intake structure (SCREENING, as illustrated in Figure 4), where sodium hypochlorite is employed as the primary disinfectant. Until recently, chlorine had been used for this purpose; however, the plant is currently transitioning toward the exclusive use of sodium hypochlorite. The disinfectant solution is introduced at the point where the raw water flow is divided into two parallel treatment lines. Maintaining appropriate levels of free chlorine in the subsequent treatment stages is essential for ensuring effective disinfection. These levels are monitored at two key locations, specifically,

Table 2: Pre-disinfection recorded parameters

Parameter Name	Parameter Type	Units	Resolution	Low limit	High limit	Recording frequency	Source
Raw water flow	External Variable	m^3/hour	1	-	-	1 minute	SCADA
Raw water Turbidity	External Variable	NTU	0.1	-	-	3 hours	Logbook
Raw water	External Variable	µS/cm	0.1	-	-	5 minutes	Conveyance

Parameter Name	Parameter Type	Units	Resolution	Low limit	High limit	Recording frequency	Source
conductivity							Monitor System
Raw water TOC	External Variable	mg/lt	0.01	-	-	15 minutes	Conveyance Monitor System
NaOCl dosage	Manipulated Variable	ppm	0.1	-	-	3 hours	Logbook
Free Cl de-gritting	Constraint Variable	mg/lt	0.01	0.7	1.2 (Summer) 0.8 (Normal)	3 hours	Logbook
Free Cl filters	Constraint Variable	mg/lt	0.01	0.2	0.4	3 hours	Logbook
Air temperature	External Variable	Celsius	0.1	-	-	1 hour	ERA5
Surface Net Solar radiation	External Variable	J/m2	1	-	-	1 hour	ERA5

the de-gritting stage and the filter inlet, through manual sampling and on-site measurements using handheld instruments at three-hour intervals. The concentration of free chlorine, expressed in mg/L, must remain within predefined operational limits (Table 2) to guarantee adequate disinfection performance. At present, the disinfectant dosage is determined in a reactive manner by DWTP operators, based primarily on the influent water flow rate and the free chlorine concentrations recorded at the downstream monitoring points. The objective of the RL algorithm is to optimize the disinfectant dosage, expressed in parts per million (ppm), to consistently maintain free chlorine concentrations within the prescribed limits, thereby enhancing both process efficiency and water safety.

2.3.2 Data sources and data preprocessing

For the pre-disinfection RL use case, the primary data source comprised the operational logbooks of the Polydendri DWTP, as the measurements recorded therein were deemed more accurate and reliable than those obtained from the online sensors. These logbook measurements also serve as the primary reference for operators when determining disinfectant dosage adjustments. From the SCADA system, only data related to untreated water flow were utilized. Compared to the beta version presented in D4.2, additional information was obtained from the EYDAP conveyance and transportation monitoring system. Furthermore, meteorological variables, such as air temperature and cloudiness, were incorporated into the dataset, as these factors have been reported in the literature to influence the chlorination process, particularly when disinfection occurs in open tanks. A comprehensive overview of all variables employed in this study is presented in Table 2.

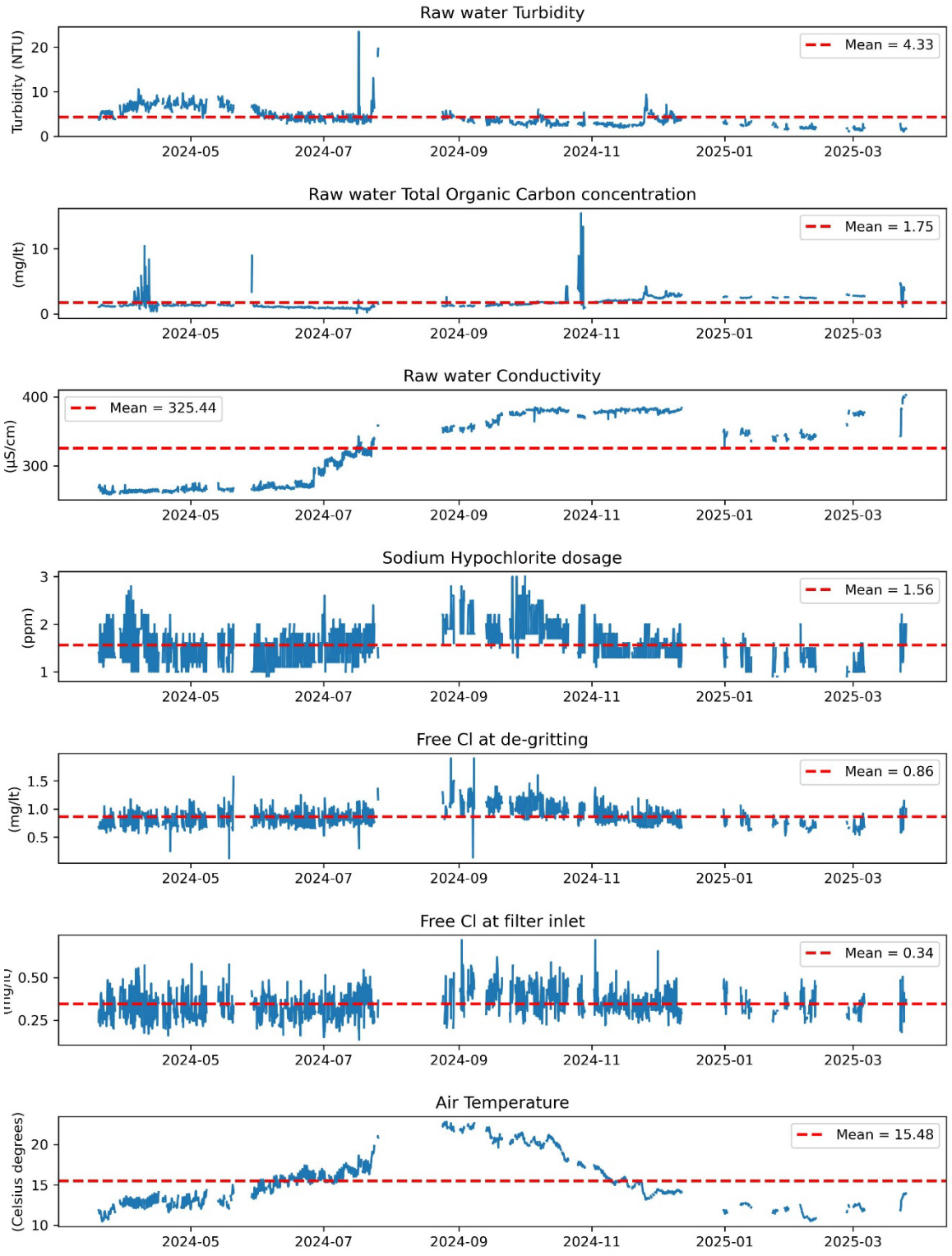

Figure 7: Pre-disinfection recorded parameters behavior over time.

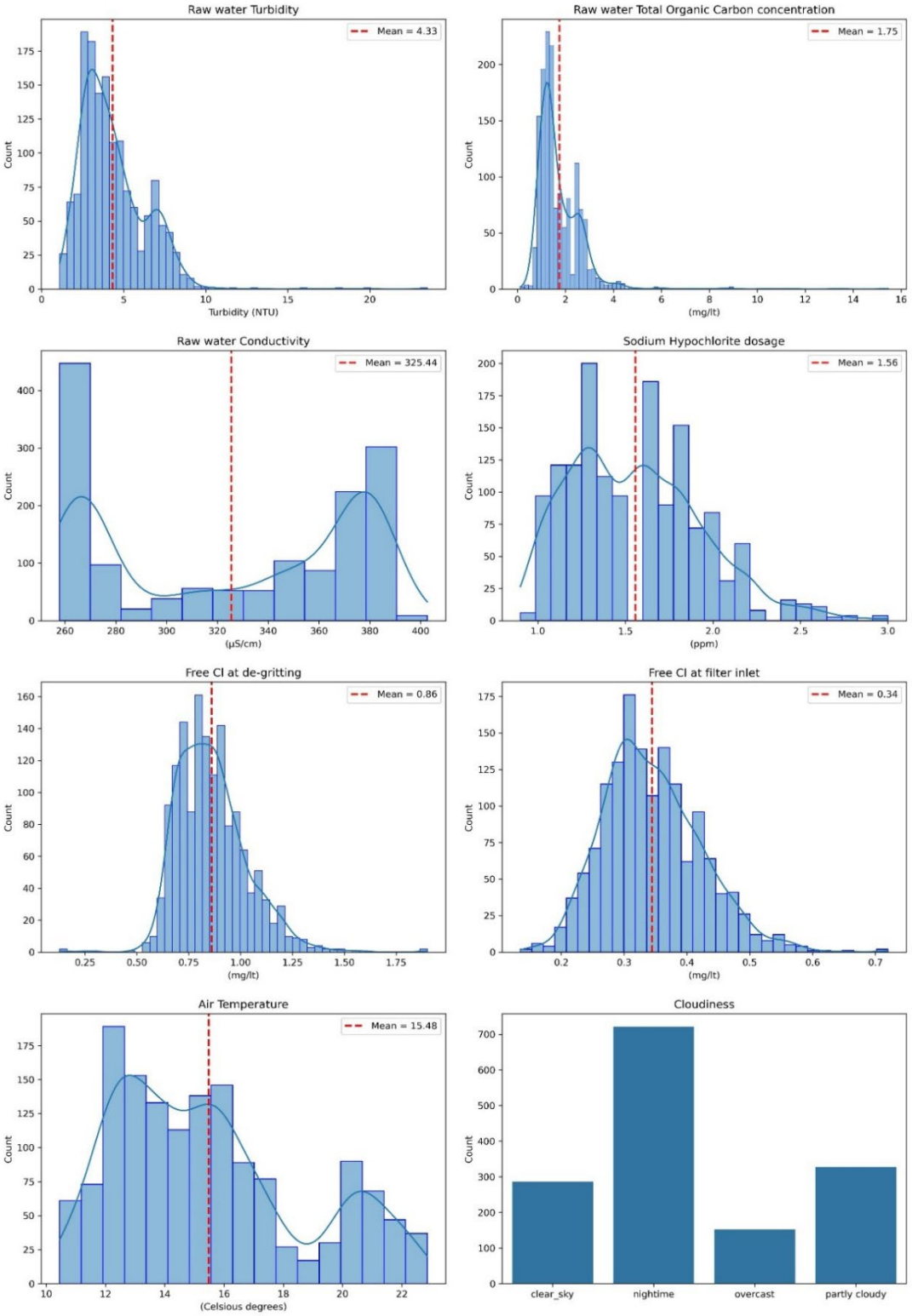

Figure 8: Exploratory Analysis of the pre-disinfection dataset.

Raw water turbidity, conductivity, TOC and flow rate are categorized as external variables, as their values are influenced by external environmental factors and remain unaffected by the pre-disinfection process. The same can be said for the weather-related variables, air temperature and Surface Net Solar radiation. In contrast, chlorine (Cl) dosage is classified as a manipulated variable, as it directly impacts the pre-disinfection process and can be adjusted to control the system. Furthermore, free chlorine concentration values at the de-gritting stage and at the filter inlet are identified as constraint variables, as these values must be maintained within defined safety limits to ensure smooth and safe operation of the treatment process.

A three-hour time interval was adopted as the main temporal resolution, and all remaining measurements were aggregated by calculating their mean values within this window. To characterize sky conditions over the Polydendri area, cloudiness was estimated using the Surface Net Solar Radiation (SSR) variable from the ERA5 reanalysis dataset. The degree of cloudiness was expressed through the clearness index (K_t), calculated as the ratio of surface irradiance to top-of-atmosphere (TOA) irradiance, with the latter determined from the solar constant and the solar zenith angle. Low K_t values (< 0.3) correspond to overcast conditions, intermediate values (0.3–0.55) indicate partial cloud cover, and high values (> 0.55) represent clear-sky conditions. This method offers a straightforward yet physically grounded means of estimating cloud cover based on radiation data.

The final dataset consists of a total of 1,789 data points, each recorded at a three-hour time interval, corresponding to approximately 7.5 months of plant operation covering the period from 19 March 2024 to 25 March 2025. The reduction in the overall temporal coverage was necessitated by the occasional use of chlorine as a disinfectant during certain operational periods. Since the RL framework developed in this study was specifically designed to model the sodium hypochlorite disinfection scenario, only the corresponding time intervals were retained for further analysis.

A detailed exploratory analysis of the main process variables included in the dataset is presented in Figures 7 and 8, providing an overview of their distribution and variability throughout the selected operational period. The gaps observed in Figure 7 correspond to operational periods during which the DWTP employed chlorine as the primary disinfectant. Apart from these observations, and with the exception of a high turbidity event recorded at the end of July 2024 and a pronounced increase in Total Organic Carbon (TOC) concentrations observed toward the end of October of the same year, the collected dataset demonstrates relatively smooth temporal variation, indicative of stable and well-regulated process conditions. From the initial dataset, a new dataset of MDP state transitions was subsequently constructed. Each data point in this derived dataset represents a complete transition tuple, consisting of the current MDP state (comprising a superset of all relevant process variables), the action executed (i.e., the applied dosage setpoint), and the resulting next MDP state.

This MDP transition dataset was then divided into two subsets: a training dataset comprising 1,489 samples and an evaluation dataset consisting of 300 samples. The training dataset was used for training the RL agent and for optimizing the hyperparameters of the TD3-BC algorithm. Conversely, the evaluation dataset was employed to assess the performance of the learned policy and to refine the formulation of the optimal MDP components, namely the state representation and the reward function.

2.3.3 Materials and methods

With a more extensive dataset—both in terms of data sources (logbooks, SCADA, conveyance system, and meteorological data) and a longer temporal coverage compared to the beta version presented in D.4.2—a more representative MDP state was formulated, along with a new reward function was introduced. As previously discussed, the formulation of the MDP components, particularly the definition of the state space and the reward function, represents a major challenge in applying RL to real-world systems.Through an extensive review of the relevant literature (Kulkarni & Chellam, 2010) (Park, et al., 2018), consultation with domain experts, careful consideration of the available recorded process variables, and a series of exhaustive experimental evaluations, the final structure of the MDP state was defined. The resulting state representation, which encapsulates the most relevant system dynamics and operational conditions, is presented below:

Table 3: MDP state parameters of Pre-disinfection RL

Pre-Disinfection MDP state parameters	
1. Raw water flow rate (m^3/hour)	8. Free Chlorine at Filters inlet 6 hours prior (mg/lt)
2. Raw water conductivity (μS/cm^2)	9. Dosage set point 3 hours prior (ppm)
3. Raw water TOC (mg/lt)	10. Dosage set point 6 hours prior (ppm)
4. Free Chlorine at De-gritting(mg/lt)	11. Air temperature (Celsius)
5. Free Chlorine at De-gritting 3 hours prior (mg/lt)	12. Hour of the day (0-23)
6. Free Chlorine at Filters inlet (mg/lt)	13. Season of year (Autumn, Winter, Spring, Summer)
7. Free Chlorine at Filters inlet 3 hours prior (mg/lt)	14. Sky state (Clear, partly cloudy, overcast, night)

The reward function was formulated based on two principal elements: the applied disinfectant dosage, which represents the control action of the RL agent, and the measured free chlorine (FCl) concentrations at the de-gritting and filter inlets, which constitute part of the subsequent MDP state (i.e., the system state observed three hours after the action is executed). This formulation establishes a direct relationship between the operational decision (dosage) and its delayed effect on the system's disinfection performance. The mathematical expression of this relationship is defined as follows:

$$R = -\left(w_1 \cdot FreeCl_{R_{Degrit}} + w_2 \cdot FreeCl_{R_{Filters}} + w_3 \cdot Dosage\right)$$

where:

$\boldsymbol{FreeCl_{R_{degrit}}}$: is a quadratic penalty for the exceeding the Free Cl constrains at De-gritting,

$\boldsymbol{FreeCl_{R_{filters}}}$: is a quadratic penalty for the exceeding the Free Cl constrains at filter inlet,

$\boldsymbol{Dosage}$ **:** the sodium hypochlorite dosage set point.

The term $w_1 \cdot FreeCl_{R_{Degrit}} + w_2 \cdot FreeCl_{R_{Filters}}$ represents the constraint satisfaction component of the reward function, reflecting the compliance of the system with the target free chlorine concentration levels at the de-gritting and filter inlets. Conversely, the term $w_3 \cdot Dosage$ corresponds to the efficiency component, which penalizes excessive chemical usage and promotes operational optimization. The weighting coefficients w_1, w_2, w_3 were determined through extensive experimental evaluation and sensitivity analysis, as described in the previous section.

As described in Section 2.2.4, once training stability was achieved through the systematic tuning of the TD3-BC algorithm parameters, the policy evaluation phase was conducted. This step involved assessing the performance of the learned policy in comparison with the existing operational policy, using the evaluation hold-out dataset.

For this purpose, an approximate modeling framework was implemented, comprising two independent XGBoost models developed to predict the FCl concentrations at the de-gritting and filter inlets, variables representing the key process constraints in the use case. Both models were trained on the initial dataset, which was partitioned into 80% for training and 20% for testing. The input variables employed for each predictive model, along with the corresponding performance metrics, are summarized in Table 4, with as a parameter mean absolute error (MAE).

Table 4: Pre-disinfection evaluation AM

method	**Input parameters** (from MDP state)	**Output Parameter** (from next MDP state)	**MAE**
XGBoost	1. Dosage 2. Dosage 3 hours prior 3. FCl de-gritting 4. FCl de-gritting 3 hours prior 5. Flow rate 6. Air temperature 7. TOC 8. Season 9. Cloudiness 10. Hour of day	**FCl de-gritting**	**0.06 (mg/lt)**
XGBoost	1. Dosage 2. FCl de-gritting 3. FCl filters inlet 4. Flow rate 5. Air temperature 6. Conductivity 7. Season 8. Cloudiness 9. Hour of day	**FCl filters inlet**	**0.04 (mg/lt)**

During the evaluation phase, each historical MDP state transition from the evaluation dataset was used to estimate the subsequent FCl concentration values by applying the actions derived from both the learned and the existing policy. The predicted FCl values, together with the respective dosage values associated with each policy, were then used as inputs to the reward function to quantify and compare the overall policy performance. Furthermore, this approximate modeling approach enabled a relatively reliable estimation of potential constraint violations associated with these critical parameters, thereby providing valuable insights into the trade-off between operational efficiency and compliance with disinfection performance requirements. The overall reinforcement learning training performance is presented in the following section.

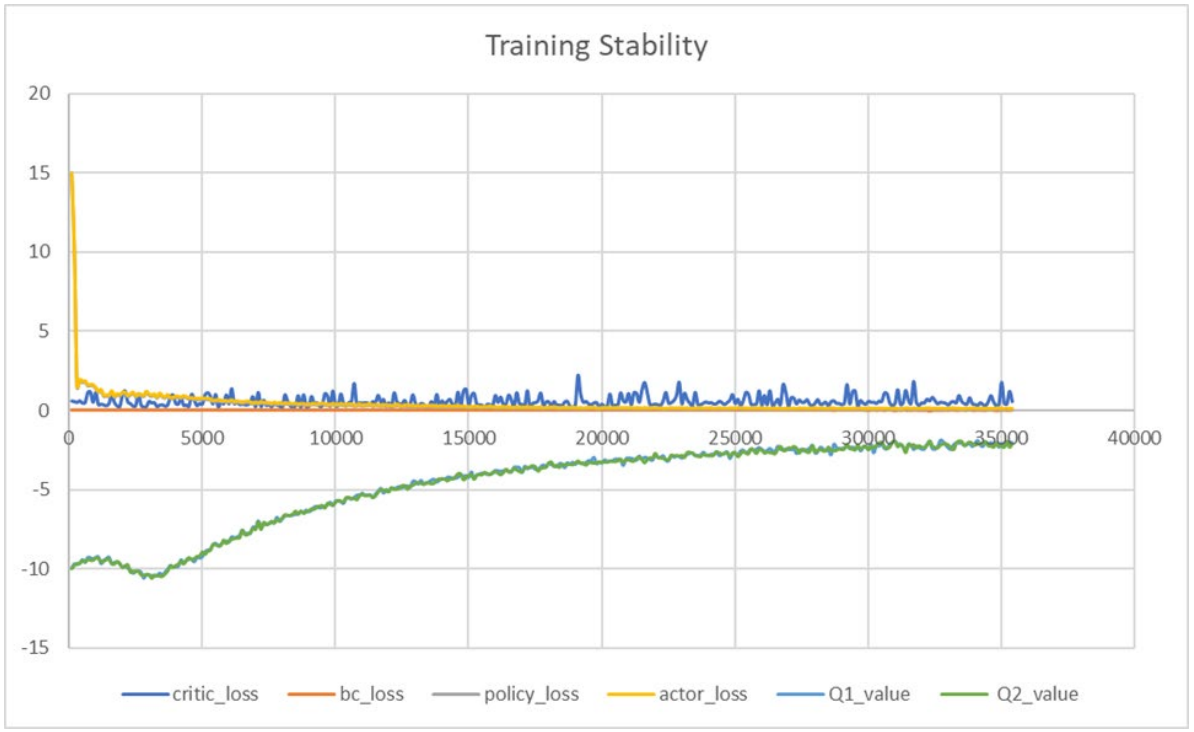

Figure 9: Actor & Critic Neural networks parameters evolution during 35000 training batches.

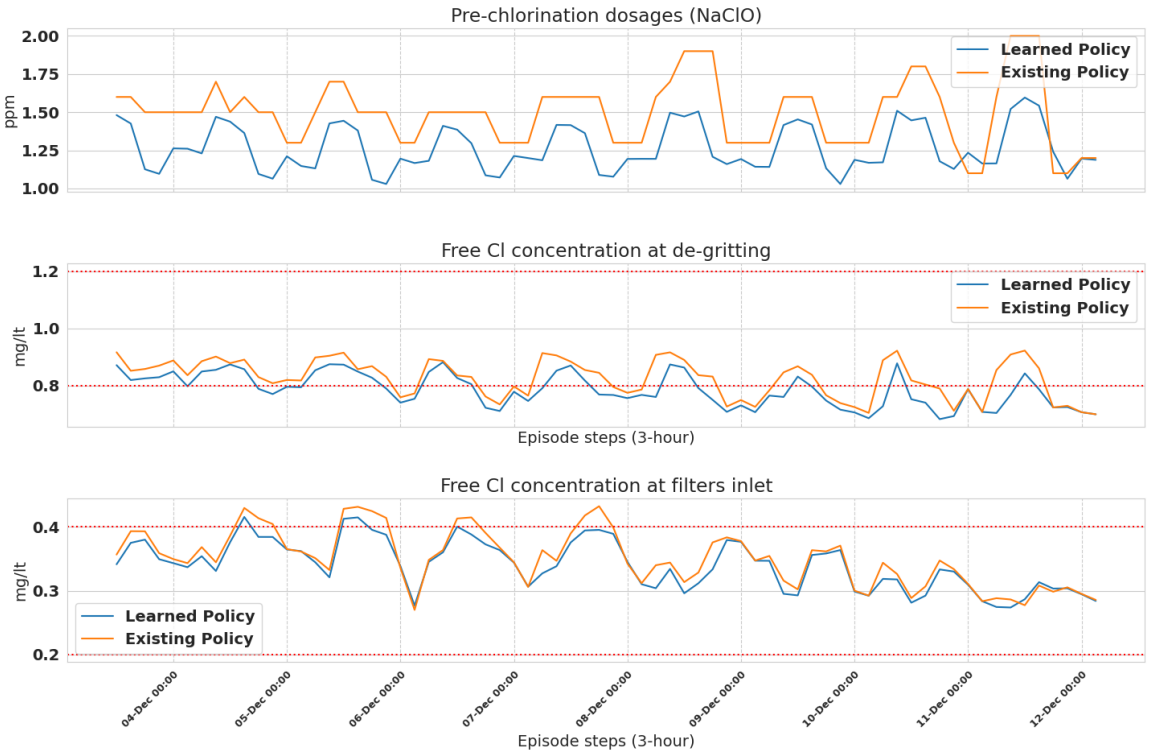

Figure 10: Learned Policy performance on a 70-step episode

2.3.4 Results

After extensive experimentation with the TD3-BC algorithm parameters, a satisfactory level of training stability was achieved. As illustrated in Figure 9, all critical parameters converged toward stable values, indicating robust learning dynamics, with only a minor exception observed in the actor loss, which exhibited slight oscillations. Furthermore, the RL agent demonstrated behaviour closely resembling the operational policy typically adopted by the DWTP personnel (Figure 10). Through iterative learning, the agent's decision-making progressively aligned with historical operational patterns, effectively replicating the actions of human operators. Remarkably, this behavioural convergence was accompanied by an approximate 3.8% reduction in disinfectant dosage, suggesting that the RL agent identified subtle process optimizations without compromising safety or operational reliability. Compared to the results of the beta version (5% reduction), this performance estimate can be considered more accurate due to the larger volume of data used and the more accurate evaluation models that were developed. Although modest, this reduction could translate into meaningful improvements in chemical usage efficiency over extended operational periods. A brief statistical analysis of the results obtained from the evaluation dataset, comprising 300 samples, is presented in Figure 11, providing quantitative support for the observed performance trends.

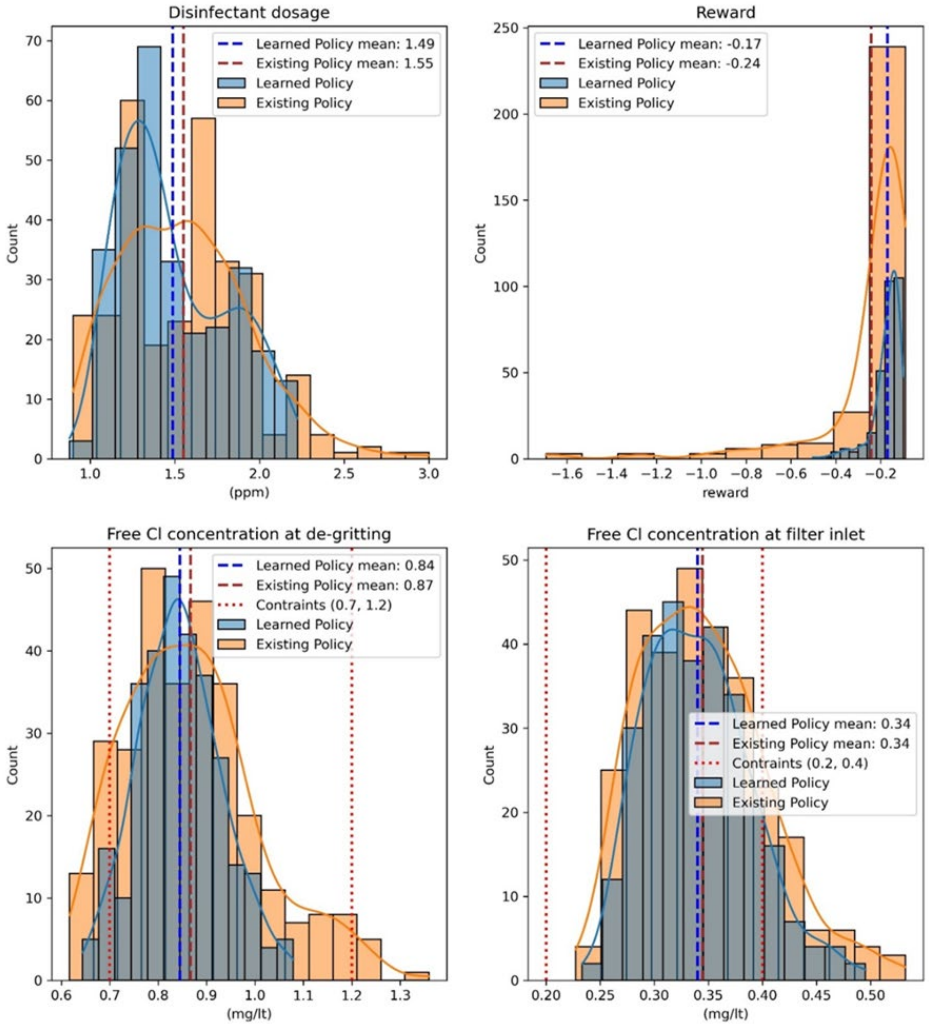

Figure 11: Performance histogram plots of the learned policy over the evaluation dataset

2.3.5 Conclusions and next steps

In summary, the developed TD3-BC–based reinforcement learning (RL) framework demonstrated stable training behaviour and successfully replicated the operational strategies typically employed by DWTP personnel, while achieving a marginal reduction in disinfectant dosage. These findings indicate the potential of RL-based decision-support tools to enhance process efficiency and support data-driven optimization of water treatment operations. It is important to note, however, that the learned policy has not yet been evaluated under real-world operating conditions. Conducting such an evaluation would entail considerable safety and regulatory risks, given the direct impact of disinfectant dosing on water quality and public health. One way to address this limitation, is to exploit the planned integration into the WP7 platform, where it will be made accessible to operators at the Polydendri DWTP. Through this platform, operators will be able to input the relevant process parameters, obtain the RL-suggested dosage set point, and subsequently provide feedback by assigning a performance score to the recommendation. This feedback will be stored in the platform's database, enabling the progressive accumulation of real-world evaluation data. Over time, this iterative human-in-the-loop approach is expected to yield a more representative assessment of the RL agent's performance and facilitate its gradual adaptation toward safe and reliable real-world deployment.

2.4 RL model #2: Optimization of coagulation stage

2.4.1 Problem statement

At the Polydendri DWTP, the coagulation and sedimentation processes, illustrated in Figure 4, are carried out in two Pulsator-type tanks, each corresponding to an independent treatment line. The performance of this stage is primarily evaluated based on two key quality indicators: the turbidity measured at the filter inlets and the concentration of residual (free) aluminium in the treated water. Both parameters must remain within the prescribed limits (Table 5) to ensure compliance with drinking water quality standards. Turbidity measurements are manually performed every three hours at the filter inlets using a handheld turbidimeter. In contrast, the free aluminium concentration was continuously monitored by online sensors installed at the inlets and outlets of the two clear water tanks (Figure 4). However, since 2024, the online aluminium sensor has been out of operation.

Currently, the coagulant dosage at the DWTP is manually adjusted in a reactive manner, based on observed variations in these quality indicators. The objective of the RL implementation in this use case is to develop a proactive and optimized coagulant dosage strategy. The RL-based approach seeks to maintain turbidity at the filter inlets within safe operational limits while minimizing total coagulant consumption, thereby improving both process efficiency and overall sustainability of the treatment operation.

2.4.2 Data sources and data preprocessing

For the coagulation RL use case, the model was developed exclusively from the operational logbooks of the Polydendri DWTP. These records were selected for their higher accuracy and reliability compared to online sensor data and are also the primary reference used by operators for coagulant dosage adjustments. In contrast to the beta version developed in D4.2, where free alum concentration measurements were available from the SCADA system, such data were not accessible in this case due to a sensor malfunction. As a result, the MDP state along with the reward function were modified accordingly compared to the beta version. A summary of all variables employed in this study is provided in Table 5.

Table 5: Coagulation recorded parameters

Parameter Name	Parameter Type	Units	Resolution	Low limit	High limit	Recording frequency	Source
Raw water Turbidity	External Variable	NTU	0.1	-	-	3 hours	Logbook
Aluminium Sulphate dosage	External Variable	ppm	1	-	-	3 hours	Logbook
Free Cl de-gritting	External Variable	mg/lt	0.01	-	-	3 hours	Logbook
Water Turbidity at filters inlet	Constraint Variable	NTU	0.01	-	0.8	3 hours	Logbook

Raw water turbidity and the free chlorine concentration at the de-gritting stage are classified as external variables, as both are determined by upstream or environmental conditions and remain unaffected by coagulation. In contrast, the aluminium sulphate dosage is treated as a manipulated variable, directly influencing process performance, while the filter inlet turbidity is defined as a constraint variable, since it must remain within prescribed operational limits to ensure safe and efficient treatment.

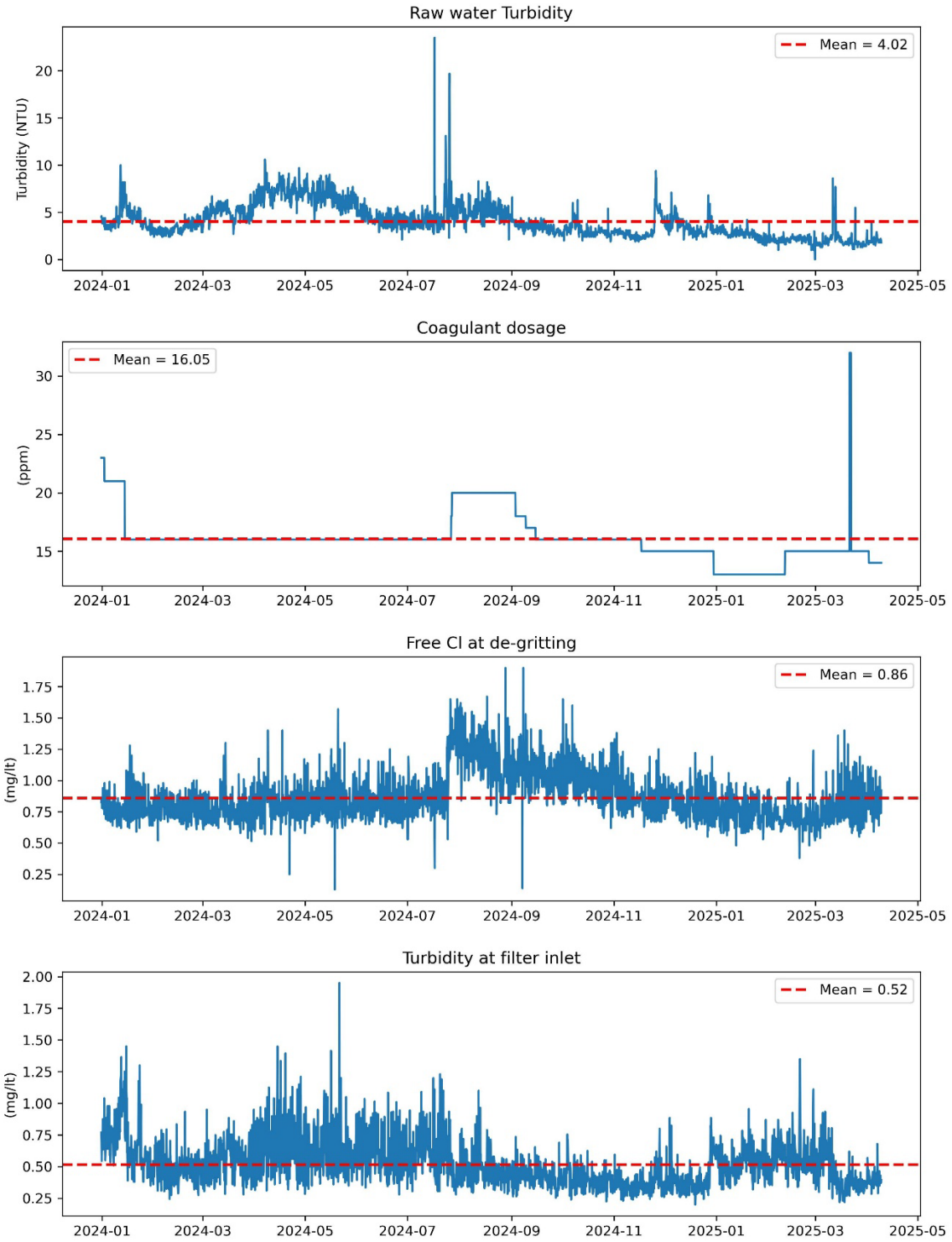

Figure 12: Recorded parameters behaviour over time.

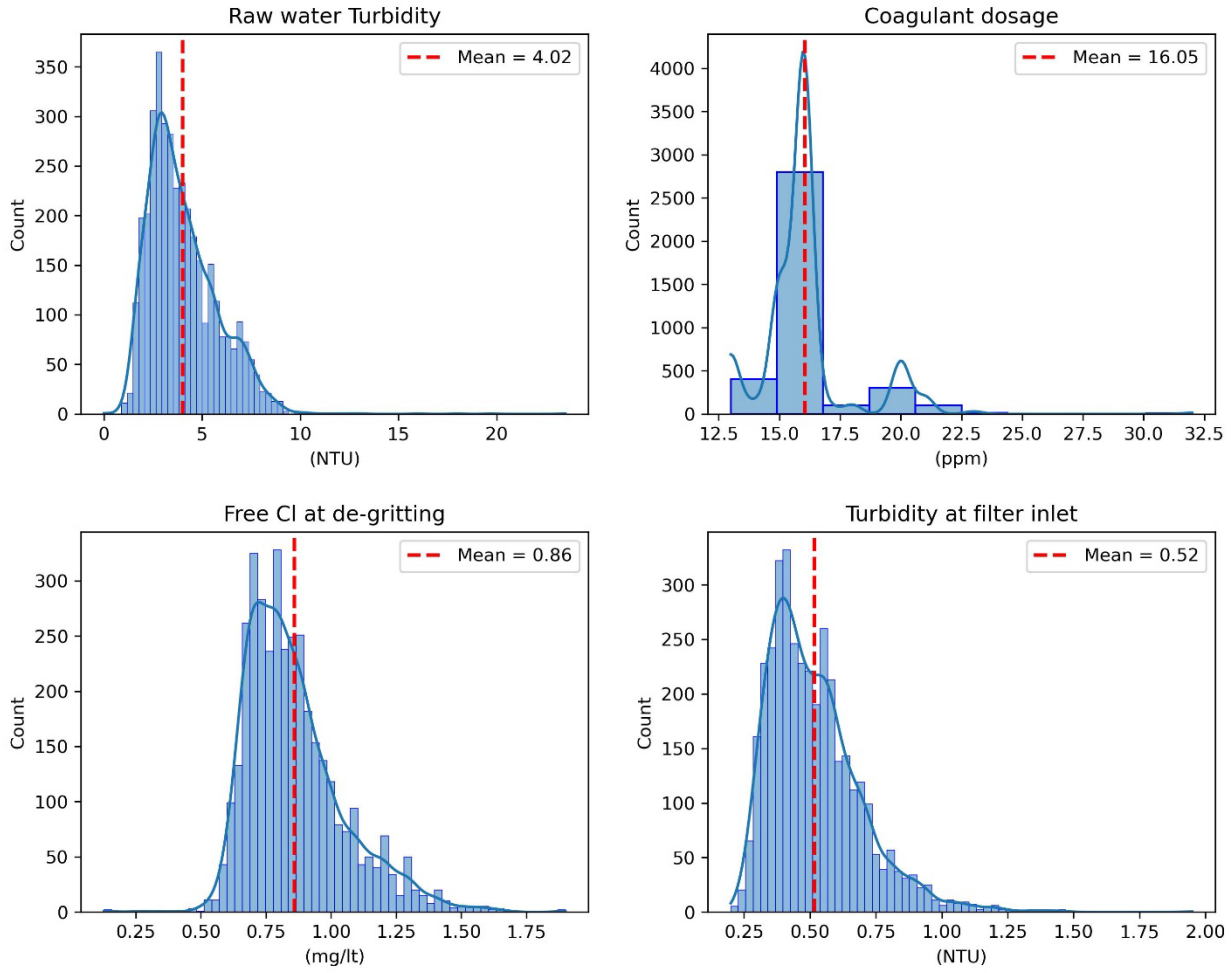

Figure 13: Exploratory Analysis of the coagulation dataset.

The final dataset contains 3,720 samples recorded at three-hour intervals, corresponding to approximately 15.5 months of operation (1 January 2024–9 April 2025) in contrast of the only one months of data that where in the beta version (D4.2). An exploratory analysis of the key process variables is presented in Figures 12 and 13. Except for elevated turbidity observed even at the end of July 2024, and a positive shift in free chlorine concentration at the degritting stage, the dataset exhibits generally smooth temporal behaviour, reflecting the stable operating conditions of the treatment process. It is important to note that the coagulant dosage setpoint is infrequently adjusted, which is consistent with the inherently slow dynamics of the coagulation process and the need to maintain operational stability. From this dataset, a MDP transition dataset was derived, where each sample represents a state–action–next-state tuple. This dataset was divided into a training set (2,976 samples) used for TD3-BC training and hyperparameter tuning, and an evaluation set (744 samples) used to assess policy performance and refine the MDP formulation.

2.4.3 Materials and methods

As previously discussed, the formulation of the MDP components, particularly the definition of the state space and the reward function, represents a major challenge in the application of RL to real-world systems. The final structure of the MDP state was defined through an extensive review of the relevant literature, consultation with domain experts, careful assessment of the available process data, and a series of comprehensive experimental evaluations. Leveraging a more extensive dataset with broader temporal coverage than the beta version presented in D4.2, a more representative MDP state was subsequently developed, along with a revised reward function. However, due to a sensor malfunction, free alum concentration data were unavailable during this period; consequently, this variable was excluded from both the MDP state and the reward function compared to the D4.2 beta version. The resulting state representation, which encapsulates the most relevant system dynamics and operational conditions, is presented in Table 6.

The reward function was formulated based on two principal elements: the applied disinfectant dosage, which represents the control action of the RL agent, and the measured FCl concentrations at the de-gritting and filter inlets, which constitute part of the subsequent MDP state (i.e., the system state observed three hours after the action is executed). This formulation establishes a direct relationship

Table 6: MDP state parameters of Coagulation RL

Coagulation MDP state parameters	
1. Raw water Turbidity (NTU)	4. Turbidity at filter inlet 3 hours prior (NTU)
2. Free Chlorine at de-gritting (mg/lt)	5. isDay (0, 1)
3. Turbidity at filter inlet (NTU)	

between the operational decision (dosage) and its delayed effect on the system's disinfection performance. The mathematical expression of this relationship is defined as follows:

$$R = -\left(w_1 \cdot Turb_{R_{Filters}} + w_2 \cdot Dosage\right)$$

where:

$\boldsymbol{Turb_{R_{Filters}}}$: is a quadratic penalty for the exceeding the turbidity constrains at filter inlet,

$\boldsymbol{Dosage}$ **:** the Aluminium Sulphate dosage set point.

The term $w_1 \cdot Turb_{R_{Filters}}$ represents the constraint satisfaction component of the reward function, reflecting the compliance of the system with the target turbidity levels at filter inlets. Conversely, the term $w_2 \cdot Dosage$ corresponds to the efficiency component, which penalizes excessive chemical usage and promotes operational optimization. The weighting coefficients w_1, w_2 were determined through extensive experimental evaluation and sensitivity analysis, as described in the previous section.

Similar to the pre-disinfection use case, once stable training was achieved through systematic tuning of the TD3-BC algorithm parameters, the policy evaluation phase was conducted. This phase assessed the performance of the learned policy against the existing operational policy using the hold-out evaluation dataset.

An approximate modeling framework was employed for this purpose, consisting of a single XGBoost model developed to predict turbidity levels at the filter inlets, representing the key process constraint in this case. The model was trained on the initial dataset, partitioned into 80% for training and 20% for testing. The input variables and corresponding performance metrics are summarized in Table 7.

Table 7:Coagulation evaluation AM

method	**Input parameters** (from MDP state)	**Output Parameter** (from next MDP state)	MAE
XGBoost	1. Dosage 2. Raw water turbidity 3. Turbidity at filter inlet 4. Turbidity at filter inlet 3 hours prior 5. FCl at de-gritting 6. isDay	**Turbidity at filter inlet**	**0.08 (NTU)**

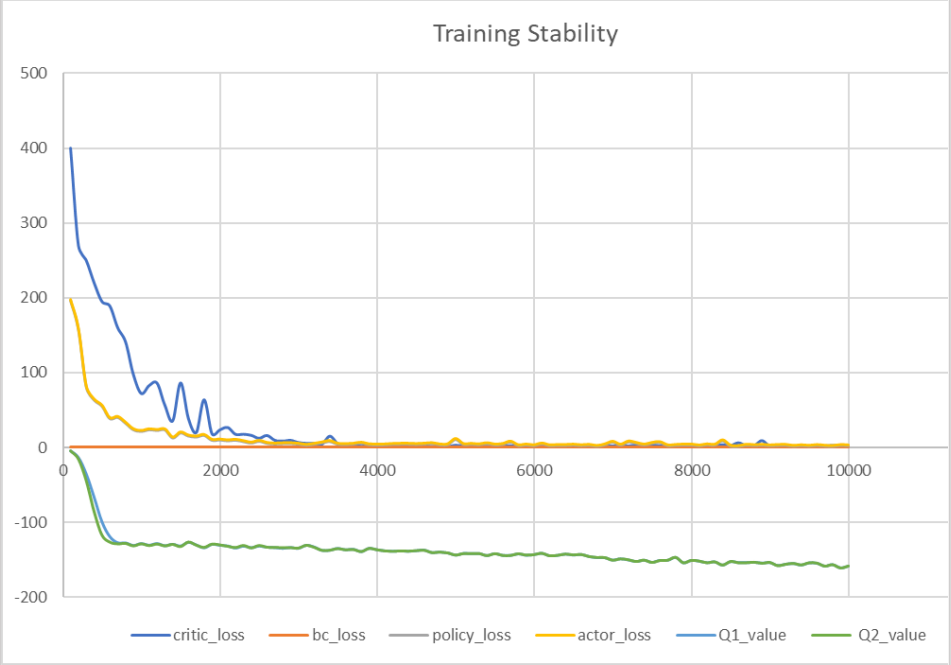

Figure 14: Actor & Critic Neural networks parameters evolution during 10000 training batches.

During policy evaluation (744 samples, see section 2.4.2), each historical MDP state transition from the evaluation dataset was used to estimate the subsequent turbidity response under the actions prescribed by both the learned and existing policies. The resulting turbidity predictions, together with the associated dosage values, were used as inputs to the reward function to quantify and compare overall policy performance. This modeling approach also facilitated a reliable estimation of potential constraint violations, providing insights into the trade-off between operational efficiency and compliance with disinfection requirements. The overall RL training performance is presented in the following section.

2.4.4 Results

After extensive experimentation with the TD3-BC algorithm parameters, a satisfactory level of training stability was achieved. As illustrated in Figure 14, all key parameters converged toward stable values, indicating that the learning process had reached equilibrium. However, despite this numerical stability, the learned policy did not fully align with historical operational patterns and was unable to effectively replicate the decision-making behaviour of human operators, as shown in Figure 15. A statistical analysis

of the evaluation dataset (744 samples), presented in Figure 16, provides quantitative evidence supporting these observed performance trends. As illustrated in Figure 12, the coagulant dosage setpoint remained largely unchanged over extended periods, in some cases persisting at a constant value for up to one month. In contrast, Figure 15 demonstrates that the RL policy induces more frequent adjustments to the dosage setpoint within a single day. Despite this increased temporal variability introduced by the RL-based control strategy, the overall reduction in coagulant dosage—shown in Figure 16—is marginal. Similarly, only minor changes are observed in turbidity levels at the filter inlet.

These findings highlight a critical distinction between convergence in the optimization process and the practical adequacy of the learned policy. In this case, stability in training did not necessarily translate into policy fidelity or operational relevance. This discrepancy may be attributed to limitations in the dataset, including its size, potential noise, and the absence of key explanatory variables that influence operator decisions but are not explicitly recorded. Furthermore, the inherent variability and heuristic nature of human decision-making in coagulation control may not be fully captured by the available state representation. Another contributing factor may lie in the constraints of the offline RL setting, where the agent is restricted to learning from a fixed dataset without the ability to explore or correct suboptimal regions of the state–action space. As a result, the learned policy may suffer from distributional shift or extrapolation errors when encountering states that are underrepresented in the training data.

2.4.5 Conclusions and next steps

Overall, these results highlight both the feasibility and current limitations of applying reinforcement learning to coagulation control. While the agent demonstrated stable learning dynamics, further improvement is required for effective policy generalization and closer alignment with expert operational strategies. The coagulation model achieved numerical training stability (Actor-Critic convergence confirmed in Figure 14) but lack to learn an exact policy that mirrors operator decisions. This is a known challenge in offline RL applied to processes with high operational inertia: coagulation dosage adjustments at Polydendri occur reactively and the relationship between dosage and turbidity outcome has a significant temporal delay that even a 3-hour state transition interval may not capture adequately. Future work will focus on continued collaboration with EYDAP process engineers to validate the TD3-BC model's recommended dosage strategies under real operational conditions. This validation will be carried out through the Web Platform developed in WP7, enabling direct assessment and refinement of the model's performance in practice. Such advancements are expected to enhance the interpretability, reliability, and practical applicability of the learned policy, paving the way for its progressive deployment in real-world water treatment operations.

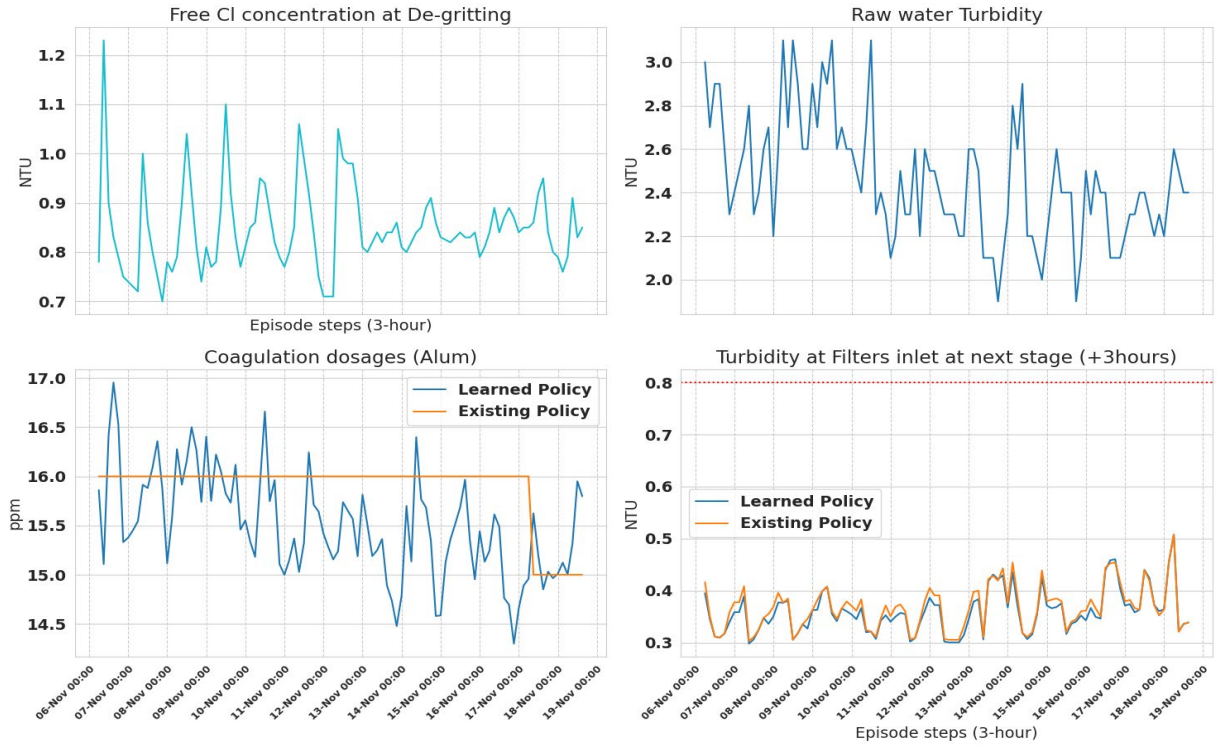

Figure 15: Learned Policy performance on a 100-step episode.

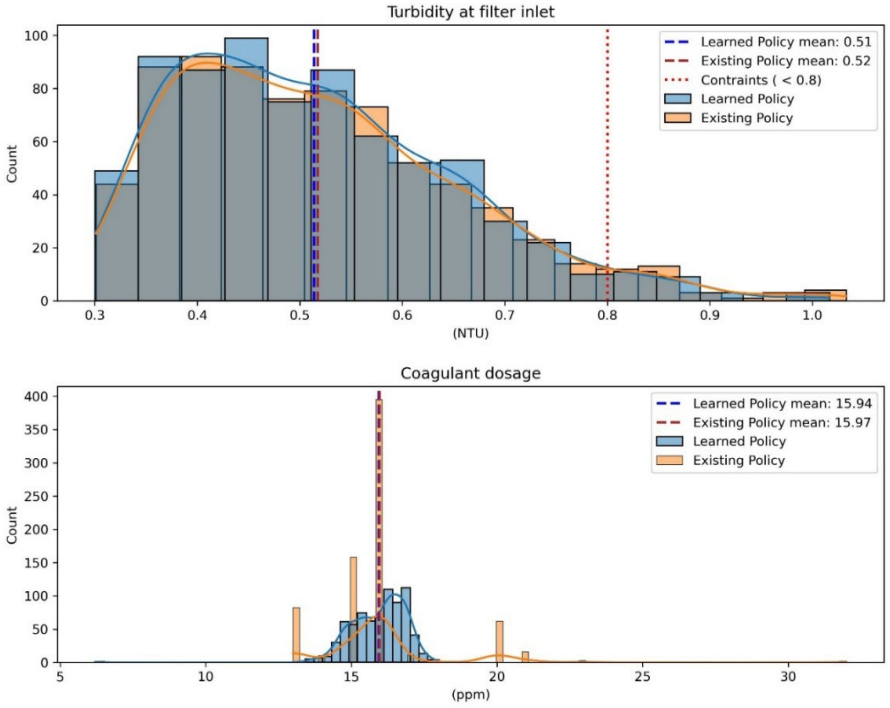

Figure 16: Performance histogram plots of the learned policy over the evaluation dataset.

2.5 RL model #3: Optimization of post-disinfection stage

2.5.1 Problem statement

At the Polydendri DWTP, the post-disinfection process occurs downstream of the filtration stage and upstream of the clear water tanks, as illustrated in Figure 4. Sodium hypochlorite is employed as the primary disinfectant. Until recently, chlorine gas had been used for this purpose; however, the plant is currently transitioning toward the exclusive use of sodium hypochlorite. Maintaining appropriate free chlorine concentrations downstream is essential to ensure effective disinfection performance.

The efficiency of this stage is primarily evaluated through continuous monitoring of free chlorine concentrations, measured by online sensors installed at both the inlet and outlet of the two clear water tanks. These measurements serve as critical indicators of disinfection efficacy and are used to verify compliance with water quality standards.

At present, disinfectant dosing at the DWTP is adjusted in a reactive manner, based solely on the observed downstream free chlorine concentration. The objective of the reinforcement learning (RL) algorithm is to optimize the disinfectant dosage, expressed in parts per million (ppm), to maintain free chlorine levels consistently within predefined operational limits. This data-driven optimization aims to enhance process efficiency while ensuring the microbiological safety and overall quality of the treated water.

Table 8: Post-disinfection recorded parameters

Parameter Name	Parameter Type	Units	Resolution	Low limit	High limit	Recording frequency	Source
Raw water flow	External Variable	m3/hour	1	-	-	1 minute	SCADA
Chlorine dosage	Manipulated Variable	ppm	0.01	-	-	1 minute	SCADA
Free Cl clear water Tank inlet	Constraint Variable	mg/lt	0.01	0.6 (Summer) 0.7 (Normal)	0.9 (Summer) 0.8 (Normal)	1 sec	SCADA

2.5.2 Data sources and data preprocessing

The primary data sources for the RL model are drawn from the DWTP existing monitoring system. The DWTP SCADA system continuously captures all necessary parameters for the post-disinfection process, with data from online sensors recorded at intervals of less than one minute. The critical parameter indicating the effectiveness of post-disinfection is the free chlorine concentration at the inlet of the clear water tank. For the development of the final version of this RL use case, data from the SCADA system has been used exclusively. The parameters that are utilized are presented in the Table 8 along with contextual information.

Raw water flow rate is categorized as external variable, as its value is influenced by external environmental factors and remains unaffected by the post-disinfection process. Chlorine dosage is classified as a manipulated variable, as it has a direct impact on the post-disinfection process and can be adjusted to regulate system performance and it is the action that the RL Agent is expected to optimize. In contrast, the free chlorine concentration at the inlet of the clear water tank is treated as a constraint

variable, as it must be maintained within established safety limits to ensure the safe and effective operation of the treatment process.

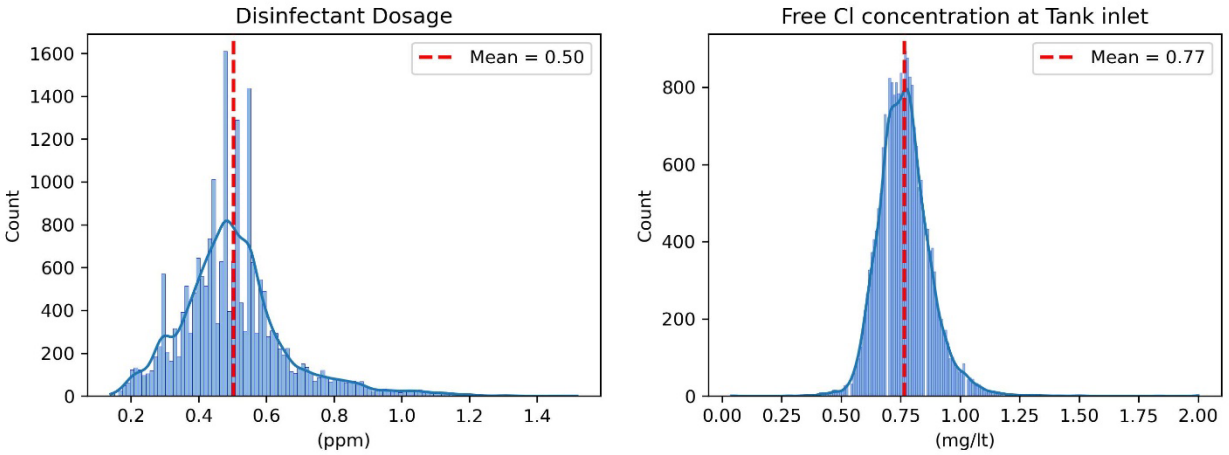

Figure 17: Exploratory Analysis of the post-disinfection dataset.

A 15-minute time interval was adopted as the main temporal resolution, and all the measurements have been down-resampled by selecting the nearest recorded value. Similarly, free chlorine concentration measurements were down-sampled to 15-minute intervals using the closest recorded values. The final dataset consists of a total **20730 data points**, each recorded at a **15-minute time interval**. Corresponding to approximately **7.5 months** of plant operation covering a period from **1 July 2024 to 28 August 2025**. The reduction in the overall temporal coverage was necessitated by the occasional use of chlorine as a disinfectant during certain operational periods. Since the RL framework developed in this study was specifically designed to model the sodium hypochlorite disinfection scenario, only the corresponding time intervals were retained for further analysis.

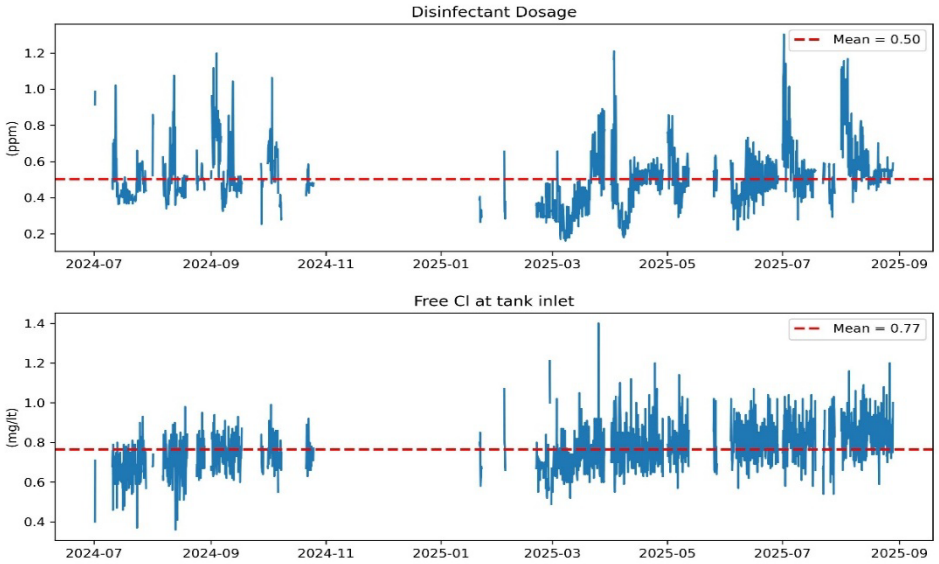

Figure 18: Post-disinfection recorded parameters behaviour over time.

A detailed exploratory analysis of the main process variables included in the dataset is presented in Figure 18 providing an overview of their distribution and variability throughout the selected operational period. The gaps observed in Figure 18 correspond to operational periods during which the DWTP employed chlorine as the primary disinfectant. From the initial dataset, a new dataset of MDP state transitions was subsequently constructed. Each data point in this derived dataset represents a complete transition tuple, consisting of the current MDP state (comprising a superset of all relevant process variables), the action executed (i.e., the applied dosage setpoint), and the resulting next MDP state.

This MDP transition dataset was then divided into two subsets: a **training dataset** comprising **16,584 samples** and an **evaluation dataset** consisting of **4,146 samples**. The training dataset was used for training the RL agent and for optimizing the hyperparameters of the TD3-BC algorithm. Conversely, the evaluation dataset was employed to assess the performance of the learned policy and to refine the formulation of the optimal MDP components, namely the state representation and the reward function.

2.5.3 Materials and methods

As previously discussed, formulating the components of the MDP, in particular the definition of the state space and the reward function, constitutes one of the main challenges in applying RL to real-world systems. Following an extensive review of the relevant literature (Kulkarni & Chellam, 2010) (Park, et al., 2018), consultations with domain experts, careful examination of the available process data, and a series of comprehensive experimental evaluations, andleveraging a more extensive dataset with broader temporal coverage than the beta version presented in D4.2, a more representative MDP state was subsequently developed, along with a revised reward function.The resulting state representation, which captures the most relevant system dynamics and operational conditions, is presented in Table 9.

Table 9: MDP state parameters of Post-disinfection RL

Post-Disinfection MDP state parameters	
1. Raw water flow rate (m^3/hour)	5. Dosage set point 3 hours prior (ppm)
2. Free Chlorine at tank inlet (mg/lt)	6. Dosage set point 6 hours prior (ppm)
3. Free Chlorine at tank inlet 3 hours prior (mg/lt)	7. Hour of the day (0-23)
4. Free Chlorine at tank inlet 6 hours prior (mg/lt)	

The reward function was formulated based on two principal elements: the applied disinfectant dosage, which represents the control action of the RL agent, and the measured FCl concentrations at the de-gritting and filter inlets, which constitute part of the subsequent MDP state (i.e., the system state observed three hours after the action is executed). This formulation establishes a direct relationship between the operational decision (dosage) and its delayed effect on the system's disinfection performance. The mathematical expression of this relationship is defined as follows:

$$R = -\left(w_1 \cdot FreeCl_{R\,Tank\ inlet} + w_2 \cdot Dosage\right)$$

where:

$\boldsymbol{FreeCl_{R\,Tank\ inlet}}$: is a quadratic penalty for the exceeding the Free Cl constrains at tank inlet,

$\boldsymbol{Dosage}$ **:** the sodium hypochlorite dosage set point.

The term $w_1 \cdot FreeCl_{R\,Tank\ inlet}$ represents the constraint satisfaction component of the reward function, reflecting the compliance of the system with the target free chlorine concentration levels at the tank inlet. Conversely, the term $w_2 \cdot Dosage$ corresponds to the efficiency component, which penalizes excessive chemical usage and promotes operational optimization. The weighting coefficients w_1, w_2

were determined through extensive experimental evaluation and sensitivity analysis, as described in the previous section.

As outlined in Section 2.2.3, after achieving training stability through systematic tuning of the TD3-BC algorithm parameters, the policy evaluation phase was initiated. This phase aimed to assess the performance of the learned policy relative to the existing operational strategy, using a dedicated hold-out evaluation dataset. To support this analysis, an approximate modeling framework was developed, consisting of an XGBoost-based predictive model designed to estimate the FCl concentration at the tank inlet, a key process constraint in the studied use case. The model was trained on the initial dataset, which was divided into 80% for training and 20% for testing. The selected input variables and the corresponding model performance indicators are presented in Table 10. During the evaluation process, each historical MDP state transition from the evaluation dataset was used to predict the subsequent FCl concentration based on the actions prescribed by both the learned and the baseline policies. The predicted FCl values, together with the respective dosage levels for each policy, were then incorporated into the reward function to quantify and compare their overall performance. This approximate modeling approach further enabled a reliable estimation of potential constraint violations for critical process parameters, offering valuable insight into the trade-offs between operational efficiency and compliance with disinfection performance requirements. The overall results of the reinforcement learning training are discussed in the following section.

Table 10: Poste-disinfection evaluation AM

Method	**Input parameters** (from MDP state)	**Output Parameter** (from next MDP state)	**MAE**
XGBoost	1. Dosage 2. Dosage 3 hours prior 3. Dosage 6 hours prior 4. FCl tank inlet 5. FCl tank inlet3 hours prior 6. Hour of day	**FCl tank inlet**	**0.03 (mg/lt)**

2.5.4 Results

After extensive experimentation with the TD3-BC algorithm parameters, a satisfactory level of training stability was achieved. As shown in Figure 19, all critical parameters converged toward stable values, confirming the robustness of the learning dynamics. Furthermore, the RL agent exhibited behaviour closely resembling the operational policy typically adopted by DWTP personnel (Figure 20). Through iterative learning, the agent's decision-making progressively aligned with historical operational patterns, effectively replicating the actions of human operators. Notably, this behavioural convergence was accompanied by an approximate **7.8% reduction** in disinfectant dosage, suggesting that the RL agent successfully identified subtle process optimizations without compromising safety or operational reliability. Although modest, this reduction could yield meaningful improvements in chemical use efficiency when sustained over extended operational periods. A statistical analysis of the evaluation dataset (4,146 samples), presented in Figure 21, provides quantitative evidence supporting these observed performance trends.

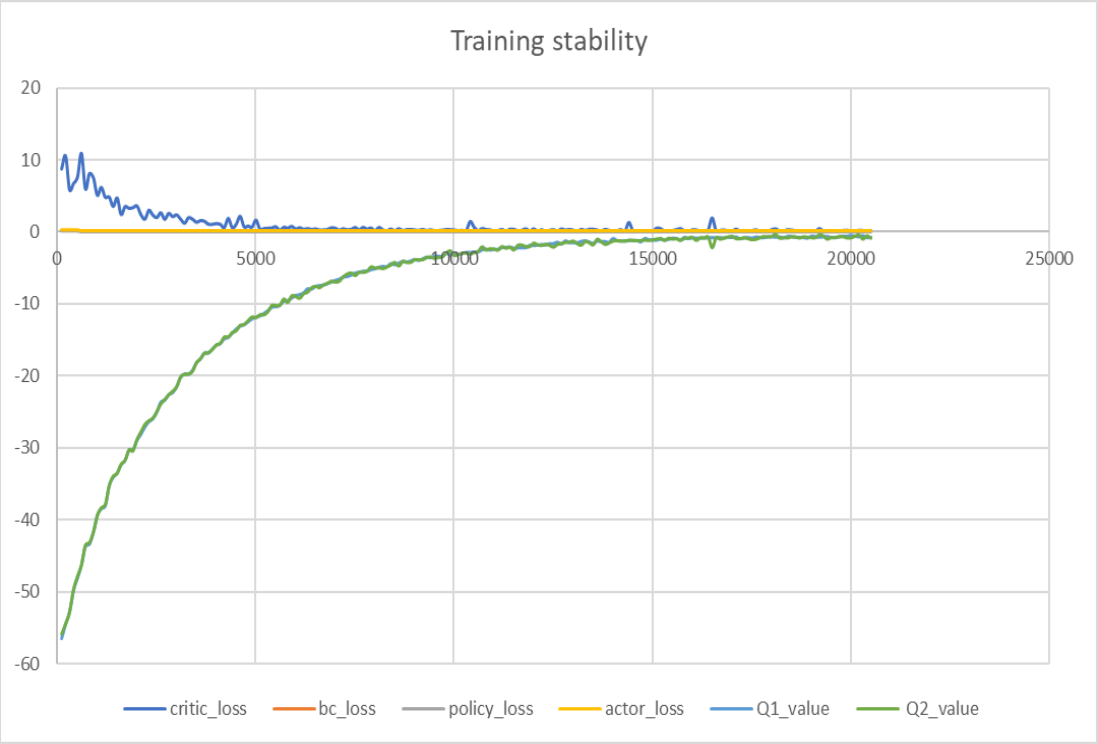

Figure 19: Actor & Critic Neural networks parameters evolution during 20000 training batches.

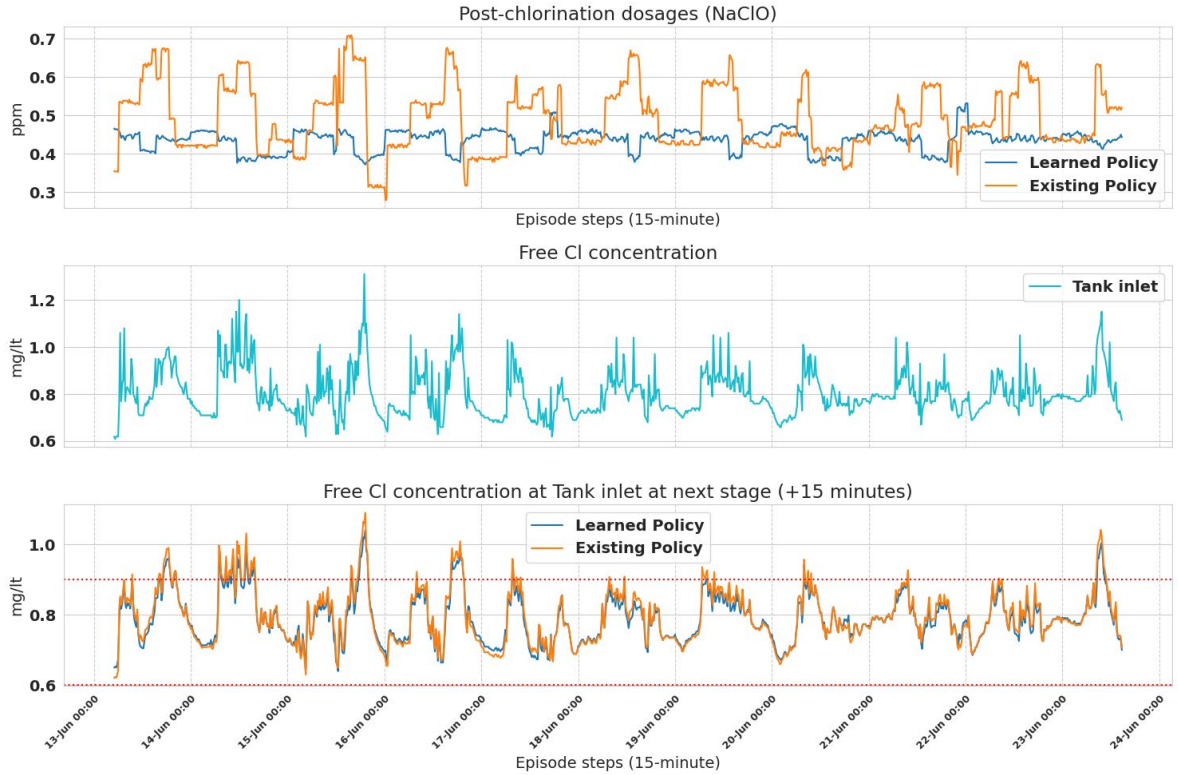

Figure 20: Learned Policy performance on a 1000-step episode.

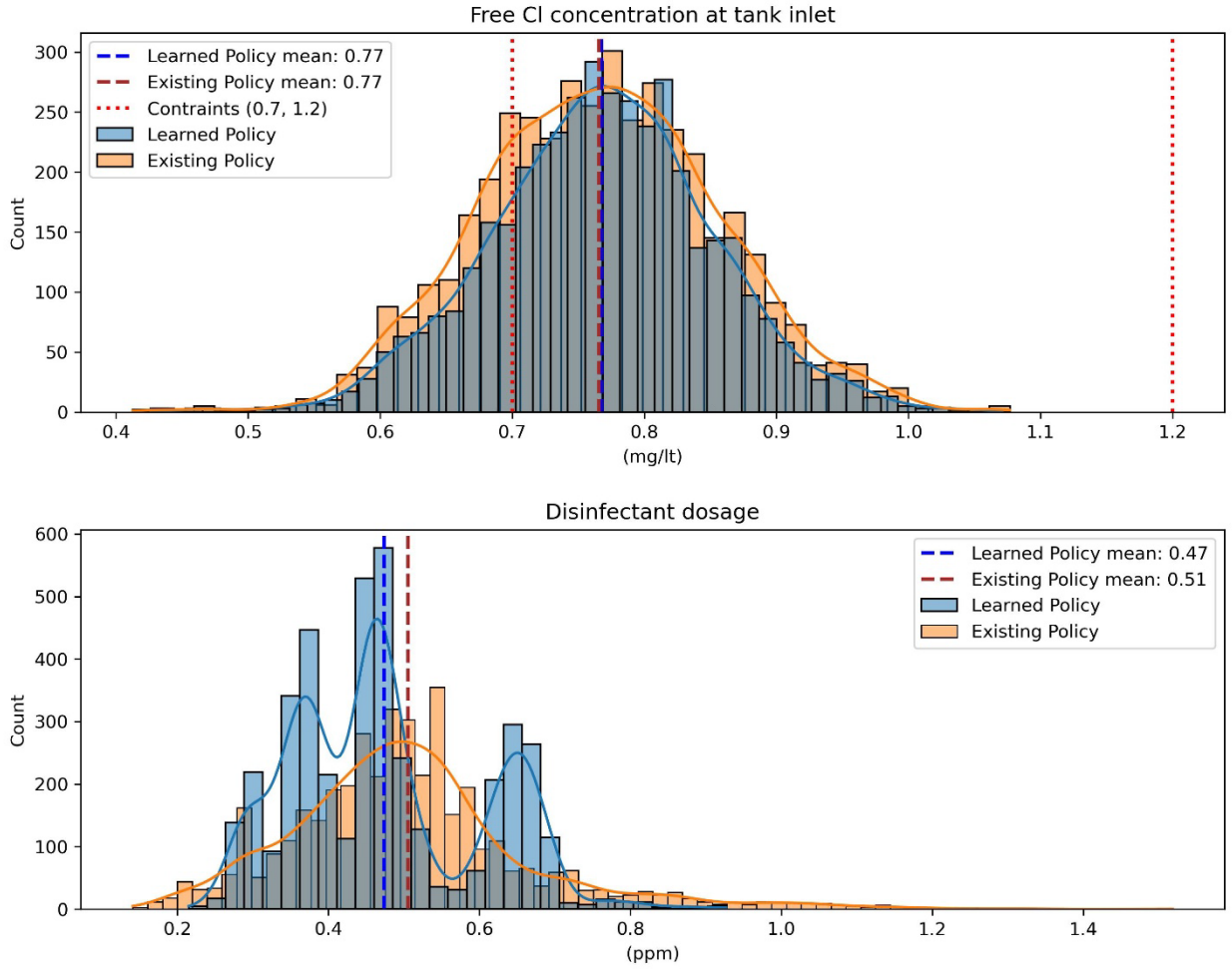

Figure 21: Performance histogram plots of the learned policy over the evaluation dataset.

Conclusions and next steps

In summary, the developed TD3-BC–based RL framework demonstrated stable training performance and effectively reproduced the operational strategies typically employed by DWTP personnel, while achieving a marginal reduction in disinfectant usage. These results highlight the potential of RL-based decision-support tools to enhance process efficiency and promote data-driven optimization of water treatment operations.

It is important to emphasize, however, that the learned policy has not yet been validated under real-world operating conditions. Conducting such an evaluation entails considerable safety and regulatory challenges due to the direct influence of disinfectant dosing on water quality and public health. To address this limitation, the RL framework is planned for integration into the WP7 platform, where it will be made accessible to operators at the Polydendri DWTP. Through this platform, operators will be able to input relevant process parameters, receive RL-generated dosage recommendations, and provide feedback by assigning a performance score to each suggestion. This feedback will be stored in the platform's database, enabling the progressive accumulation of real-world evaluation data. Over time, this human-in-the-loop approach is expected to support a more representative assessment of the RL agent's performance and facilitate its gradual adaptation toward safe, interpretable, and reliable real-world deployment in water treatment operations.

2.6 RL model #4 pyStimela pre-chlorination RL environment

2.6.1 Problem statement

As outlined in section 2.2.3, the core functionalities of the Stimela package have been re-implemented in Python, resulting in the development of the pyStimela package. This provided an opportunity to leverage Stimela's chlorination simulation capabilities to emulate the dynamic behaviour of the disinfection process within an online RL environment. In this configuration, the RL agent interacts with the simulated process in real time, receives feedback in the form of rewards, and iteratively improves its decision-making through experience-based learning. This approach contrasts with the offline RL framework, where the agent learns from historical data representing past operational decisions made by DWTP personnel.

It is important to acknowledge that the parameterization of pyStimela may not fully reproduce the specific operational and water quality characteristics of the Polydendri DWTP pre-disinfection process. Several key input parameters required by pyStimela are not available for the Polydendri demonstration case, which introduces inherent uncertainties in the simulation results. Furthermore, pyStimela currently supports only chlorine as a disinfectant, whereas the Polydendri plant is in the process of phasing out chlorine and transitioning to sodium hypochlorite. This discrepancy further limits the direct representativeness of the simulation with respect to the actual disinfection practices at the facility.

Despite these limitations, the development of the pyStimela-based online simulation environment was pursued as a proof of concept. The primary objective was to demonstrate the technical feasibility of coupling physics-based process simulators with reinforcement learning frameworks. This integration enables the creation of interactive environments where agents can engage in real-time learning, thereby supporting the development and evaluation of intelligent control policies for optimized water treatment operations.

2.6.1 Data sources and data preprocessing

Table 11 presents the recorded parameters utilized in this reinforcement learning (RL) use case. The historical measurements of these parameters served two primary purposes. First, the parameters identified as external variables were used to develop statistical models that describe the temporal dynamics of the influent water quality in the Polydendri demonstration case in order to incorporated to the RL environment implementation for the pre-disinfection process. These variables are classified as external because their values are independent of the chlorination process and instead reflect fluctuation ns in the raw water entering the treatment plant.

Table 11: Available recorded parameters for the pre-chlorination RL Environment

Parameter Name	Parameter Type	Units	Resolution	Low limit	High limit	Recording frequency	Source
Raw water flow	External Variable	m^3/hour	1	-	-	1 minute	SCADA
Raw water Temperature	External Variable	Celsius	0.1	-	-	5 minutes	Conveyance Monitor System
Raw water TOC	External Variable	mg/lt	0.01	-	-	15 minutes	Conveyance Monitor System
Raw water	External	1/m	0.01	-	-	15 minutes	Conveyance

Parameter Name	Parameter Type	Units	Resolution	Low limit	High limit	Recording frequency	Source
uv254	Variable						Monitor System
Cl dosage	Manipulated Variable	ppm	0.1	-	-	3 hours	Logbook
Free Cl de-gritting	Constraint Variable	mg/lt	0.01	0.7	1.2 (Summer) 0.8 (Normal)	3 hours	Logbook

In contrast, parameters directly associated with the chlorination process, categorized in Table 11 as constraint and manipulated variables, were employed, together with the external variables, to construct a validation dataset of 1054 samples with 3-hour timestep for assessing the performance of the learned RL policy. This dataset enabled the quantitative evaluation of the agent's decision-making behaviour under realistic process conditions, ensuring that the resulting control strategy aligns with both operational constraints and process objectives.

2.6.2 Materials and methods

The PyStimela simulation framework requires a defined set of water quality parameters as input to estimate key chlorination-related outputs, including Free Cl concentration, total trihalomethanes (TTHMs), and haloacetic acids (HAAs). In addition, the model has the capability to simulate microbial inactivation, specifically the concentrations of Giardia and viruses, based on appropriate input parameters. However, this functionality was excluded from the current demonstration case. The main influent parameters required by PyStimela include water temperature (°C), flow rate (m^3/h), pH, dissolved organic carbon (DOC), UV254 absorbance (1/m), bromide ion concentration (mg/L), and ammonia concentration (mg/L), with Giardia and viruses only necessary for microbial risk assessment modules.

Table 12: MDP state parameters

Coagulation MDP state parameters	
1. Raw water flow rate (m^3/h)	4. Raw water DOC (mg/lt)
2. Raw water temperature (Celsius)	5. Free Cl downstream (mg/lt)
3. Raw water pH	

The MDP state variables presented in Table 12 constitute a subset of the previously described Stimela input and output parameters. Several parameters, specifically TOC, bromide, and ammonia, are not routinely monitored at the Polydendri DWTP. To compensate for these data gaps, representative literature-based average values were assigned to bromide and ammonia concentrations (both set to 0.1 mg/L), reflecting typical conditions for surface water sources; consequently, these variables were excluded from the formal MDP state definition. The DOC concentration was inferred from TOC using a constant conversion factor of 0.8, while UV254 was approximated as 2.5 times the estimated DOC value, consistent with their well-documented empirical correlation. These assumptions enabled the construction of a coherent and practically meaningful state-space formulation in the absence of complete monitoring data. As a result, the RL agent can be trained and deployed under realistic operational conditions despite inherent data limitations. In addition, a 3-hour timestep was adopted for the state transitions, consistent with the configuration used in RL Model #1.

Table 13: Initial MDP state generation and transitioning mechanism

Parameter	Initialization Method	Transitioning Method
Raw Water temperature	Linear regression model based on date and flow rate (R2=0.82, MAE=1.2C) + random noise	Adding a small random noise
Flow rate	Randomly from a statistical model (multinomial distribution)	Adding small random noise
pH	Randomly from a statistical model (mixture of two Gaussians)	Adding small random noise
DOC	Randomly from a statistical model of TOC * 0.8 (mixture of 6 Gaussians)	Adding small random noise
UV254	Randomly From DOC statistical model * 2.5 + random noise	Adding small random noise
Free Cl at de-gritting	Randomly from a statistical model of Free Cl at de-gritting (Normal)	From pyStimela + Linear regression model

PyStimela also requires specification of the chlorination tank characteristics to perform its calculations. In particular, the model needs the effective tank volume, which for the Polydentri DWTP is 253 m³. The initialization procedures associated with each PyStimela input, together with the transition dynamics defined for the reinforcement learning environment, are summarized in Table 13. It should be noted that, for the Free Cl concentration at the de-gritting stage, a simple regression model was developed based on historical data. This adjustment was necessary because the initial analysis revealed a relatively consistent offset between the PyStimela-simulated values and the corresponding measurements in the historical dataset. The regression model therefore serves to align the simulated outputs more closely with the observed data. A plausible explanation for this discrepancy is that the de-gritting facility is located approximately 100 meters downstream of the pre-disinfection tank's outlet, which may introduce differences between the modelled and measured chlorine concentrations.

The reward function was designed around two principal components: the applied disinfectant dosage, representing the control action of the RL agent, and the resulting free chlorine (FreeCl) concentration at the system output. The latter forms part of the subsequent MDP state—specifically, the state observed three hours after the control action is applied—and must remain within the constraint limits presented in Table 11. This formulation establishes a direct linkage between the operational decision (dosage) and its delayed impact on the disinfection performance of the system. The mathematical definition of this relationship is provided below:

$$R = w_1 \cdot e^{-\frac{1}{0.1}\cdot(FreeCl-1.1)^2} - w_2 \cdot \frac{Dosage_{used} - Dosage_{max}}{Dosage_{max} - Dosage_{min}}$$

where:

$\boldsymbol{FreeCl}$: the Free Chlorine concentration at the output,

$\boldsymbol{Dosage_{used}}$: the Chlorine dosage set point,

$\boldsymbol{Dosage_{min}}$, $\boldsymbol{Dosage_{max}}$: the Chlorine dosage minimum and maximum set point,

$\boldsymbol{w_1}$: the weight for the constrain reward term,

$\boldsymbol{w_2}$: the weight for the efficiency reward term.

Figure 22: Actor & Critic Neural networks parameters evolution during 10000 training batches.

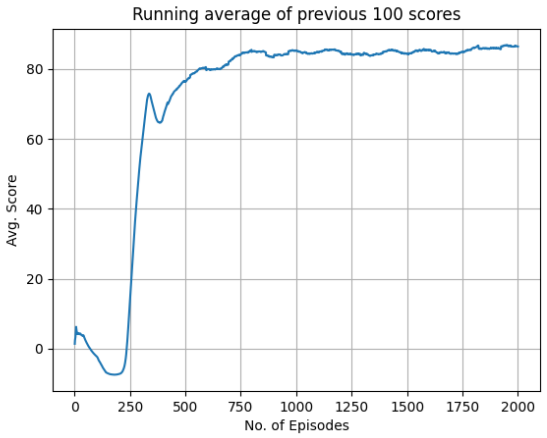

Figure 23: RL Agent average performance during training.

The RL agent was trained using the TD3 algorithm(Fujimoto et al., 2018). Extensive experimentation was conducted with TD3-specific hyperparameters, including the dimensionality of neural-network hidden layers, learning rates for both actor and critic networks, network architecture choices, state preprocessing strategies, and gradient-clipping thresholds. Through systematic tuning and the incorporation of stability-enhancing techniques, a robust and stable training procedure was achieved. The model's performance was subsequently validated using an evaluation dataset comprising 1,054 historical samples from the Polydendri DWTP. The results of the training and evaluation process are presented in the following section.

2.6.3 Results

The RL agent was trained for 2,000 episodes, each comprising 50 three-hour timesteps, equivalent to approximately six days of simulated system operation per episode. As illustrated in Figure 24, the agent's average return—computed over a rolling window of 100 episodes—stabilized at approximately 87 after roughly 1,000 training episodes, with 100 representing the maximum achievable value. This plateau indicates that the policy consistently approaches near-optimal performance. Moreover, Figure 25 shows that all key state variables and internal parameters converge toward steady values throughout training, confirming stable and robust learning dynamics.

The evaluation results, obtained using the historical dataset of the Polydendri DWTP, further highlight the effectiveness of the learned policy. As shown in Figure 25 and Figure 26, the mean disinfectant dosage recommended by the RL agent across 1,054 evaluation samples is higher than the dosage applied under the existing operational strategy. Nonetheless, this increase yields a clear benefit in terms of water-quality performance: the Free Cl concentrations at the de-gritting stage remain substantially more consistent within the prescribed regulatory limits of 0.7–1.2 mg/L. **These findings indicate that the learned policy achieves a more effective balance between operational dosing and compliance requirements, resulting in markedly improved reliability in satisfying downstream water-quality constraints.**

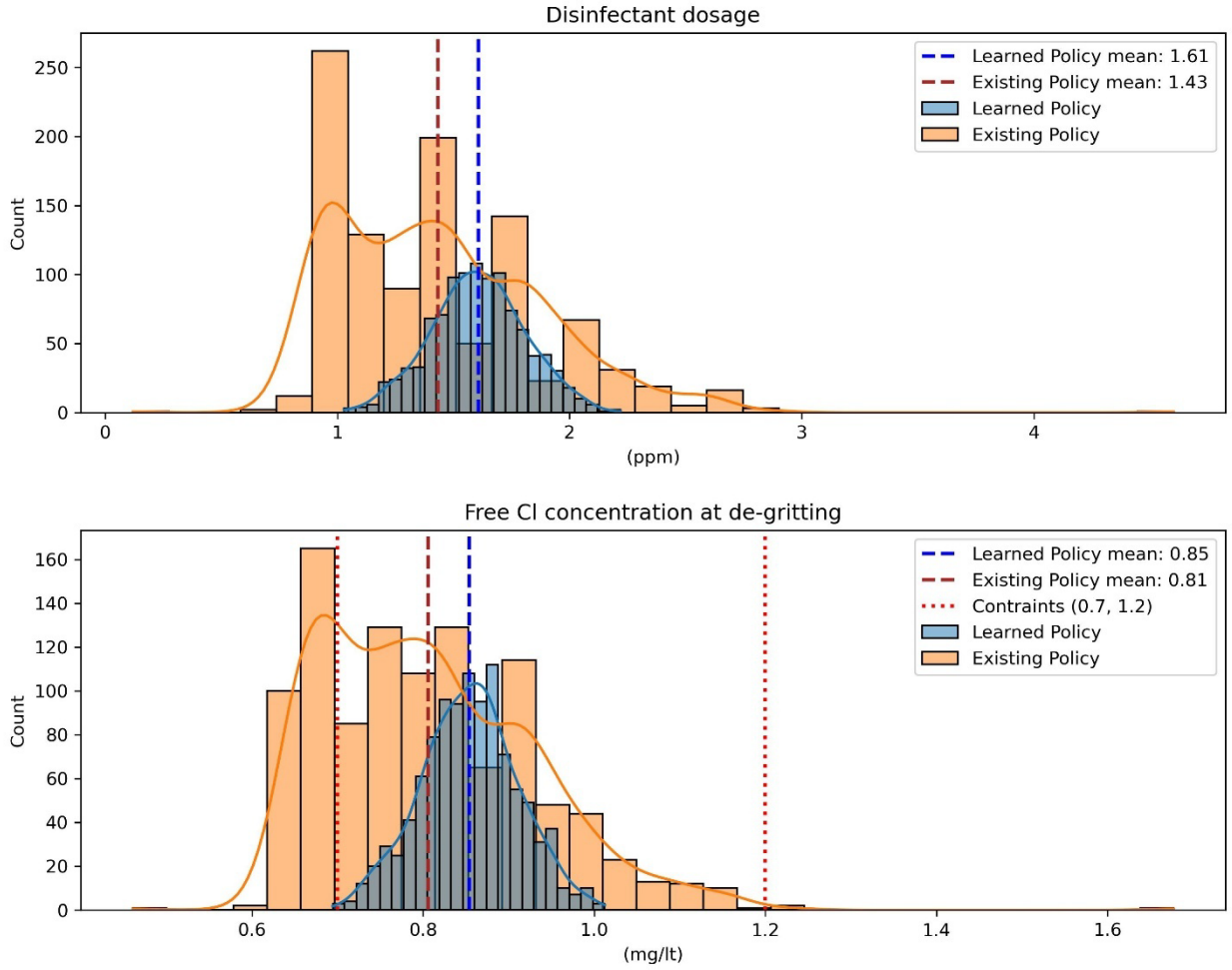

Figure 24: Performance histogram plots of the learned policy over the evaluation dataset.

2.6.4 Conclusions and next steps

This work demonstrated the feasibility and potential of integrating physics-based process simulation with reinforcement learning for intelligent control of drinking water treatment processes. The re-implementation of key Stimela functionalities in Python enabled the development of the pyStimela package, which served as the foundation for constructing an online RL training environment capable of simulating the dynamic behaviour of the chlorination process. Although the parameterization of pyStimela cannot fully replicate the specific operational characteristics of the Polydendri DWTP—due to unavailable input parameters and the plant's ongoing transition from chlorine to sodium hypochlorite—the simulation framework provided a robust proof of concept. The RL agent exhibited stable and consistent learning dynamics, with performance converging near optimal return values. When subsequently evaluated on real historical data, the learned policy produced dosages that, while higher on average, resulted in significantly improved compliance with Free Cl concentration constraints at de-gritting, demonstrating the potential of data-driven policies to enhance operational reliability.

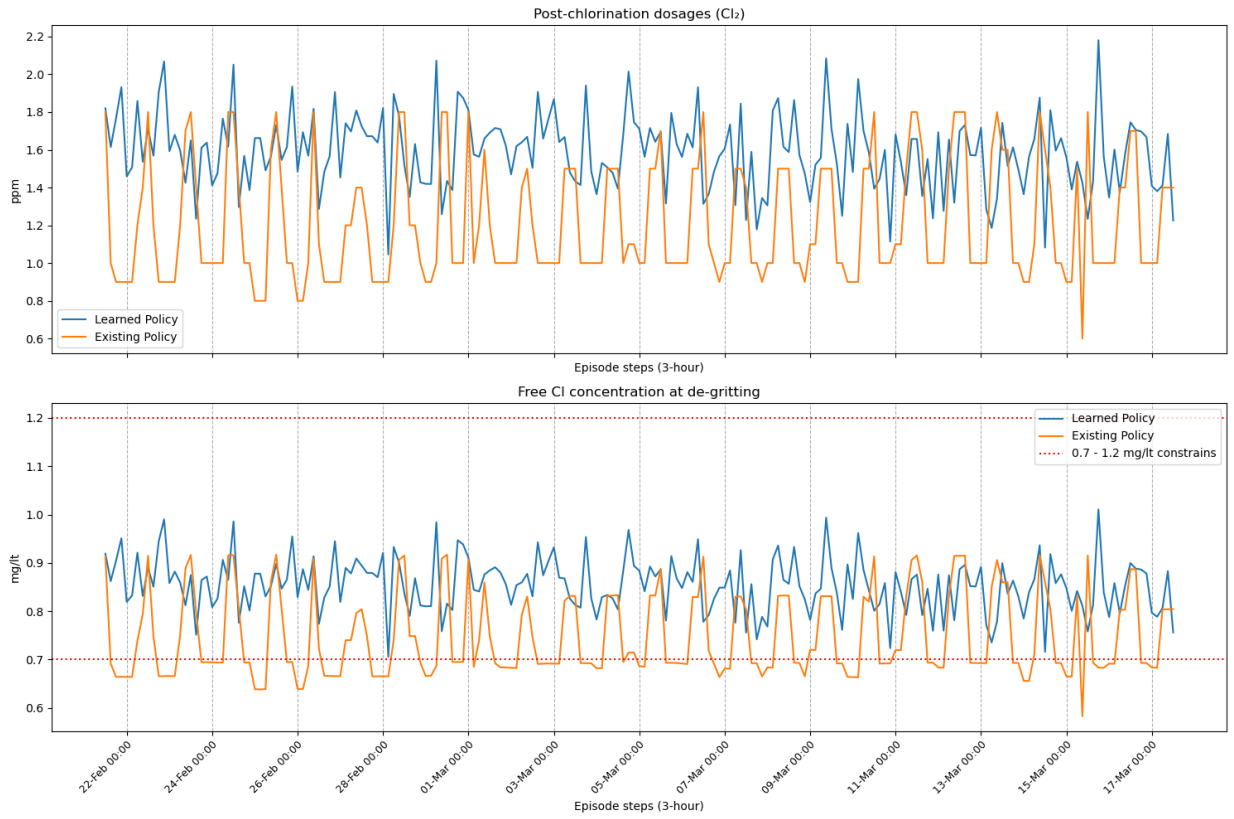

Figure 25: Performance of the Learned Policy Over a One-Month Window of Historical Data.

Given that the major development efforts have been completed in this project, the next steps primarily focus on operational evaluation, validation, and integration. A key priority is the incorporation of this RL use case into the WP7 web platform, enabling human-in-the-loop assessment by Polydendri DWTP operators. This integration will allow practitioners to interact directly with the learned policy, explore its recommendations under real operational scenarios, and provide expert feedback on its practical relevance and usability. Such operator-driven evaluation is essential for assessing the interpretability, trustworthiness, and decision-support potential of RL-based control strategies in real-world water treatment settings.

3. Reinforcement Learning model for Amsterdam demo case (DC#1)

3.1 Demo case description, Leiduin DWTP

WTNT supplies more than 90 million cubic meters of drinking water annually to consumers in the Amsterdam area. The water is produced at two different DWTPs, Leiduin and Weesperkarspel. Leiduin is the main DWTP as it produces approximately 70% of the water that feeds the Amsterdam area. The main source of the drinking water produced in the Leiduin plant is river water from the Lek Canal, supplemented by natural dune water. The treatment process consists of the pretreatment phase that takes place in Nieuwegein and the main treatment that takes place in Leiduin. The pretreatment consists of 2 stages, coagulation and rapid sand filtration and then the pre-treated water is transported to the Amsterdam Water supply Dunes (AWD) through 3 pipelines of 210 km length using 8 pumps. The main treatment process starts in the AWD with the infiltration of pretreated water. Thereafter, the following stages are rapid sand filtration, ozonation, that is used for both oxidation and disinfection, softening of the water, carbon filtration and slow sand filtration. The treated water is then stored in two different service reservoirs (storage tanks). Finally, the water is distributed in the Amsterdam area using pumps and 3 large pipelines. A schematic of this plant is presented in the following Figure 26.

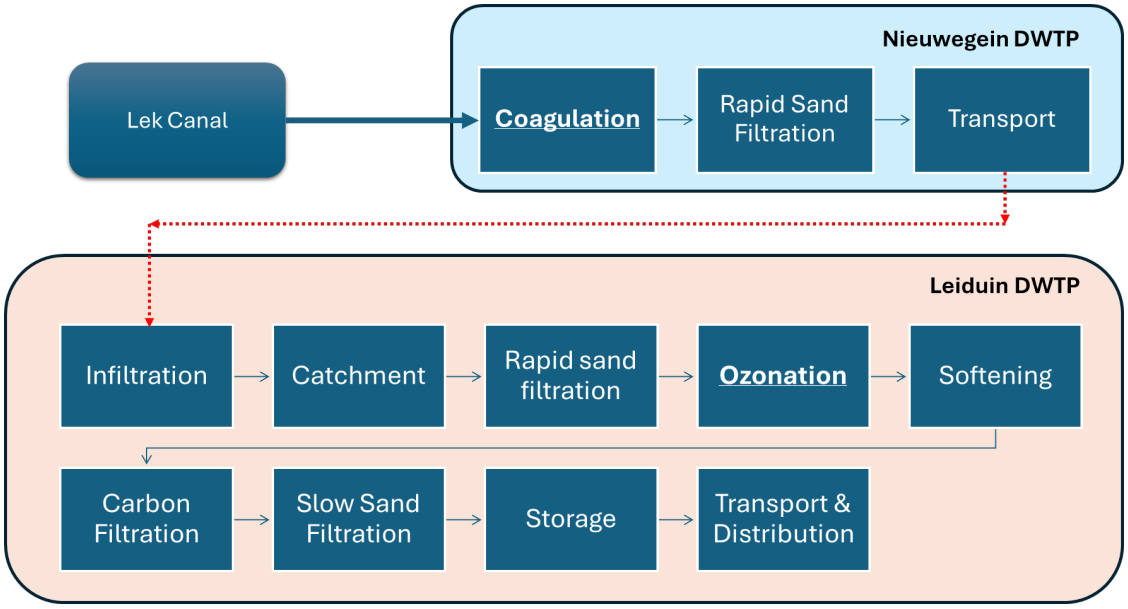

Figure 26: A schematic of the Leiduin drinking water treatment plant

3.2 RL model 5: Optimal $FeCl_3$ dosage for coagulation process

3.2.1 Problem statement

In the deliverable 4.1, a reinforcement learning (RL) framework was developed to optimise ferric chloride ($FeCl_3$) dosing in the coagulation–flocculation process at a DWTP. The study highlighted the limitations of traditional empirical dosing approaches and reactive operational strategies based on delayed turbidity measurements. To address this, the supervised hybrid model developed by KWR for deliverable 4.1 was

introduced to predict turbidity with a six-hour lead time using SCADA data, enabling operators to anticipate water quality changes. Building on this, an RL-based model was proposed to recommend optimal coagulant dosing decisions by integrating predicted turbidity, water quality variables, and flow conditions. The approach demonstrated the potential of combining data-driven forecasting with RL to support more proactive and efficient process control, while remaining adaptable to site-specific conditions through recalibration and reward tuning.

In this present deliverable, this initial RL framework is further developed and refined by comparing different RL methodologies, one with discrete suggestions and one with continuous suggestions. Moreover, improved reward formulations were implemented to enhance decision-making performance and better capture operational objectives in the coagulation-flocculation process.

3.2.2 Data sources and data preprocessing

The data sources used in this work are well presented in the deliverable 4.2.

3.2.3 Materials and methods

Overview

This reinforcement learning (RL) framework aims to optimise the $FeCl_3$ dosage in the coagulation–flocculation process, with the hybrid supervised model (Soft Sensor 6) presented in Deliverable D4.3 acting as the environment. The supervised model provides turbidity predictions that serve as essential feedback for the RL agent during training. In the current implementation, the RL agent can be based on either a Deep Q-learning Network (DQN ; Mnih et al., 2015) or a TD3 (Fujimoto et al., 2018) algorithm. These two approaches will be compared and further refined to improve stability and performance.

Hybrid supervised model for turbidity prediction

The supervised model predicts turbidity at the outlet of the coagulation–flocculation process with a six-hour time lag. It uses input variables including raw water pH, temperature, turbidity, flow in each of the three coagulation–flocculation lanes, and lane number. The model follows a hybrid data-driven and physics-informed approach, as detailed in Deliverables 4.1 and 4.3. In this RL framework, this pre-trained model is used to forecast future water quality conditions, and its turbidity predictions are used as the environmental state for the reinforcement learning (RL) agent. This enables the RL framework to anticipate the impact of different $FeCl_3$ dosing strategies on outlet turbidity and supports more informed operational decision-making.

Deep Q-learning Network (DQN) for coagulant optimization

DQN is a value-based (sequential discrete actions RL) that belongs in the family of the temporal difference (TD) category of RL algorithms(Mnih et al., 2015). It integrates Q-learning with deep neural networks to approximate the state–action value function, $Q(s, a)$, allowing it to learn optimal dosing policies from high-dimensional input data. DQN learns from sequential interactions with the environment, updating its value estimates using the reward received at each time step and the predicted Q-value of the next state. The key elements of the DQN algorithm are described as follows:

1. **Experience:** The agent stores state–action–reward–next state tuples and samples them randomly to break temporal correlations.
2. **Target Network**: This is an additional network that is updated periodically to capture the key state – action relationships.
3. **Epsilon greedy policy**: An epsilon-greedy policy (ε) is used to define the relationship between exploitation - selecting an action from the ones that exist – and exploration –

exploring a new action. The epsilon used in this work was initially set equal to 1 and was then reduced in every new episode by 0.01.

4. **Q-value update**: The agent updates its Q-values based on the observed reward and the estimated future return.

Twin Delayed Deep Deterministic Policy Gradient (TD3)

TD3 is an actor–critic, policy-based RL algorithm designed to address the overestimation bias and instability of earlier deterministic methods like Deep Deterministic Policy Gradient (Fujimoto et al., 2018). It simultaneously learns two critic networks to provide a more accurate estimation of the Q-value and delays policy (actor) updates to improve training stability.

TD3 is particularly suitable for continuous action spaces, such as fine-tuning $FeCl_3$ dosage levels, providing smoother and more precise control compared to discrete-action methods like DQN.

RL model architecture and flow-chart

The developed RL model follows the same steps described in the deliverable 4.2. However, in this new deliverable, a modified reward function is introduced to better capture the operational constrains and the penalties. More specifically, these workflow steps for the identification of the optimal $FeCl_3$ dosage are as follows:

1. **Step 1 – Environment and state generation:** The supervised model (soft sensor no6) acts as the environment in our case, in other words, it simulates the coagulation – flocculation process. The model predicts a turbidity in the coagulation flocculation outlet using the inputs (inlet pH, inlet temperature etc.). These inputs and outputs define the current state in the RL model.

2. **Step 2 – Action selection: The RL agent (DQN or TD3) selects the** optimal $FeCl_3$ dosage **based on the current state.**
   a. For DQN, the action is chosen from a discrete set of possible dosages: $\{1.5, 2, 2.5, 3, 3.5, 4, 5, 6\}$ mg/L
   b. For TD3, the dosage can be continuously adjusted within the defined range of 1.5 to 6 mg/l, allowing finer control**.**

3. **Step 3 – Reward function:** The reward function is crucial part of the methodology as it defines the objective guidance for the RL agents learning process. In this updated version of the RL framework, a different non-linear reward for each agent was introduced to provide smoother guidance and penalise extreme events and high coagulant dosages. The different design of the reward function per RL agent was selected to reflect their different learning mechanisms.

   A) DQN Reward Function

   DQN model operates in discrete action space, and the agent selects from 8 different possible actions – coagulation dosages. The reward promotes turbidity levels in an optimal range (5-6 FTU), penalises high coagulant dosages and turbidity above the threshold of 6 FTU. This structure allows for explicit control of behaviour across different turbidity regions, which suits the value-based and discrete learning nature of DQN. The mathematical formulation of the DQN reward is as follows:

$$DQN_{reward} = \begin{cases} -S * \dfrac{T_{pred} - T_{thresh}}{T_{thresh}}, & if\ T_{pred} > 6 \\ -w_1 \dfrac{|T_{pred} - T_{thresh}|}{T_{thresh}} - w_2 \dfrac{D - D_{min}}{D_{max}}, & otherwise \end{cases}$$

Where:

- T_{pred}: predicted turbidity (FTU)
- T_{thresh}: turbidity threshold (6 FTU)
- S: Severe penalty weight (S=1.5)
- D: the suggested $FeCl_3$ dosage (mg/l)
- D_{max}, D_{min}: the maximum and minimum dosage bounds (6 and 1.5 mg/l respectively)
- W_1 and w_2 weights to balance the contribution of turbidity and dosage contributions in the reward ($w_1 = w_2 = 0.5$)

Rewards were clipped between -20 and 20 to prevent extreme gradients that could destabilize learning.

B) TD3 Reward Function

The TD3 model operates in a continuous action space, allowing smoother adjustments to coagulant dosage. Therefore, a continuous and differentiable reward function was designed to ensure stable gradient propagation during training with a smoother reward function. Therefore, instead of using additional penalty terms when the turbidity threshold is exceeded, a quadratic penalty on turbidity was introduced. This approach ensures stronger negative feedback for higher turbidity values, while the linear dosage penalty discourages excessive coagulant use. In this reward function, the rewards were clipped between -20 and 20 to prevent extreme gradients that could destabilize learning. The mathematical expression of this reward is expressed as follows:

$$\begin{cases} -w1 * {T_{pred}}^2, & if\ T_{pred} > 6 \\ -w_1 * T_{pred} - w_2 * D, & otherwise \end{cases} \quad TD3_{reward} = -w_1 * {T_{pred}}^2 - w_2 * D$$

Where:

w_1 and w_2 weights to balance the contribution of turbidity and dosage contributions in the reward ($w_1 = 1$ $w_2 = 0.5$)

4. **Step 4 – Updating policy:** The RL agent (DQN or TD3) updates its policy based on the received reward. Over multiple episodes, the agent learns to select $FeCl_3$ dosages that achieve the target turbidity with minimal chemical usage. A schematic of the RL flowchart is presented in Figure 27.

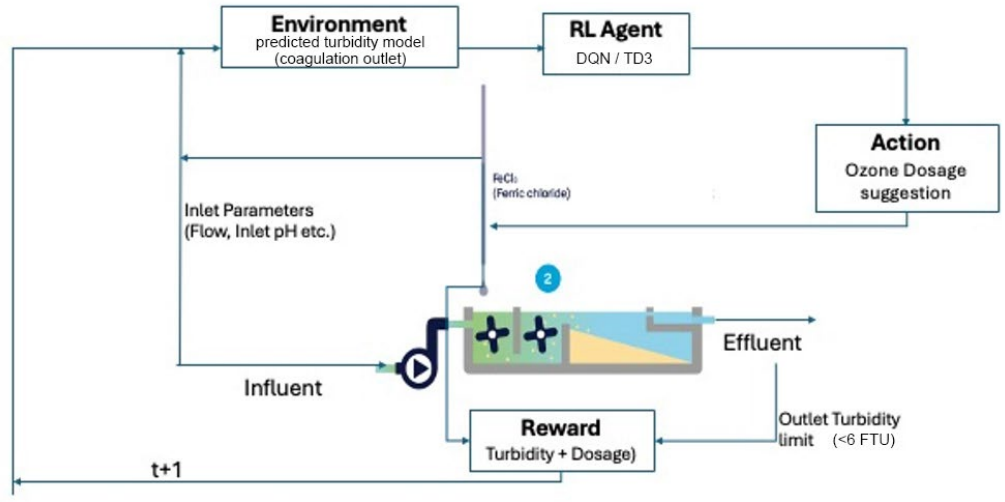

Figure 27: Flow chart of the RL model

3.2.4 Results

The two different RL agents were trained in a dataset of around 4000 samples for roughly 1000 episodes. Once trained, the 3 best performing models per agent in terms of lower reward achieved, were then used in an unseen new dataset of around 1000 samples. The objective was to recommend under known water quality conditions (turbidity in the inlet etc.) a minimum coagulant dosage required to reduce turbidity in the outlet below the threshold of 6FTU, guarantee stable coagulation – flocculation process and protect the filters of the downstream rapid sand filtration process.

In Figure 28, the recommended dosages from the 3 best episodes for both the DQN and TD3 agents are compared with the actual dosages used for the coagulation. This figure shows that TD3 recommended only 2 possible dosage suggestions, which were the upper and the lower limits given to the agent. This result indicates that for this optimisation problem this continuous control agent is not recommended. The TD3 algorithm exhibited limited performance in this study due to the nature of the problem's reward landscape and the characteristics of the underlying process model. Although TD3 is well suited for continuous control tasks, the coagulation dosage optimisation presents a bimodal reward surface, where extreme low or high dosage actions tend to yield locally optimal rewards. Consequently, despite the use of smooth reward formulations to stabilize learning, the TD3 agent consistently converged to boundary actions rather than identifying intermediate, balanced dosages. Furthermore, the supervised turbidity prediction model, used as the environment, was trained predominantly (approximately 95% of the dataset) on data corresponding to discrete coagulant dosages, which likely limited the RL ability to generalize predictions accurately across continuous dosage values. This mismatch between the training distribution and the continuous action space of TD3 further constrained the agent's capacity to learn stable and optimal control policies.

The DQN agent suggests in some cases higher dosages than those measured. However, as shown in the table below, the overall average of the suggested by the DQN agent dosages over the testing period is lower than the average dosage used in the coagulation process without increasing the average turbidity significantly and keeping it in the outlet below the constraint threshold. Moreover, the average response to the spikes is higher than the current response by the DWTP operators indicating that it could adapt into new deteriorated water quality conditions faster than the current approach.

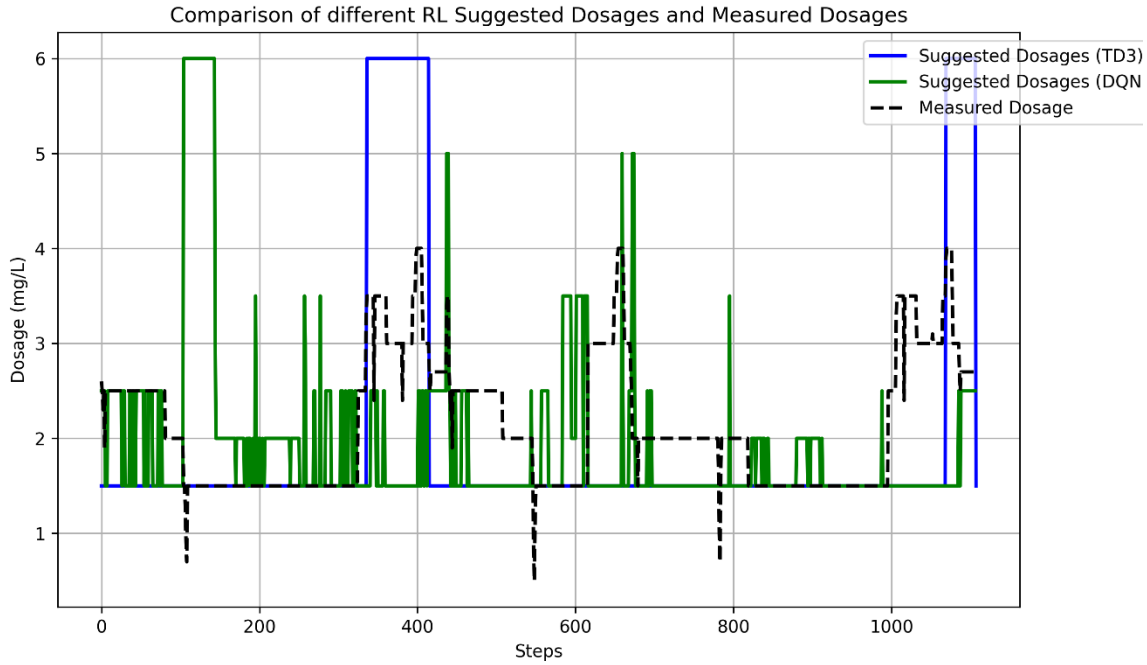

Figure 28: Best RL episode suggested coagulant dosages vs measured coagulant dosages

Table 14: Comparison of models performance

MODEL	Mean Turbidity (FTU)	Average dosage (mg/l)	Response to spikes	High turbidity occurrences (>6FTU)
DQN	**5.035**	**1.949**	**0.369**	201
TD3	5.090	1.975	0.024	301
Current approach	4.332	2.132	0.043	**125**

The subsequent plot in Figure 29 compares the predicted turbidity using the RL agents suggested dosages with the actual turbidity. This figure shows that in comparison to the current approach, if the DQN is selected, the number of times where turbidity exceeds the threshold (6 FTU) is higher. However, it is important to note that these results are based on the predictions of the environment model, which tends to slightly overestimate turbidity comparing to the actual measurements. As part of future work, beyond the scope of this final 4.4 deliverable, a close collaboration with WATERNET process engineers is planned to explore the direct, real-time implementation of the DQN agent in the coagulation – flocculation process without reliance on an environment model. This approach will introduce practical challenges related to the operational stability of the process.

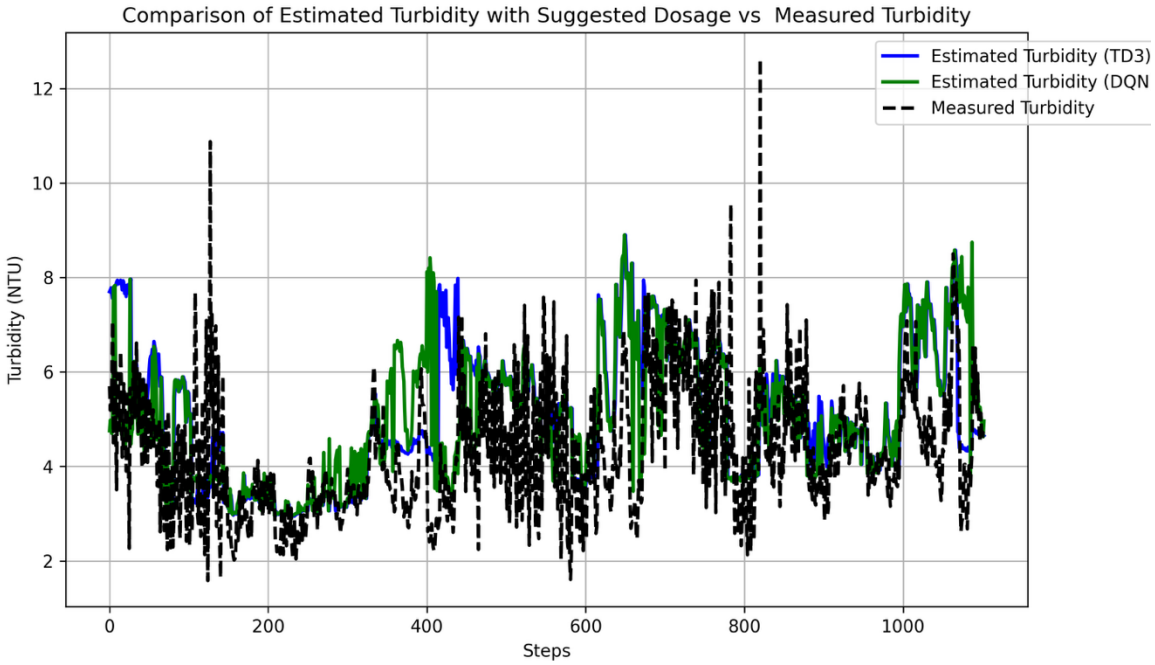

Figure 29: Actual turbidity outputs vs predicted turbidity outputs using the RL suggested dosages

3.2.5 Conclusions and pathway to deployment

An RL model is generated with the aim to optimise the $FeCl_3$ dosage coagulation-flocculation process. Two different RL agents were used, DQN a discrete control agent and TD3 a continuous control agent. The key conclusions of this work are as follows:

- The TD3 agent exhibited limited effectiveness for this application, as it consistently converged to extreme dosage values.
- The DQN agent demonstrated more adaptable and practical dosage recommendations compared to the TD3 agent, which consistently converged to boundary dosage values due to the bimodal reward surface and discrete nature of the training data.
- On average, the DQN agent suggested lower coagulant dosages than those currently applied in practice, achieving comparable turbidity outcomes and indicating potential for more efficient chemical use. Thus, this RL agent, using the supervised model presented in in the deliverable 4.3 as its environment, is the recommended approach to Waternet for supporting the management of the coagulation-flocculation process.
- Future work, beyond the scope of the deliverable 4.4, will focus on the real-time implementation of the trained RL agent without the use of the environment model.

4. Conclusions Upscaling and European Added Value

4.1 Conclusions

Water treatment plants are critical components of the water supply chain, omnipresent in every urban water system in Europe, irrespective of their complexity and size. The ToDrinQ project attempts to advance the state of play in the management and optimization of DWTPs, by leveraging approaches and techniques from the realm of RL. The four RL models developed provide real-time, data-driven recommendations for optimizing critical treatment processes such as coagulation–flocculation, pre-chlorination, and post-disinfection, appear in any typical DWTP.

On the one hand, by ensuring precise chemical dosing, the RL models enable to maintain water quality parameters within regulatory and safety limits, and hence they enhance the reliability of treated water output, directly benefiting public health and consumer trust. On the other hand, the models enable the derivation of optimal dosages under raw water quality conditions, which is critical for ensuring resilience in water supply systems.

Finally, the project bridges a significant gap by introducing RL methodologies to drinking water treatment, a domain where such approaches remain underexplored compared to wastewater treatment. The technical innovations generated by ToDrinQ can serve as a foundation for future research and development in **smart water treatment**.

Three of the four RL models developed, specifically the pre-disinfection and post-disinfection models at Polydendri, and the DQN-based coagulation model at Leiduin, successfully demonstrated stable training and alignment with historical operator behaviour, with the two disinfection models achieving meaningful reductions in chemical dosage. The Athens coagulation model (RL Model 2) demonstrated stable training convergence but did not achieve full policy alignment, a limitation attributed to sensor unavailability and the resulting simplification of the reward function. This outcome should be acknowledged as an open challenge rather than a resolved result.

At the level of methodology, these four implementations collectively demonstrate that the RL development framework is transferable to diverse DWTP settings. The underlying approach, using plant-specific SCADA/logbook data to formulate and train an offline RL agent benchmarked against an Approximation Model, is data-agnostic and has been validated across two plants with different coagulants, instrumentation maturity levels, and process configurations.

At the same time, they are modular and hence they can be deployed individually depending on the treatment process exist in a treatment unit. These pilot implementations serve as proof-of-concept examples that can inspire adoption by other utilities and encourage private sector investment in similar technologies.

Finally, the RL models use data streams from SCADA systems, and this compatibility facilitates seamless integration without requiring significant changes in infrastructure, making adoption more feasible for a wide range of operators. It is also worth mentioning that the models will be standardised according to FIWARE standards (in the context of Task 4.5 and integration with legacy system in the context of Task 7.5), which make them directly compatible with any FIWARE-enabled architecture. This increases substantially their potential for further upscaling and transferability.

4.2 Pathway to upscaling

The Reinforcement Learning (RL) models developed in Task 4.4 have successfully achieved Technology Readiness Level (TRL) 5 (technology validated in a relevant environment). ToDrinQ has defined a specific upscaling pathway that leverages the project's integrated architecture:

- **Integration into the Modular Platform (WP7):** The finalized algorithms presented in this deliverable will be containerized and integrated into the ToDrinQ Modular Platform (Task 7.4). This transition moves the models from standalone analytical scripts to deployed microservices accessible via a unified dashboard (NESSIE).
- **Standardization via FIWARE**: To ensure technical scalability, the RL modules utilize FIWARE smart data models and open APIs (Task 7.2). This adherence to EU-curated open standards ensures that the advice provision modules are interoperable with legacy SCADA systems across Europe, significantly lowering the technical barrier to entry for other utilities.
- **Human-in-the-Loop Validation (WP2):** The transition to the Demonstration phase (WP2) introduces operator feedback loops. This safeguards the upscaling process by ensuring the AI remains an assistive tool ("Advice Provision") rather than a black-box controller, aligning with the project's safety-first approach and the risk management requirements of the Drinking Water Directive.

4.3 European Added Value and EU Policy Alignment

The Reinforcement Learning (RL) modules developed in ToDrinQ directly support the implementation of the revised Drinking Water Directive, specifically regarding the mandatory Risk-Based Approach to Water Safety (Article 6). **By transitioning from reactive control to proactive, data-driven advice provision**, these modules enable utilities to:

- **Enhance Operational Resilience:** The RL agents predict and mitigate process deviations (e.g., turbidity spikes) before they breach quality thresholds, ensuring consistent compliance with safety parameters (Annex I).
- **Minimize Disinfection By-Products (DBPs):** By optimizing the precise dosage of disinfectants (chlorine/sodium hypochlorite) based on real-time water quality variations, the models reduce the formation of harmful by-products like THMs and HAAs, directly addressing the Directive's stricter observation list requirements.
- **Secure Supply Continuity:** The "human-in-the-loop" design ensures that operators can make evidence-based decisions rapidly during determining events (e.g., extreme weather impacting source water), safeguarding the supply of wholesome water to consumers.

The ToDrinQ RL modules directly support key European strategic objectives defined in the Horizon Europe call HORIZON-CL6-2022-ZEROPOLLUTION-01:

- **Zero Pollution Action Plan:** By optimizing coagulant ($FeCl_3$, Alum) and disinfectant dosing, the models reduce chemical consumption without compromising water quality. This directly contributes to the Zero Pollution ambition by minimizing the environmental footprint of water treatment and reducing the formation of Disinfection By-Products (DBPs).
- **Revised Drinking Water Directive (EU 2020/2184):** The models enhance the reliability of treated water, supporting the Directive's mandate for a risk-based approach to water safety. The ability to predict and control turbidity spikes ensures higher resilience against climate-induced water

quality variations (e.g., extreme weather events affecting raw water sources), a core objective of the ToDrinQ toolkit.

- **Digital Transformation of the Water Sector:** The project bridges the gap between academic AI research and industrial water treatment. By demonstrating safe, offline RL, ToDrinQ fosters a digitalized, data-driven water sector capable of meeting future regulatory challenges with advanced monitoring and control tools.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. http://arxiv.org/abs/1606.06565

Álvarez Díez, A., Pena Rois, R., Mocanu, I., Orzan, C., Brebenel, C., Stere, J., . . . Fernández Montenegro, J. (2024). Reinforcement learning-based DSS for coagulant and disinfectant dosage selection on drinking water treatment plants. Water Supply, 24, 86-102. doi:10.2166/ws.2023.328

Bae, H., Kim, S., & Kim, Y. (2006, February 1). Decision algorithm based on data mining for coagulant type and dosage in water treatment systems. Water Science and Technology, 53, 321-329. doi:10.2166/wst.2006.137

Baouab, M., & Cherif, S. (2018, July 30). Prediction of the optimal dose of coagulant for various potable water treatment processes through artificial neural network. Journal of Hydroinformatics, 20, 1215-1226. doi:10.2166/hydro.2018.014

Boumezbeur, H., Laouacheria, F., Heddam, S., & Djemili, L. (2023, May 12). Modelling coagulant dosage in drinking water treatment plant using advance machine learning model: Hybrid extreme learning machine optimized by Bat algorithm. Environmental Science and Pollution Research, 30, 72463-72483. doi:10.1007/s11356-023-27224-6

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.

Croll, H. C., Ikuma, K., Ong, S. K., & Sarkar, S. (2023). Systematic Performance Evaluation of Reinforcement Learning Algorithms Applied to Wastewater Treatment Control Optimization. *Environmental Science and Technology*, *57*(46), 18382–18390. https://doi.org/10.1021/acs.est.3c00353.

Chen, K., Wang, H., Valvelde-Perez, 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.

Delgrange, N., Cabassud, C., Cabassud, M., Durand-Bourlier, L., & Lainé, J. (1998, November). Neural networks for prediction of ultrafiltration transmembrane pressure – application to drinking water production. Journal of Membrane Science, 150, 111-123. doi:10.1016/s0376-7388(98)00217-8

Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). *Challenges of Real-World Reinforcement Learning*. http://arxiv.org/abs/1904.12901

Fujimoto, S., & Gu, S. S. (2021). *A Minimalist Approach to Offline Reinforcement Learning*. http://arxiv.org/abs/2106.06860

Fujimoto, S., van Hoof, H., & Meger, D. (2018). *Addressing Function Approximation Error in Actor-Critic Methods*. http://arxiv.org/abs/1802.09477

Gagnon, C., Grandjean, B. P. A., & Thibault, J. (1997). Modelling of coagulant dosage in a water treatment plant. In *Arrijicid Intelligence in Engineering* (Vol. 11).

Gomes, L. S., Souza, F. A. A., Pontes, R. S. T., Neto, T. R. F., & Araújo, R. A. M. (2015). Coagulant Dosage Determination in a Water Treatment Plant Using Dynamic Neural Network Models. *International Journal of Computational Intelligence and Applications*, *14*(3). https://doi.org/10.1142/S1469026815500133

Griffiths, K. A., & Andrews, R. C. (2011). The application of artificial neural networks for the optimization of coagulant dosage. *Water Science and Technology: Water Supply*, *11*(5), 605–611. https://doi.org/10.2166/ws.2011.028

Gu, S., Knoll, A., & Jin, M. (2024). TeaMs-RL: Teaching LLMs to Teach Themselves Better Instructions via Reinforcement Learning. *arXiv preprint arXiv:2403.08694*.

Haghiri, S., Daghighi, A., & Moharramzadeh, S. (2018). Optimum coagulant forecasting by modeling jar test experiments using ANNs. *Drinking Water Engineering and Science*, *11*(1), 1–8. https://doi.org/10.5194/dwes-11-1-2018

Helm, W., Zhong, S., Reid, E., Igou, T., & Chen, Y. (2024). Development of gradient boosting-assisted machine learning data-driven model for free chlorine residual prediction. Frontiers of Environmental Science & Engineering, 18, 17. doi:10.1007/s11783-024-1777-6

Kim, C., & Parnichkun, M. (2017). MLP, ANFIS, and GRNN based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant. Journal of Water Supply: Research and Technology - Aqua, 66, 49-61. doi:10.2166/aqua.2016.022

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*(6), 4909-4926.

Koch, Lucas, Dennis Roeser, Kevin Badalian, Alexander Lieb, and Jakob Andert. "Cloud-Based Reinforcement Learning in Automotive Control Function Development." *Vehicles* 5, no. 3 (2023): 914-930.

Kulkarni, P., & Chellam, S. (2010, September 1). Disinfection by-product formation following chlorination of drinking water: Artificial neural network models and changes in speciation with treatment. Science of The Total Environment, 408, 4202-4210. doi:10.1016/j.scitotenv.2010.05.040

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). *Continuous control with deep reinforcement learning*. http://arxiv.org/abs/1509.02971.

Leng, Jinling, Xingyuan Wang, Shiping Wu, Chun Jin, Meng Tang, Rui Liu, Alexander Vogl, and Huiyu Liu. "A multi-objective reinforcement learning approach for resequencing scheduling problems in automotive manufacturing systems." *International Journal of Production Research* 61, no. 15 (2023): 5156-5175.

Maier, H., Morgan, N., & Chow, C. (2004, May). Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. Environmental Modelling & Software, 19, 485-494. doi:10.1016/s1364-8152(03)00163-4

Milot, J., Rodriguez, M., & Sérodes, J. (2002, September). Contribution of Neural Networks for Modeling Trihalomethanes Occurrence in Drinking Water. Journal of Water Resources Planning and Management, 128, 370-376. doi:10.1061/(asce)0733-9496(2002)128:5(370)

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015a). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran,

D., Wierstra, D., Legg, S., & Hassabis, D. (2015b). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mohammadi, E., Stokholm-Bjerregaard, M., Hansen, A. A., Nielsen, P. H., Ortiz-Arroyo, D., & Durdevic, P. (2024). Deep learning based simulators for the phosphorus removal process control in wastewater treatment via deep reinforcement learning algorithms. *Engineering Applications of Artificial Intelligence*, *133*. https://doi.org/10.1016/j.engappai.2024.107992

Park, J., Lee, C., Cho, K., Hong, S., Kim, Y., & Park, Y. (2018). Modeling trihalomethanes concentrations in water treatment plants using machine learning techniques. DESALINATION AND WATER TREATMENT, 111, 125-133. doi:10.5004/dwt.2018.22353

Patel, K. M. (2023). A practical Reinforcement Learning implementation approach for continuous process control. *Computers and Chemical Engineering*, *174*. https://doi.org/10.1016/j.compchemeng.2023.108232

Peleato, N. M., Legge, R. L., & Andrews, R. C. (2018). Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. *Water Research*, *136*, 84–94. https://doi.org/10.1016/j.watres.2018.02.052

Plappert, M., Andrychowicz, M., Ray, A., McGrew, B., Baker, B., Powell, G., Schneider, J., Tobin, J., Chociej, M., Welinder, P., Kumar, V., & Zaremba, W. (2018). *Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research*. http://arxiv.org/abs/1802.09464

Puiutta, E., & Veith, E. M. (2020). *Explainable Reinforcement Learning: A Survey*. http://arxiv.org/abs/2005.06247.

Sewak, Mohit (2019). Deep Reinforcement Learning, Frontiers of Artificial Intelligence. Springer Nature Singapore, https://doi.org/10.1007/978-981-13-8285-7.

Singh, K. P., & Gupta, S. (2012). Artificial intelligence based modeling for predicting the disinfection by-products in water. *Chemometrics and Intelligent Laboratory Systems*, *114*, 122–131. https://doi.org/10.1016/j.chemolab.2012.03.014.

Singh, P., & Yadav, A. (2021). "Reinforcement learning in water management and sustainability." Water, 13(17), 2372.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Dieleman, S., 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, *529*(7587), pp.484-489.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

Syafiie, S., Tadeo, F., Martinez, E., & Alvarez, T. (2011). Model-free control based on reinforcement learning for a wastewater treatment problem. *Applied Soft Computing Journal*, *11*(1), 73–82. https://doi.org/10.1016/j.asoc.2009.10.018

Tang, S., & Wiens, J. (2021). Model Selection for Offline Reinforcement Learning: Practical Considerations for Healthcare Settings. In K. Jung, S. Yeung, M. Sendak, M. Sjoding, & R. Ranganath (Ed.), Proceedings of the 6th Machine Learning for Healthcare Conference. 149, pp. 2–35. PMLR. Retrieved from https://proceedings.mlr.press/v149/tang21a.html

Van der Helm, A. W., & Rietveld, L. C. (2002). Modelling of drinking water treatment processes within the Stimela environment. Water Science and Technology: Water Supply, 2, 87–93.

Voloshin, C., Le, H. M., Jiang, N., & Yue, Y. (2021). Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning. Retrieved from https://arxiv.org/abs/1911.06854

Wang, Y., Jiang, Z., & Jiang, J. (2018). "Data-driven predictive control for water treatment process based on reinforcement learning." IEEE Transactions on Industrial Electronics, 66(4), 2815-2823.

Wu, G.-D., & Lo, S.-L. (2008, December). Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. Engineering Applications of Artificial Intelligence, 21, 1189-1195. doi:10.1016/j.engappai.2008.03.015

Zhang, K., Achari, G., Li, H., Zargar, A., & Sadiq, R. (2013). Machine learning approaches to predict coagulant dosage in water treatment plants. *International Journal of System Assurance Engineering and Management*, *4*(2), 205–214. https://doi.org/10.1007/s13198-013-0166-5

Zhang, Q., & Stanley, S. J. (1999). REAL-TIME WATER TREATMENT PROCESS CONTROL WITH ARTIFICIAL NEURAL NETWORKS. In *JOURNAL OF ENVIRONMENTAL ENGINEERING* (Issue 153).

Zheng, H., Ding, R., & Li, X. (2020). "A survey of reinforcement learning applications in industrial automated control." IEEE Transactions on Industrial Informatics, 17(5), 3265-3275.

Zhu, Z., Lin, K., Jain, A. K., & Zhou, J. (2020). *Transfer Learning in Deep Reinforcement Learning: A Survey*. http://arxiv.org/abs/2009.07888.

The revised EU Drinking Water Directive promotes a risk assessment and risk management approach for securing drinking water supply in the context of climate change and increased pollution. However, this approach is challenged by insufficient information that is available to operators, especially in real time, on compounds and organisms of emerging concern, such as pesticides, pharmaceuticals, disinfection by-products, heavy metals and pathogenic microorganisms. We argue that if drinking water treatment could leverage novel technologies and design philosophies, and more agile operational actions could be supported, drinking water supply systems could become more adaptable and robust without expensive infrastructural investments. In this context, ToDrinQ develops and tests a compendium of modular, complementary, innovative solutions (the 'ToDrinQ Toolkit') that provide new information and better support tools to operators and designers to adapt to (short- and long-term) changes in water quality, while obtaining high drinking water quality at the tap. ToDrinQ develops novel real time sensing and water quality monitoring technologies, innovative treatment systems (especially suitable for small-scale/modular, adaptable treatment plants) and interoperable decision tools that support resilient, evidence-based treatment plant design and improved overall water system operational awareness and response.

ToDrinQ

Funded by
the European Union