Reinforcement Learning-driven advice provision module for advice provision based on Soft Sensors: beta version

Deliverable 4.2

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	Reinforcement Learning-driven advice provision module for advice provision based on Soft Sensors: beta version		
ACTUAL DATE OF DELIVERY	29-11-2024		
NATURE	R	DISSEMINATION LEVEL	Sensitive
LEAD BENEFICIARY	National Technical University of Athens		
RESPONSIBLE AUTHOR	Prof. Christos Makropoulos, NTUA		
CONTRIBUTIONS FROM	Iosif Spartalis, NTUA Greg Kyritsakas, TUD Jasper Imnik, KWR Panagiotis Kossieris, NTUA George Bariamis, NTUA		

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	Review	Lydia Vamvakeridou-Lyroudia	KWR Water Research Institute
V0.3	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
V0.4	Review	Christos Makropoulos, NTUA	National Technical University of Athens
V1.0	Final version	Luuk Rietveld, TUD	Delft University of Technology

Quality control

Author	Organisation short name	Role	Date
Christos Makropoulos	National Technical University of Athens	Deliverable Leader	01-11-2024
Christos Makropoulos	National Technical University of Athens	Work Package Leader	07-11-2024
Lydia Vamvakeridou-Lyroudia	KWR Water Research Institute	Reviewer 1	19-11-2024
Luuk Rietveld	Delft University of Technology	Scientific project coordinator	29-11-2024
Danitsja van Heusden	Delft University of Technology	Project coordinator	29-11-2024

Abbreviations

AWD	-	Amsterdam Water supply Dunes
BC	-	Behavioural Cloning
CA	-	Consortium Agreement
Cl	-	Chlorine
DC	-	Demo Case
DCL	-	Demo Case Leader
DDPG	-	Deep Deterministic Policy Gradient
DoA	-	Description of the Action
DPG	-	Deterministic Policy Gradient
DQN		Deep q-learning network
DWTP	-	Drinking Water Treatment Plant
EAB	-	External Advisory Board
EC	-	European Commission
EU	-	European Union
$FeCl_3$		Ferric Chloride
GA	-	General Assembly
LLM		Large Language Model
MAE	-	Mean absolute error
MDP	-	Markov Decision Process
ML		Machine Learning
NOM	-	Natural Organic Matter
PC	-	Project Coordinator
RL	-	Reinforcement learning
SC	-	Steering Committee
SCADA	-	Supervisory Control and Data Acquisition
TD		Temporal Difference
TD3-BC	-	Twin Delayed Deep Deterministic Policy Gradient
THMs	-	Trihalomethanes
WP	-	Work Package
WPL	-	Work Package Leader

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 to end users. However, the large-scale complexity of WTPs, along with the variability in the quality of raw water, imposes significant operational challenges. To address these issues, the ToDrinQ project is developing advanced tools and services to support plant operators in the optimal management of DWTPs by delivering real-time, stage-specific treatment advice.

The project leverages Reinforcement Learning (RL), a branch of machine learning in which agents learn to make sequential decisions by interacting with an environment to maximize cumulative rewards. While RL applications in wastewater treatment have been extensively explored, their use in WTP operations remains nascent. In this context, we can argue that ToDrinQ developments go well beyond the current state of play in the automated management and control of DWTP processes.

This deliverable focuses on the development of beta versions of four RL models under Task 4.4 of Work Package 4 (WP4). These models are designed to optimize key treatment processes in two demonstration cases: the **Leiduin plant** serving the city of Amsterdam (DC#1) and the **Polydendri plant** serving the metropolitan area of Athens (DC#2).

The Leiduin plant supplies approximately 70% of the water for the Amsterdam area, treating river water from the Lek Canal supplemented by natural dune water. The focus is on optimizing the **coagulation–flocculation process**, which is critical for destabilizing particles and forming flocs that can be removed in subsequent filtration stages. 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. Among the various processes in a DWTP, coagulation–flocculation plays a pivotal role in destabilizing particles and facilitating their aggregation into larger flocs, which can be removed in subsequent filtration stages. The success of this process depends heavily on the dosage of the coagulant, which in this case is ferric chloride ($FeCl_3$). In this context, the RL model for this process (RL Module 4) aims to optimize the dosage of $FeCl_3$, and specifically to suggest the optimal dosage for the next six-hour interval using as inputs water quality and water flow data, provided by the SCADA system, along with prediction on turbidity, as obtained by the relevant prediction model (soft sensor) developed within Task 4.3. Particularly, input data includes temperature, turbidity and pH of raw water, water flow in coagulation–flocculation lanes, turbidity at the coagulation–flocculation inlet, while recorded $FeCl_3$ dosages are used for model training. In the current form, the RL model uses the Deep Q-Learning Network, while additional approaches will be tested during the third year of the project to improve model performance.

With respect to the Athens demo case, the focus is on the Polydendri WTP, that is one of the four DWTPs which serve the metropolitan area of Athens (Greece). The plant consists of two identical units, each with a nominal capacity of 100,000 m^3 per day. 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.

For this case, three RL models are being developed to optimize critical water treatment processes, and specifically the pre-chlorination, coagulation, and disinfection. With respect to pre-chlorination (RL model #1), which takes place at pre-disinfection/oxidation stage, an RL model is being developed to provide advice on the optimal chlorine dosage (the disinfectant dosage) using as inputs critical parameters, such as the turbidity and temperature of the raw water at the DWTP main inlet, the free Cl concentration at de-gritting state and at the filter's inlet. With respect to coagulation (RL model #2), an RL model is being

developed to provide advice on the optimal the aluminum sulphate (coagulant) dosage, maintaining turbidity and free aluminum concentration within safe operational limits. The RL model uses as inputs the turbidity and temperature of the raw water, the turbidity at filters inlet, the free Alum at clear water tank inlet, and the free Cl at degritting stage. Finally, RL model #3 targets the post-disinfection process, which takes place after the filtration stage and before the clear water tanks, and provides advice on the optimal chlorine dosage. In all three RL models developed for Athens demo case, the Twin Delayed Deep Deterministic Policy Gradient with Behaviour Cloning algorithm has been used.

The four above-mentioned beta versions of RL models will be further developed and improved during the third year of the project. This includes fine-tuning of algorithms, re-training with new data, as well as further experimentation with alternative reward-functions and input parameters.

1. Introduction

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 behaviors. 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).

This deliverable includes beta versions of the Reinforcement Learning (RL) algorithms developed as part of Task 4.4 in Work Package 4 (WP4). These algorithms are designed for future integration into a model that will support operators of drinking water treatment plants (DWTPs) by providing operational guidance. The goal is to enhance and optimize specific treatment processes within these facilities.

Within this task three reinforcement learning (RL) agents are being developed by the partner NTUA to optimize critical water treatment processes: pre-chlorination, coagulation, and disinfection. Each agent is designed with a model-free approach, utilizing data from direct sensors and SCADA system of the Polydendri DWTP (Demo Case -DC #2) to capture real-time water quality parameters, enabling the optimization of dosing strategies and process adjustments without needing an explicit model of the treatment plant. These agents focus on achieving specific performance goals defined by plant operators, such as minimizing chemical usage and achieving target water quality standards. By continuously learning from the plant's operational data, each agent dynamically adapts its recommendations to improve treatment efficiency and operational stability under varying conditions.

In parallel, KWR has developed a RL agent to optimize ferric chloride dosing specifically for the coagulation stage. This agent leverages real-time data from the plant's 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.

1.1 RL and DWTP processes

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 labeled input data, and unsupervised learning aims to discover patterns or clusters within unlabeled 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).

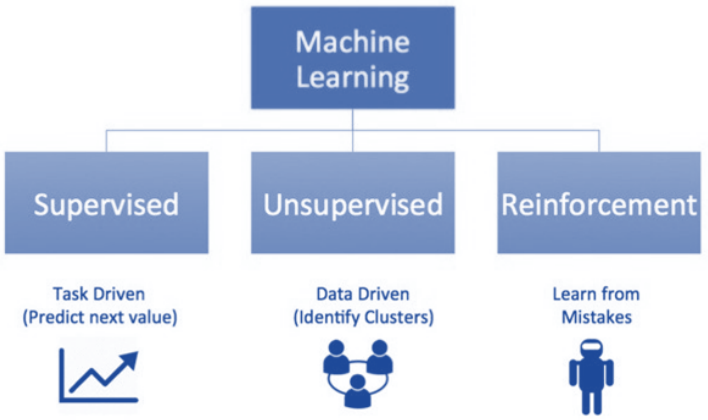

Figure 1: Types of Machine Learning

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.

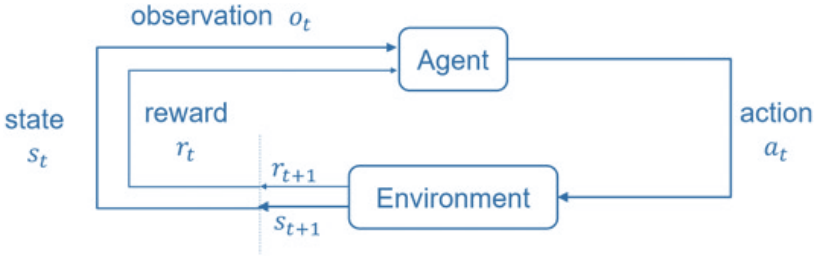

Figure 2: RL algorithm main components

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 behaviors, 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 defense. 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.

In the present work, many of the challenges associated with applying RL in the drinking water treatment sector are addressed. 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 (DC#2)

2.1 Demo case description (Polydendri DWTP)

As a DC for developing the RL model the Polydendri Drinking Water Treatment Plant (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 Supervisory Control and Data

Acquisition (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 and the concentration of free aluminium are 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.

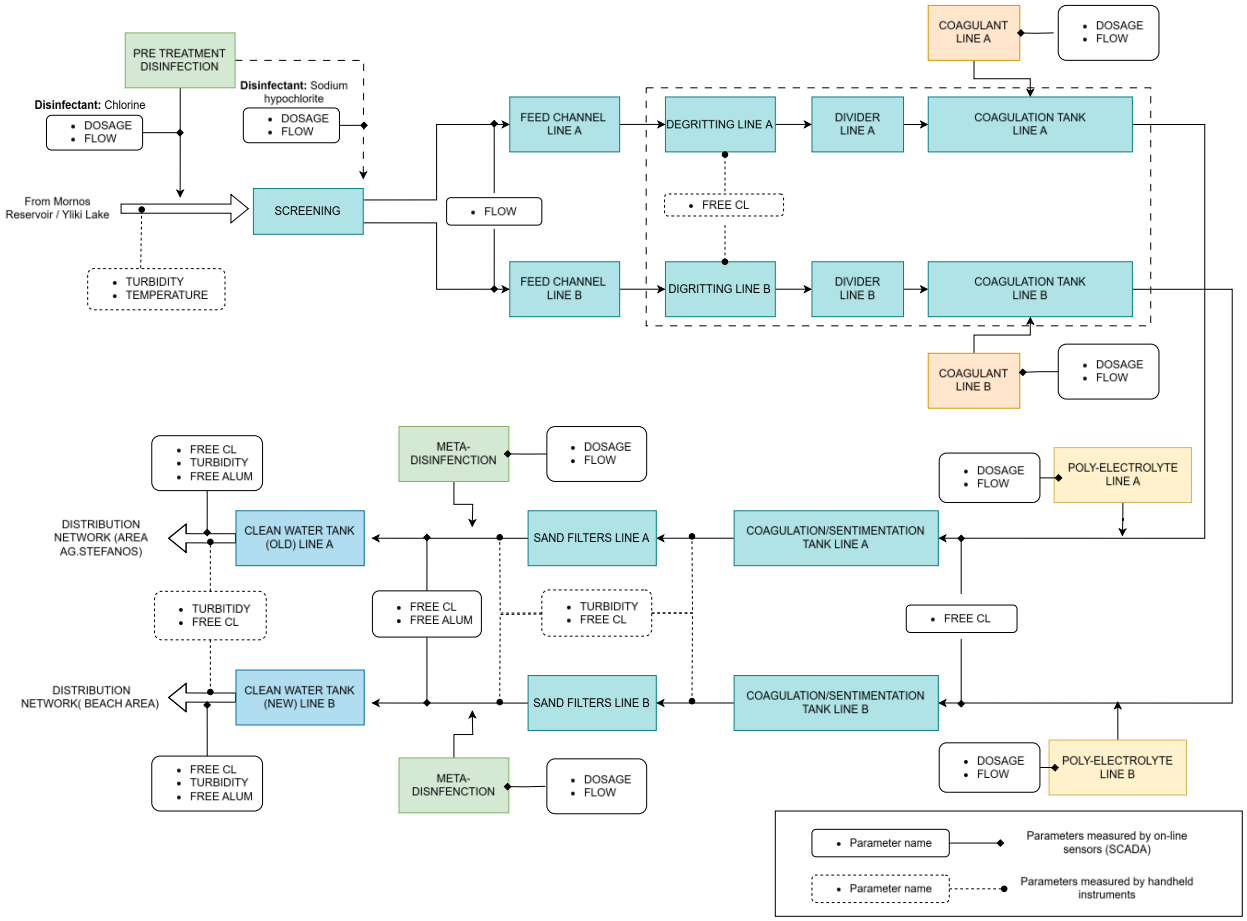

Figure 4: Polydendri DWTP main treatment processes and monitored parameters map

Three key processes were selected as appropriate use cases for the development of RL algorithms: pre-disinfection, coagulation, 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. The

subsequent step focused on assessing the availability of data for these parameters. 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.

As a start, one month of available data (November of 2018) was retrieved from the SCADA system and the corresponding logbooks were manually scanned and digitized to complement the data collection process. This initial dataset has been established as the foundation for developing the beta versions of the reinforcement RL use cases. Because the drinking water treatment processes are low varying and highly stable procedures, the data points collected in just one month's period were vary sparse. In the current stage of RL development, the parameters from EYDAP's Chemistry Laboratory have not been included, primarily due to their infrequent recording intervals.

2.2.2 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 overcome these challenges, the use of simulation software as an RL environment presents a viable alternative. However, most available simulation tools for water treatment processes are licensed commercially, with significant costs associated, limiting their accessibility. An exception to this is the open-source software Stimela (Van der Helm & Rietveld, 2002). An alternative to this is the adoption of a purely data-driven approach, which enables the description of system dynamics by exploring the patterns available in the collected data. Specifically, in the development of beta RL prototypes we utilized real-world data from the DWTP, while conducting the RL agent's training offline. In this alternative approach, the agent is trained by analysing historical data from the DWTP, along with the corresponding actions taken by the plant's operators.

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 behavior. 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 and accuracy 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 meta-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 beta versions of the three use cases, was the Twin Delayed Deep Deterministic Policy Gradient with Behaviour Cloning (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 shown in Figure 5below.

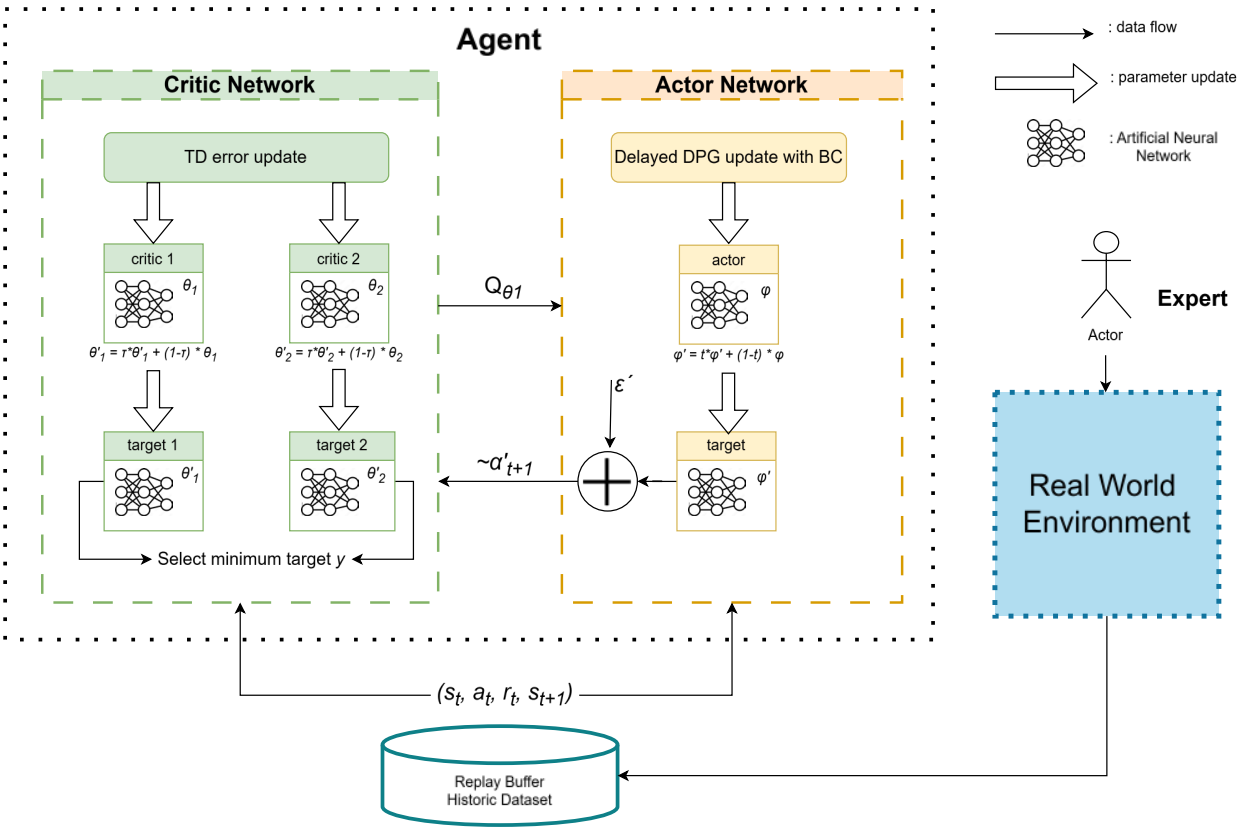

Figure 5: The TD3-BC Architecture

The BC component works by adding a regularization term that minimizes the difference between the actions chosen by the learned policy and those in the offline dataset. This regularization is scaled by a hyperparameter, which balances between imitating the behaviour in the dataset and pursuing the reward-driven objective of the original TD3 algorithm. As a result, TD3-BC can leverage TD3's stability and robustness while being more suitable for offline tasks where exploration is limited to past experiences.

The training procedure begins by sampling a batch of state transitions from the Replay Buffer, which stores historical data collected from expert human interactions within the real-world environment. These state transitions consist of the MDP state at time t, denoted as s_t, the corresponding action at time t, denoted as a_t, the reward received at time t, r_t, and the subsequent MDP state at time t, s_{t+1}. The parameters of the Critic Neural Networks are denoted as θ_1, θ_2, while the parameters of the target networks are represented by θ_1' and θ_2'. Similarly, φ and φ' represent the parameters of the actor and its corresponding target network. τ serves as the "soft" update parameter, controlling the rate at which target networks are updated. To compute the temporal-difference (TD) error for the action-value Q at the next state s_{t+1}, the action a_{t+1} is determined by the actor target network with the addition of a small Gaussian noise ε'. For the Deterministic Policy Gradient (DPG) update, however, the action-value Q is directly provided by the Critic 1 network.

2.2.3 Development of the RL models workflow

As previously discussed, formulating both the MDP state and reward function presents significant challenges in real-world RL applications. Determining these parameters requires an iterative approach involving the training and evaluation of the Agent's performance. Within this iterative process, there exists an additional optimization loop in which specific hyperparameters require fine-tuning. These include the learning rate of the Deterministic Policy Gradient (DPG) and TD error update, the dimensions of the hidden layers in the Actor-Critic network, parameters of the Behavioural Cloning (BC) regularization component, among others.

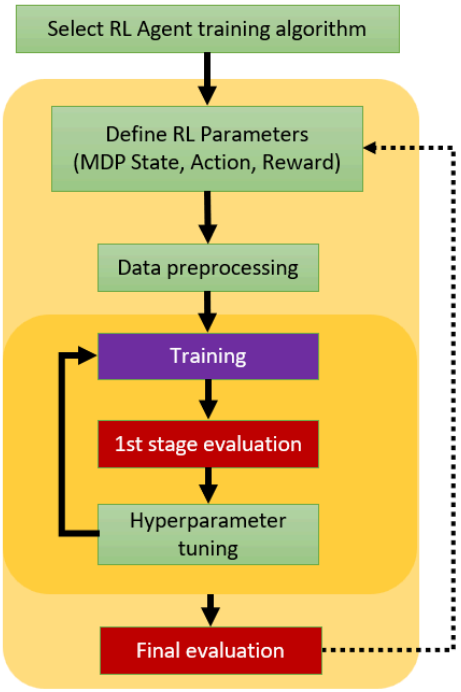

Figure 6: Evaluation process workflow

The evaluation within this inner loop, where RL hyperparameter tuning is performed, is referred to as the first stage evaluation. In contrast, the outer evaluation process, which assesses the MDP state formulation and reward function, is designated as the Final Evaluation. The entire workflow of the RL development process, detailing these stages, is shown in Figure 6. Since real-world DWTP processes serve as RL environments and RL agent training is conducted offline, evaluations for each use case are performed at both the initial and final stages using dedicated Random Forest regression models (Breiman, 2001) as RL environment approximation models (AMs) (Voloshin, Le, Jiang, & Yue, 2021). These models are trained on a reserved holdout portion (20%) of the historical dataset specific to each use case. This methodology facilitates a comparative assessment of the RL agent's performance against the current operational policy implemented by DWTP operators (Tang & Wiens, 2021).

To formulate the beta versions for the three RL use cases, a common approach has been adopted, treating each case as a problem in industrial continuous process control rather than as an episodic MDP. Unlike typical episodic MDPs, which are common in RL applications such as video or board games and consist of multiple time steps, continuous process control lacks a terminal state with a significant reward or penalty (Patel, 2023). In episodic MDPs, reaching a terminal state with a high reward or penalty back-propagates this reward to previous states, aided by the discount factor γ to enable learning. This absence of terminal rewards in continuous control tasks necessitates a different strategy for learning and optimization.

In continuous process control, the primary objective is to identify and maintain a safe, steady-state operating point as efficiently as possible. The system must adapt to disturbances and dynamically changing process characteristics over time, while also aiming to maximize operational profitability. The RL agent, therefore, is required to maintain robust, stable control under varying conditions rather than simply maximizing reward based on terminal states. Considering these factors, the MDP states and reward functions for each RL use case were designed with specific attention to maintaining critical parameters within defined safety limits throughout each DWTP process (Patel, 2023). Additionally, given the absence of a terminal state in the MDP framework, a discount factor γ of 1 was applied uniformly across all three use cases. This choice reflects the continuous nature of the process control problem, where learning is not reliant on terminal rewards but instead focuses on sustaining steady-state operation while navigating real-time variations and disturbances in the system.

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

2.3.1 Problem statement

The pre-disinfection/oxidation stage in *the* Polydendri DWTP occurs immediately before the major intake structure (SCREENING as depicted in Figure 4), utilizing chlorine as the primary disinfectant. Alternatively, sodium hypochlorite can be used as a disinfectant. In such cases, the disinfectant mixture is introduced just before the water is divided into the two separate processing lines. Critical to the successful operation of this process is maintaining adequate free chlorine levels further downstream. These levels are measured at two distinct points—specifically at the *de*-gritting stage and the filter inlet—using handheld instruments at three-hour intervals. The free chlorine concentration, expressed in mg/L, must remain within predefined limits (Table 1) to ensure effective disinfection. At present, the disinfectant dosage is determined reactively by the DWTP operators, based solely on the concentrations of free chlorine measured at the downstream locations mentioned earlier. The objective of the RL algorithm is to optimize the chlorine dosage, set in parts per million (ppm), to consistently maintain free chlorine levels within these critical limits, thus ensuring both operational efficiency and water safety.

2.3.2 Data sources and data preprocessing

The primary sources of data used for the RL models originate from the existing DWTP monitoring system. While numerous parameters are monitored at intervals of less than one minute by online sensors – in this use case the disinfectant dosage - and stored in the DWTP's SCADA system, critical parameters are measured using handheld instruments at three-hour intervals and recorded in handwritten logbooks. Those critical parameters are the turbidity and the temperature of the raw water at the DWTP main inlet, the free Cl concentration at de-gritting state and at the filter's inlet. For this study, we utilized one month of SCADA data from November 2018, during which Yliki Lake served as the primary water supply. Additionally, the handwritten logbooks from that period were photographed, manually digitized, and the relevant parameters were extracted for analysis.

Table 1 presents the retrieved parameters relevant to the pre-disinfection process, along with associated contextual information. Notably, critical parameters such as raw water pH and conductivity - essential for the chemical disinfection process—are recorded once a day by EYDAP's Chemistry Laboratory and they were not utilized in the current RL beta version due to the low recording frequency.

Table 1: Pre-disinfection 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
Raw water Temperature	External Variable	Celsius	0.1	-	-	3 hours	Logbook
Cl dosage	Manipulated Variable	ppm	0.1	-	-	30 sec	SCADA
Free Cl de-gritting	Constraint Variable	mg/lt	0.01	0.8	1.2	3 hours	Logbook
Free Cl filters	Constraint Variable	mg/lt	0.01	0.2	0.4	3 hours	Logbook

Raw water turbidity and temperature are categorized as independent variables, as their values are influenced by external environmental factors and remain unaffected by the pre-disinfection process. 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 concentrations at the de-gritting stages 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.

It should be noted that, although chlorine dosage is recorded at 30-second intervals, the dosage set point is adjusted approximately every 3 hours, so re-sampling has been done taking the median value for more accurate representation. For the remaining variables, a basic outlier removal procedure has been applied to ensure data quality and consistency. Additionally, a straightforward linear interpolation method has been employed to address missing values, provided the time gap does not exceed a maximum duration of 3 to 6 hours. The final dataset consists of 202 data points, with 161 allocated for RL training and 41 reserved for training the evaluation model.

2.3.3 Material and methods

As previously discussed, defining the MDP state in real-world RL problems presents a significant challenge. Table 2 presents the alternative MDP state configurations examined, differentiated with respect to the features used as inputs. In this setup, the state transition interval is set at 3 hours, while the agent's action for this use case is the adjustment of the chlorine dosage set point.

Table 2: Pre-disinfection MDP state versions

Version 1	Version 2	Version 3	Version 4	Version 5
Raw water Turbidity	Raw water Turbidity	Raw water Turbidity	Raw water Turbidity	Raw water Turbidity
Free Cl de-gritting	Free Cl de-gritting	Free Cl de-gritting	Free Cl de-gritting	Free Cl de-gritting
Free Cl filters	Free Cl filters	Free Cl filters	Free Cl filters	Free Cl filters
	Hour of the Day	Hour of the Day (0-23)	Hour of the Day (0-23)	Hour of the Day (0-23)
		Previous Cl dosage	Previous Cl dosage (- 3hours)	Previous Cl dosage (- 3hours)

The reward function is calculated by the following relation:

$$R = -\left(w_1 \cdot FreeCl_{R_{Degrit}} + w_2 \cdot FreeCl_{R_{Filters}} + w_3 \cdot \frac{Cl\ dosage_{used}}{Cl\ dosage_{max}}\right) \quad (2.1)$$

where:

$\boldsymbol{FreeCl_{R_{degrit}}}$: is derived from the extent to which the free chlorine concentration at the de-gritting stage exceeds defined upper or lower limits,

$\boldsymbol{FreeCl_{R_{filters}}}$: is similarly based on free chlorine concentration levels at the filter inlet.

$\boldsymbol{\frac{Cl\ dosage_{used}}{Cl\ dosage_{max}}}$**:** pertains to the effectiveness of the agent's actions, represented as the ratio of the chlorine dosage set by the agent to the maximum allowable dosage of 20 ppm.

Finally, the weights w_1, w_2, and w_3 are assigned to these terms, and they are treated as hyperparameters requiring tuning for optimal performance.

The evaluation of the RL Agent's performance is described by the following methodology:

1. **Objective**: To evaluate the RL Agent's performance, two distinct Random Forest regression models were developed:
    - The first model predicts the Free Chlorine concentration at the de-gritting stage at time *t+1*.
    - The second model forecasts the Free Chlorine concentration at the filter inlets at time *t+1*, where the time step is 3 hours.

2. **Independent Variables**: Both models utilize two primary features, identified through a feature selection process based on the Pearson correlation matrix, as illustrated in Figure 16:
    - Raw water turbidity at time *t*.
    - Chlorine dosage applied at time *t*.
3. **Evaluation Dataset and Model Training**:
    - The evaluation dataset was divided into a 50-50 split, with one half used for training and the other for evaluating the model performance (41 data points in total).
    - Model performance was quantified using the Mean Absolute Error (MAE), with the following results:
        - MAE for the Free Cl at de-gritting model: 0.13 mg/L.
        - MAE for the Free Cl at filter inlet model: 0.54 mg/L.
    - Although these MAE values indicate that the evaluation models may not exhibit optimal predictive accuracy, they serve primarily as a comparative benchmark.
4. **Comparative Nature of the Evaluation**:
    - It is essential to highlight that the evaluation does not assess the RL Agent's absolute real-world performance. Instead, it provides a relative performance comparison between the RL Agent's proposed actions and the existing operational policy, as documented in historical data from DWTP.
5. **Reward Calculation**:
    - The predicted Free Chlorine concentrations from these regression models are subsequently used to compute the reward function for the RL Agent (Eq. 2.1).
    - This framework enables the RL Agent's behavior to be compared effectively against established operator policies, facilitating the evaluation of potential performance improvements.

2.3.4 Results

The most effective MDP state configuration identified was Version 4 in Table 4. Under this onfiguration, theRL Agent successfully maintained all key constraint variables—namely, turbidity at the filter inlet (Figure 12)and free aluminum concentration at the clear water tank inlet (Figure 13)—within established safe operational thresholds on average. This outcome demonstrates that the chosen MDP state structure was well-suited to meet the regulatory and safety standards essential to the water treatment process.

Moreover, the RL Agent exhibited a behavior that closely emulated the policy typically followed by DWTP personnel (Figure 8). Through repeated iterations and learning, the agent's actions aligned progressively with the historical operational patterns, effectively "cloning" the behavior of human operators. Notably, this alignment was achieved with close to 7% reduction in chlorine dosage (Figure 8), suggesting that the RL Agent was able to identify slight optimizations without compromising on safety or operational standards. This marginal dosage reduction, although minor, could represent a valuable improvement in terms of chemical usage efficiency over time.

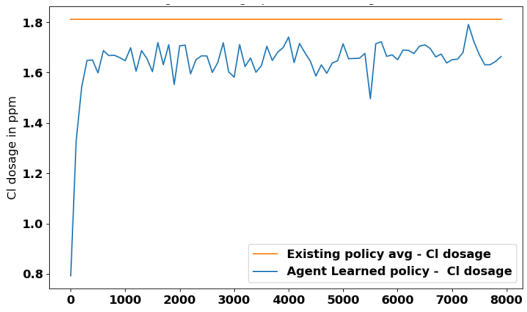

Figure 7: Average Cl dosage per 100 training iterations.

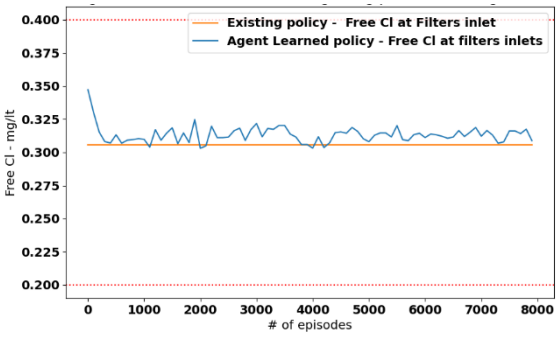

Figure 8: Average Free Cl concentration at filter's inlet per 100 training iterations.

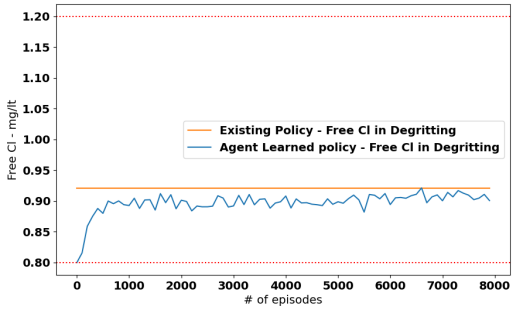

Figure 9: Average Free Cl concentration at De-gritting per 100 training iterations.

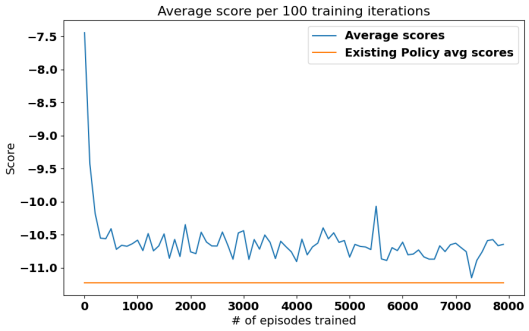

Figure 10: Average score (reward) per 100 training iterations.

2.3.5 Conclusions and next steps

The initial results of the beta version of the RL Agent are promising, particularly given the constraints of the limited historical data used for training. Despite these limitations, the RL Agent demonstrated a strong ability to replicate the operational policy typically employed by DWTP personnel. Furthermore, it achieved close to 5% improvement in performance, evidenced by almost 7% reduction in chlorine dosage, which points to the agent's potential for optimizing chemical usage without deviating from established safety and quality standards.

The next phase in development will involve expanding the dataset with historical records spanning a longer period, ideally over three years, drawn from both the SCADA system and the handwritten logbooks currently in use. Additionally, efforts will be directed toward incorporating supplementary parameters, such as raw water pH and conductivity, which are sporadically measured by EYDAP's Chemistry Laboratory. Integrating these variables into the MDP state could enhance the agent's capacity to respond to a broader range of operational conditions.

With an expanded dataset, it will be possible to refine the existing MDP state configurations and explore new versions of MDP states, thus supporting a more rigorous evaluation of the RL Agent's performance. A larger and more diverse dataset would also enable the development of more robust evaluation models, which are essential for accurately comparing the learned policy against the current policy followed by

DWTP operators. This comprehensive approach aims to facilitate a well-informed assessment of the RL Agent's performance and support any necessary adjustments for its deployment in real-world operations.

2.4 RL model #2: Optimization of coagulation stage

2.4.1 Problem statement

At the Polydendri DWTP, the coagulation and sedimentation processes, as shown in Figure 4, are conducted within two Palsuator tanks, each serving a distinct treatment line. The performance of this stage is primarily assessed using two critical parameters: turbidity at the filter inlet and free aluminum concentration downstream of the treatment process. Both parameters must remain within specified limits (Table 3) to ensure compliance with water quality standards.

Turbidity measurements are manually taken every three hours at the filter inlets using a handheld device. In contrast, the free aluminum concentration is continuously monitored by online sensors located at the inlet and outlet of the two clear water tanks, as shown in Figure 4. These measurements are vital for evaluating the treatment process's effectiveness and maintaining operational integrity.

Currently, coagulant dosage at the DWTP is adjusted reactively based on the observed values of these parameters. The goal of the RL implementation in this use case is to proactively determine an optimized coagulant dosage strategy. This approach aims to maintain turbidity and free aluminum concentration within safe operational limits while minimizing the total amount of coagulant used, thereby enhancing both efficiency and sustainability in the treatment process.

2.4.2 Data sources and data preprocessing

The primary data sources utilized for the RL model are drawn from the existing DWTP monitoring system. The DWTP's SCADA system continuously monitors the aluminum sulphate dosage and the free Alum concentration with a frequency of less than one minute, capturing data via online sensors. However, certain critical parameters are measured using handheld instruments at three-hour intervals and recorded manually in handwritten logbooks. These critical parameters are the raw water turbidity and the turbidity of water at the filter inlet.

For this study, data were sourced from SCADA records spanning November 2018, during which Yliki Lake was the primary water supply, i.e. the same time period as in Section 2.3 for the Optimization of pre-disinfection stage. Additionally, handwritten logbook entries from the same period were photographed, manually digitized, and the relevant parameters extracted for subsequent analysis.

Table 3 presents the retrieved parameters relevant to the coagulation process, along with associated contextual information. As mentioned in Section 2.3.2, critical parameters such as raw water pH and conductivity — essential for the coagulation process — are recorded once a day by EYDAP's Chemistry Laboratory and they were not utilized in the current RL beta version due to the low recording frequency.

Table 3: 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
Raw water Temperature	External Variable	Celsius	0.1	-	-	3 hours	Logbook
Aluminum Sulphate dosage	External Variable	ppm	1	-	-	30 sec	SCADA
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
Free Alum clear water Tank inlet	Constraint Variable	mg/lt	0.1	-	200	1 sec	SCADA

Raw water turbidity and temperature are categorized as external variables, as they are influenced by external environmental factors and remain unaffected by the coagulation process. Additionally, the free chlorine concentration at the de-gritting stage is also considered an external variable, given that the de-gritting stage precedes coagulation and is therefore not influenced by the coagulation procedure.

In contrast, aluminum sulphate dosage is classified as a manipulated variable, as it directly impacts the coagulation process and can be adjusted to control system performance. Meanwhile, water turbidity at the filter inlet and the concentration of free aluminum at the clear water tank inlet are identified as constraint variables, as these values must be maintained within defined safety limits to ensure safe and efficient operation of the treatment process.

It is important to note that, while alum dosage is recorded at 30-second intervals, its set point is adjusted approximately every 3 hours. Accordingly, the data were resampled, selecting the nearest recorded value to provide a more accurate representation. Likewise, free aluminum concentration measurements were down-sampled to a 3-hour interval using the nearest value. For the remaining variables, a basic outlier removal procedure has been applied to ensure data quality and consistency. Additionally, a straightforward linear interpolation method has been employed to address missing values, provided the time gap does not exceed a maximum duration of 3 to 6 hours. The resulting dataset contains 185 data points, with 147 used for training the RL model and 38 reserved for training the evaluation model.

2.4.3 Material and methods

As previously discussed, defining the MDP state in real-world RL problems presents a significant challenge. Table 4 outlines the parameters that constitute the various MDP state configurations used for training and evaluating the RL agent's performance. In this setup, the state transition interval is set at 3 hours, while the agent's action for this use case is the adjustment of the aluminum sulphate dosage set point.

Table 4: Coagulation MDP state versions

Version name	Version 1	Version 2	Version 3	Version 4
	Raw water Turbidity	Raw water Turbidity	Raw water Turbidity	Raw water Turbidity
	Turbidity filter inlet	Turbidity filter inlet	Turbidity filter inlet	Turbidity filter inlet
	Free Alum clear tank inlet	Free Alum clear tank inlet	Free Alum clear tank inlet	Free Alum clear tank inlet
		Free Cl de-gritting	Free Cl de-gritting	Free Cl de-gritting
			Previous raw water Turbidity (-3hours)	Previous raw water Turbidity (-3hours)
				Previous Free Alum clear water tank (-3hours)

The reward function is calculated by the following relation:

$$R = -\left(w_1 \cdot Turb_{R_{Filters}} + w_2 \cdot FreeAlum_{R_{Tank\ in}} + w_3 \cdot \frac{Alum\ dosage_{used}}{Alum\ dosage_{max}}\right) \quad (2.2)$$

The term $Turb_{R_{Filters}}$ is derived from the extent to which the water turbidity at the filter's inlet exceeds defined upper limit, while the $FreeAlum_{RTank\ in}$ term is similarly based on free Alum concentration levels at the clear water tank inlet. The third term pertains to the effectiveness of the agent's actions, represented as the ratio of the aluminum sulphate dosage set by the agent to the maximum allowable dosage of 70 ppm. The weights w_1, w_2, and w_3 are assigned to these terms, and they are treated as hyperparameters requiring tuning for optimal performance.

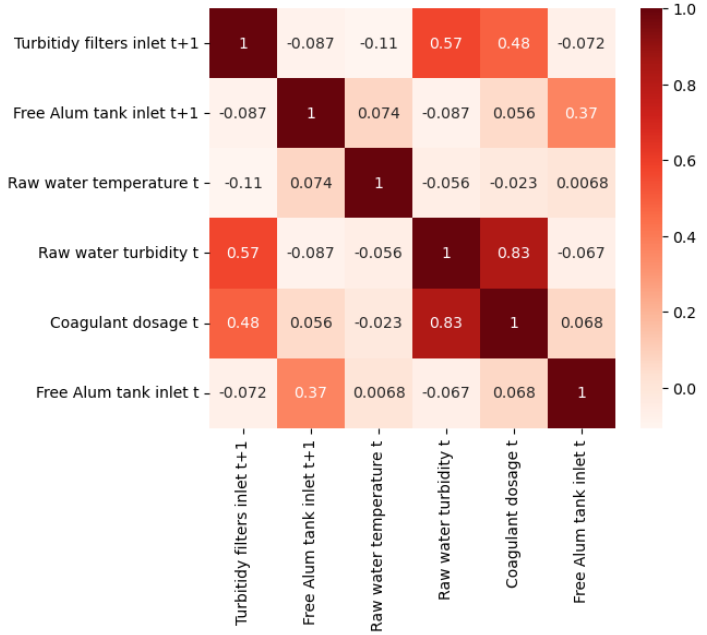

Figure 11: Pearson correlation matrix of coagulation monitored parameters

The evaluation of the RL Agent's performance is described by the following methodology:

1. **Objective**: To evaluate the RL Agent's performance, two distinct Random Forest regression models were developed:
    - The first model forecasts the water turbidity at the filter inlets at time *t+1*, where the time step is 3 hours.
    - The second model predicts the Free Alum concentration at the clear water tank inlet at time *t+1*, where the time step is 3 hours.
2. **Independent Variables for the Turbidity at filters model**: This model utilizes two primary features, identified through a feature selection process based on the Pearson correlation matrix, as illustrated in Figure 11:
    - Raw water turbidity at time *t*.
    - Aluminum Sulphate dosage applied at time *t*.
3. **Independent Variables for the Free alum model**: This model utilizes two primary features, determined after a feature selection process:
    - Free alum at clear water tank at time *t*.
    - Aluminum Sulphate dosage applied at time *t*.
4. **Evaluation Dataset and Model Training**:
    - The evaluation dataset was divided into a 50-50 split, with one half used for training and the other for evaluating the model performance (38 data points in total).
    - Model performance was quantified using the Mean Absolute Error (MAE), with the following results:
        - MAE for the Turbidity at filters model: 0.18 NTU.
        - MAE for the Free Alum at clear water tank inlet model: 16.5 mg/L.
    - Although these MAE values indicate that the evaluation models may not exhibit optimal predictive accuracy, they serve primarily as a comparative benchmark.
5. **Comparative Nature of the Evaluation**:
    - It is essential to highlight that the evaluation does not assess the RL Agent's absolute real-world performance. Instead, it provides a relative performance comparison between the RL Agent's proposed actions and the existing operational policy, as documented in historical data from the DWTP.
6. **Reward Calculation**:
    - The predicted Turbidity at the filter's inlet and the Free Alum concentration from these regression models are subsequently used to calculate the reward function for the RL agent, as defined in Eq. 2.2.

This framework enables the RL Agent's behavior to be compared effectively against established operator policies, facilitating the evaluation of potential performance improvements.

2.4.4 Results

The most effective MDP state configuration identified was Version 4 in Table 4. Under this configuration, the RL Agent successfully maintained all key constraint variables—namely, turbidity at the filter inlet (Figure 12)and free aluminum concentration at the clear water tank inlet (Figure 13)—within established safe operational thresholds on average. This outcome demonstrates that the chosen MDP state structure was well-suited to meet the regulatory and safety standards essential to the water treatment process.

Furthermore, the RL Agent exhibited behavior closely aligned with the operational policy traditionally followed by DWTP personnel. Through repeated learning cycles, the agent's actions progressively converged with historical operational patterns, effectively emulating the behavior of human operators (Figure 15). Notably, this alignment was achieved with a slight increase (2-3%) in aluminum sulphate dosage usage (Figure 14). This increase, however, may be attributable to the limited volume of historical data available for training, which may have constrained the agent's ability to optimize dosage efficiency fully.

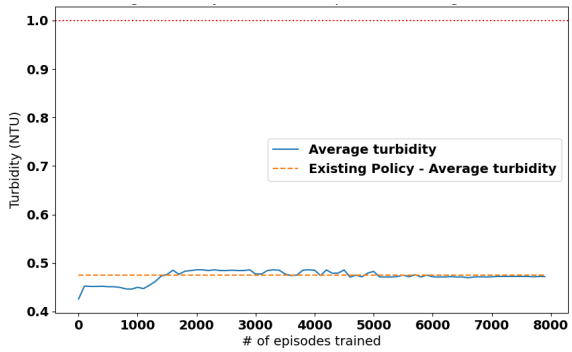

Figure 12: Average Turbidity at filters inlet per 100 training iterations.

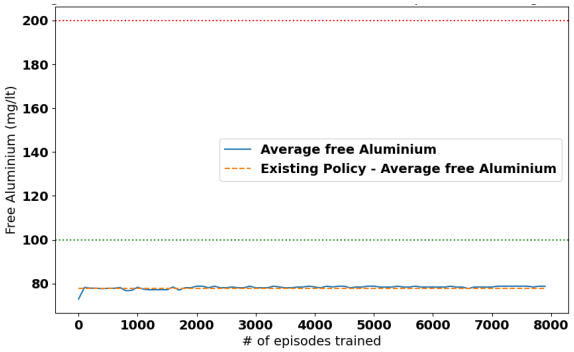

Figure 13: Average Free Aluminum Concentration at Tank inlet per 100 training iterations.

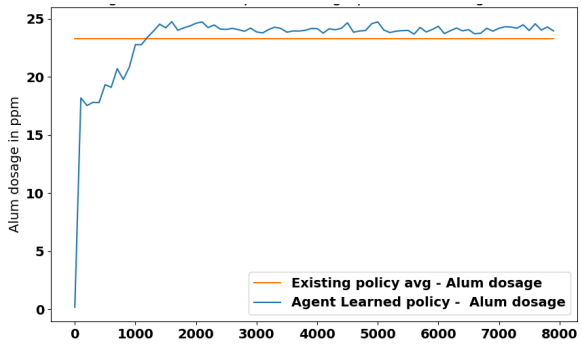

Figure 14: Average Aluminum Sulphate dosage per 100 training iterations.

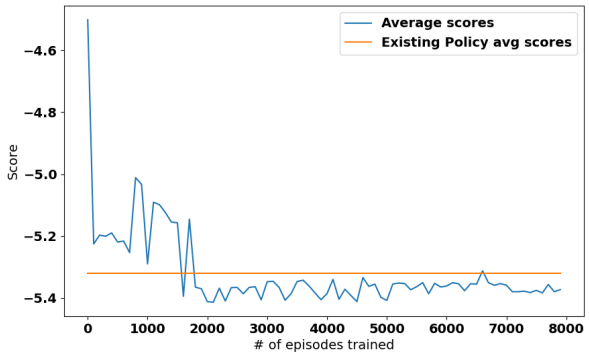

Figure 15: Average score(reward) per 100 training iterations.

2.4.5 Conclusions and next steps

The preliminary outcomes from the beta version of the RL Agent are encouraging, particularly given the constraints imposed by the limited historical data available for training. Despite this limitation, the RL Agent effectively emulated the operational policies traditionally followed by DWTP personnel, though with a slightly increased (2-3%) usage of aluminum sulphate. This may stem from the gradual nature of the coagulation process, combined with the limited training window of one month's historical data, which may have restricted the agent's capacity to achieve fully optimized dosing.

The next phase of development will focus on expanding the dataset with historical records covering a longer period, ideally spanning three years, sourced from both the SCADA system and the handwritten logbooks. This expanded dataset would provide a more comprehensive training basis, allowing the agent to capture seasonal and long-term variations. Additionally, efforts will be made to incorporate supplementary parameters, such as raw water pH and conductivity, which are periodically recorded by EYDAP's Chemistry Laboratory. Including these additional variables in the MDP state has the potential to enhance the agent's responsiveness to a broader range of operational conditions.

With an enriched dataset, it will become feasible to refine the existing MDP configurations and experiment with new versions, facilitating a more rigorous evaluation of the RL Agent's performance. The availability of a larger and more varied dataset will also enable the development of more robust evaluation models, essential for accurately benchmarking the RL Agent's learned policy against the current DWTP operational policy. This comprehensive approach is designed to provide a thorough assessment of the RL Agent's performance and inform any necessary modifications for successful deployment in real-world operations.

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

2.5.1 Problem statement

At the Polydendri DWTP, the post-disinfection process takes place after the filtration stage and before the clear water tanks, shown in Figure 4, using chlorine as the primary disinfectant. Alternatively, the DWTP has the capability to utilize sodium hypochlorite as a disinfectant. Maintaining adequate free chlorine levels downstream is essential to the effective operation of this process. The efficiency of the disinfection stage is primarily assessed by monitoring free chlorine concentration, continuously measured by online sensors positioned at both the inlet and outlet of the two clear water tanks. These measurements provide vital indicators of treatment performance, ensuring compliance with water quality standards.

Currently the DWTP operators set the disinfectant dosage reactively, ie by taking account only the Free Cl concentration downstream. The objective of the RL algorithm is to optimize the chlorine dosage, set in parts per million (ppm), to consistently maintain free chlorine levels within specified critical limits. This optimization ensures both operational efficiency and the safety of the treated water.

2.5.2 Data sources and data preprocessing

The primary data sources for the RL model are drawn from the DWTP's existing monitoring system. The DWTP's 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 this study, SCADA data from November 2018 were used, a period when Yliki Lake served as the primary water source and chlorine was used as disinfectant for that period. Table 5 presents the parameters relevant to post-disinfection, along with contextual information. As mentioned in the sections 2.3.2 and 2.4.2, critical parameters such as raw water pH and conductivity, which are essential for the chemical disinfection process, are measured once daily by EYDAP's Chemistry Laboratory. However, these parameters were not incorporated into the current RL beta version due to their low recording frequency.

Table 5: Post-disinfection recorded parameters

Parameter Name	Parameter Type	Units	Resolution	Low limit	High limit	Recording frequency	Source
Chlorine dosage	Manipulated Variable	ppm	0.1	-	-	30 sec	SCADA
Free Cl clear water Tank inlet	Constraint Variable	mg/lt	0.01	0.70	0.75	1 sec	SCADA

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.

It is noteworthy that, while chlorine dosage data is recorded every 30 seconds, its set point is typically adjusted at intervals of approximately one hour. To create a more representative dataset, this data was resampled down to a 1-hour time interval by selecting the nearest recorded value. Similarly, free chlorine concentration measurements were down-sampled to 30-minute intervals using the closest recorded values. For the remaining variables, a basic outlier removal procedure has been applied to ensure data quality and consistency. Additionally, a straightforward linear interpolation method has been employed to address missing values, provided the time gap does not exceed a maximum duration of 3 to 6 hours. The final dataset consists of 719 data points, with 574 allocated for training the RL model and 145 reserved for training the evaluation model.

2.5.3 Material and methods

As previously highlighted, defining the Markov Decision Process (MDP) state in real-world reinforcement learning (RL) applications poses a considerable challenge. Table 6 details the parameters that make up the different MDP state configurations used for training and assessing the RL agent's performance. In this framework, the state transition interval is set to one hour, and the agent's action involves adjusting the chlorine dosage set point to optimize system performance.

Table 6: Post-disinfection MDP state versions

Version 1	Version 2	Version 3
Free Cl Water Tank inlet	Free Cl Water Tank inlet	Free Cl Water Tank inlet
	Free Cl Water Tank inlet (-30 min)	Free Cl Water Tank inlet (-30 min)
		Previous Cl dosage set point (-hour)

The reward function is calculated by the following equation:

$$R = -(w_1 \cdot FreeCl\_R_{Tank\ in} + w_2 \cdot \frac{Cl\ dosage_{used}}{Cl\ dosage_{max}}) \quad (2.3)$$

The term $FreeCl_{R_{Tank\ in}}$ is derived from the extent to which the free chlorine concentration at clear water tank inlet exceeds defined upper or lower limits, while second term pertains to the effectiveness of the agent's actions, represented as the ratio of the chlorine dosage set by the agent to the maximum allowable dosage of 2 ppm. The weights w_1, w_2 are assigned to these terms, and they are treated as hyperparameters requiring tuning for optimal performance.

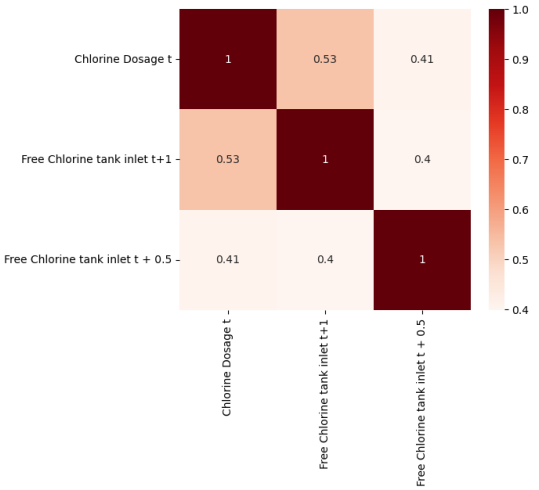

Figure 16: Pearson correlation matrix of post-disinfection monitored parameters

The evaluation of the RL Agent's performance is described by the following methodology:

1. **Objective**: To evaluate the RL Agent's performance, a Random Forest regression model was developed as AM:
    - This model predicts the Free Chlorine concentration at the clear water tank inlet at time *t+1*.
2. **Independent Variables**: The model utilizes two primary features, identified through a feature selection process based on the Pearson correlation matrix, as illustrated in Figure 16:
    - The Free Cl concentration at clear water tank inlet 30 minutes prior to the time t.
    - Chlorine dosage applied at time *t*.
3. **Evaluation Dataset and Model Training**:
    - The evaluation dataset was divided into a 50-50 split, with one half used for training and the other for evaluating the model performance (145 data points in total).
    - Model performance was quantified using the Mean Absolute Error (MAE), with the following results:
        - MAE for the Free Cl at clear water tank inlet model: 0.07 mg/L.
    - Although these MAE values indicate that the evaluation models may not exhibit optimal predictive accuracy, they serve primarily as a comparative benchmark.

4. **Comparative Nature of the Evaluation**:
    - It is essential to highlight that the evaluation does not assess the RL Agent's absolute real-world performance. Instead, it provides a relative performance comparison between the RL Agent's proposed actions and the existing operational policy, as documented in historical data from DWTP.
5. **Reward Calculation**:
    - The predicted Free Chlorine concentration from the regression model is subsequently used to calculate the reward function for the RL agent, as defined in Eq. 2.3.

This framework enables the RL Agent's behavior to be compared effectively against established operator policies, facilitating the evaluation of potential performance improvements.

2.5.4 Results

The most effective MDP state configuration identified in this use case was Version 3 in Table 6. Despite this, the RL agent did not achieve optimal performance as observed in Figure 19. Under Version 3, the RL agent was able to maintain the Free Cl concentration constraint within safe operational thresholds on average (Figure 17), but only for the initial 1,300 training iterations. After this period, a significant degradation in performance was observed, with the recommended dosage output persistently set to 0 ppm (Figure 18), an outcome well outside acceptable operational limits. This persistent dosing failure suggests that further optimization is required to enhance the RL agent's capacity for sustained, reliable decision-making.

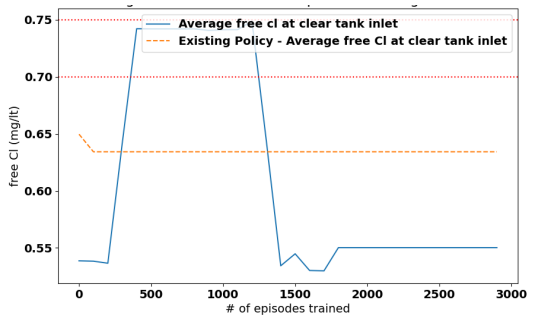

Figure 17: Average Free Cl concentration at clear water tank inlet per 100 training iterations.

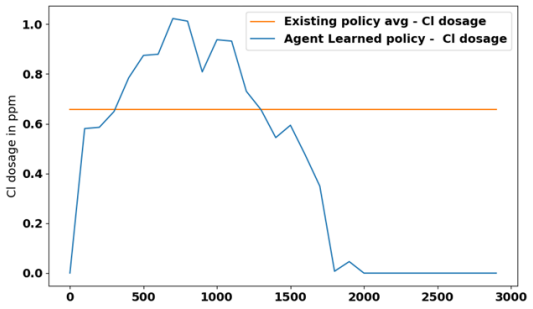

Figure 18: Average Cl dosage per 100 training.

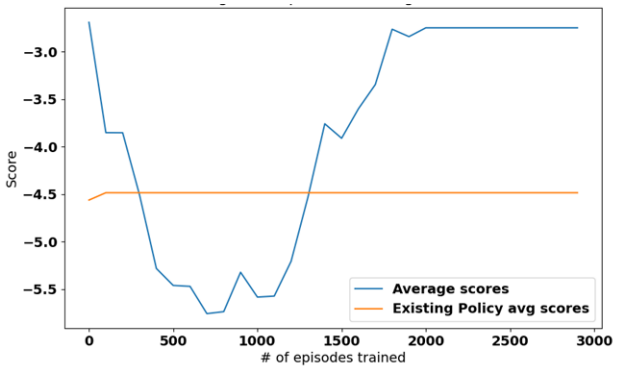

Figure 19: Average score (reward) per 100 training iterations.

2.5.5 Conclusions and next steps

The preliminary outcomes of the beta version of the RL Agent are currently suboptimal, primarily due to constraints imposed by limited historical data and a restricted number of recorded parameters available for training. Despite these limitations, the RL Agent was able to partially emulate the operational policies typically followed by DWTP personnel, though this emulation was sustained only over a limited number of training iterations. These results underscore the need for an expanded dataset to enable more robust and stable learning outcomes.

The next development phase will focus on extending the dataset to include historical records spanning an ideal period of three years, sourced not only from the SCADA system but also from handwritten logbooks, where Free Cl concentration is measured with handheld instruments of higher accuracy at the clear water tank inlet. This comprehensive dataset would allow the RL Agent to capture long-term and seasonal variations in operational conditions. Additionally, the dataset will be enriched by incorporating supplementary parameters, such as raw water pH and conductivity from periodic records maintained by EYDAP's Chemistry Laboratory, as well as relevant weather data—such as air temperature and cloud coverage—to broaden the RL Agent's contextual understanding. With this enriched dataset, it will be feasible to refine existing MDP configurations and experiment with new ones, fostering a rigorous evaluation process. Moreover, a larger, more varied dataset will support the development of robust evaluation models, essential for benchmarking the RL Agent's learned policy against current DWTP operational standards. This comprehensive approach is intended to facilitate a thorough assessment of the RL Agent’s performance and inform essential refinements, setting the stage for successful deployment in real-world operations.

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 drinking water treatment plants (DWTP), 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 20.

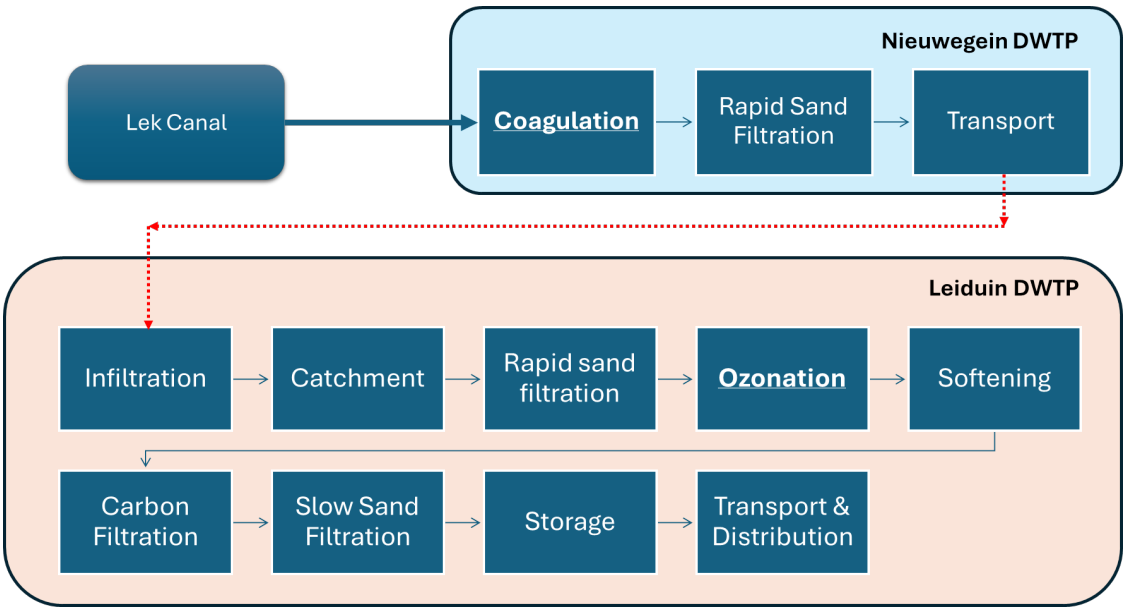

Figure 20: A schematic of the Leiduin drinking water treatment plant

3.2 RL model No4: Optimal $FeCl_3$ dosage for coagulation process

3.2.1 Problem statement

Among the various processes in a DWTP, coagulation–flocculation plays a pivotal role in destabilizing particles and facilitating their aggregation into larger flocs, which can be removed in subsequent filtration stages. An effective coagulation process removes harmful microorganisms, pathogens, and suspended solids. However, the success of this process depends heavily on the dosage of the coagulant, which in

WTNT's case is ferric chloride ($FeCl_3$). Insufficient dosage reduces the process's efficiency in lowering turbidity—an indicator of suspended solids and natural organic matter (NOM)—while overdosing can result in high iron concentrations entering the distribution system, increased chemical costs, higher energy consumption, and a greater environmental impact.

Traditionally, coagulant dosage is determined through empirical methods like jar testing, which require significant time and cannot respond to short-term water quality fluctuations. To address this, WTNT's process engineers currently adjust the $FeCl_3$ dosage every six hours (corresponding to the average retention time) based on turbidity measurements collected through their SCADA system from the outlet of the coagulation-flocculation lanes. This reactive approach relies on water quality fluctuations that have already occurred, adjusting dosage only after observing changes.

Data-driven models offer robust frameworks for water quality forecasting, for instance, by predicting random excels in turbidity, based on historical water quality data. ToDrinQ project, through WP4, aims to develop soft sensors to help water utilities prevent water quality deterioration and ensure high-quality drinking water. KWR proposes a supervised hybrid model (combining data-driven and physics-based methods) that predicts turbidity with a six-hour time-lag, giving WTNT's operational staff time to adjust the dosage if necessary. However, while this model provides an estimate of water quality at the coagulation-flocculation stage, it does not offer any guidance on optimal coagulant dosage. As a result, operators still rely on manuals and their experience for dosage decisions.

To fulfill this gap, the proposed reinforcement learning (RL) model aims to develop an optimized dosing strategy for $FeCl_3$. Reinforcement learning (RL) has shown promise in optimizing operational decisions in dynamic environments, enabling proactive and cost-effective interventions. In this case study, the RL model suggests the optimal $FeCl_3$ dosage for the next six-hour interval using as inputs the WQ and water flow data provided by the SCADA system and the outputs of the turbidity predictive model. By optimizing coagulant dosage, this model seeks to improve treatment efficiency and enhance the overall performance of the DWTP.

The proposed model builds upon well-established ML and RL frameworks, utilizing open-source tools, such as TensorFlow1 and PyTorch2, making replication feasible for utilities with similar SCADA infrastructure and same data availability. By incorporating domain knowledge at different DWTPs and leveraging SCADA data already available, the methodology ensures scalability across different plants. Site-specific customization is required, such as fine-tuning of the reward functions and recalibration of the predictive turbidity model, to simulate the conditions of each different DWTP. Nevertheless, this necessity is achievable with further collaborations between experts in the sector but also with operational staff whose knowledge of the local conditions at each specific DWTP is valuable.

3.2.2 Data sources and data preprocessing

The data used in this study consist of various water quality indicators and flow measurements taken from the intake and the inlet and outlet of the coagulation–flocculation process at the Nieuwegein DWTP. These parameters were recorded by sensors with a 5-minute frequency and stored in the plant's SCADA system. The specific parameters measured include:

- Temperature of the raw water (°C)
- Turbidity of raw water (FTU)

[1] https://www.tensorflow.org/api_docs
[2] https://pytorch.org/

- pH of the raw water
- Water flow in coagulation–flocculation lanes 1, 2, and 3 (m^3/h)
- Turbidity at the coagulation–flocculation inlet (FTU)
- $FeCl_3$ dosage in lanes 1, 2, and 3 (mg/L)

The dataset spans 3.5 years, covering the period from January 2020 to June 2024. Data pre-processing followed the methodology outlined in section 6.2 of Deliverable 4.1, as the RL model utilizes the same data as those employed in the development of soft sensor for predicting turbidity. In summary, the pre-processing involved removing and replacing outliers in temperature and flow measurements, followed by aggregating the 5-minute frequency data into 6-hour intervals.

3.2.3 Material and methods

Overview

This RL model aims to optimise the $FeCl_3$ dosage of the coagulation – flocculation process with the supervised model presented in the Deliverable D4.1 (soft sensor 1) acting as an agent. The supervised model provides turbidity predictions are essential feedback for the training of the RL. In the current form the RL model used is a deep q-learning network (DQN; Mnih et al., 2015), during the coming months of the implementation of this project additional approaches will be tested for improving the model performance.

Hybrid supervised model for turbidity prediction

The supervised model predicts the turbidity in the outlet of the coagulation – flocculation process with a 6-hour time-lag. The input data that model uses are pH, temperature and turbidity of the raw water, flow in each one of the three coagulation flocculation lanes and the lane number. The model is a hybrid approach. For more details check the relevant section in the Deliverable D4.1.

In this work, the already pre-trained model is used for the prediction of the future water quality conditions, as indicated by the turbidity measurements. Thus, this model allows the RL agent to anticipate the impact of different $FeCl_3$ dosages on the outlet turbidity.

Deep Q-learning Network (DQN) for coagulant optimization

DQN belongs in the family of the temporal difference (TD) category of RL algorithms. TD algorithms learn from interactions with the environment and update the state-action value estimation based on real-time estimates. In TD, the new experience is used to update the estimated reward, associated with a particular state at the next time step. DQN combines the use of deep neural networks with the q-learning methodology which aims to learn policies for a large variety of different simulations and approximates the future rewards for taking certain actions. The key steps of the DQN algorithm are described as follows:

1. **Experience:** The DQN collects different experiences based on the state, action, rewards and next states in a continuous approach.
2. **Target Network**: This is an additional network that is updated less frequently to capture the key state – action relationships
3. **Action selection and exploration**: An epsilon-greedy policy 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**: during each episode the q value (state and action) is updated using the reward obtained (predicted turbidity – threshold turbidity in our case) and the q value of the next state.

RL model architecture and flow-chart

The developed RL model is following these steps for the identification of the optimised dosing of $FeCl_3$:

1. **Step 1 – Setting the environment and generating the state:** The supervised model acts as the environment in our case, in other words, it simulates the conditions inside the coagulation – flocculation tank. The model predicts a turbidity in the coagulation flocculation outlet using the inputs (inlet pH, inlet temperature etc.). These inputs and outputs represent the current state in the RL model.
2. **Step 2 - Action:** The action is the selection of the the optimal dosage by the RL agent (DQN in our model) based on the state. The optimal dosage is selected from a set of actions that were given to the RL. In our case, the given actions were 8 different possible dosages (1.5, 2, 2.5, 3, 3.5, 4, 5 and 6 mg/l).
3. **Step 3 – Setting the reward:** Firstly, a threshold of a maximum turbidity of 4.5 FTU ($Turb\_thresh$) in the coagulation – flocculation outlet was set. The selected $FeCl_3$ dosage ($selected\ dosage$) is applied, and a new turbidity is predicted by the supervised model. The aim of this RL model is to achieve the 4.5 FTU threshold with the minimum possible coagulation dosage. Therefore, the reward is calculated as the sum of turbidity award and the dosage reward if the turbidity threshold is achieved and equal to -3 otherwise. So, the total reward and the turbidity and dosage rewards are calculated as follows:

   If predicted turbidity >4.5

   $$Total\ Reward = -3 \quad (3.1)$$

   If predicted turbidity <4.5

   $$Turb\ reward = \frac{Turb_{thresh} - predicted\ turbidity}{Turb\_thresh} \quad (3.2)$$

   $$Dosage\ Reward = \frac{Max\ dosage - selected\ dosage}{Max\ dosage} \quad (3.3)$$

   $$Total\ Reward = Turb\ reward + Dosage\ Reward \quad (3.4)$$

   where, $Max\ dosage$ is the maximum dosage from the set of actions that in our case is 6 mg/l.
4. **Step 4 – Updating policy:** DQN takes the reward of the step 3 to update the policy of dosage selection in the next step.

A schematic of the RL flowchart is presented in Figure 21.

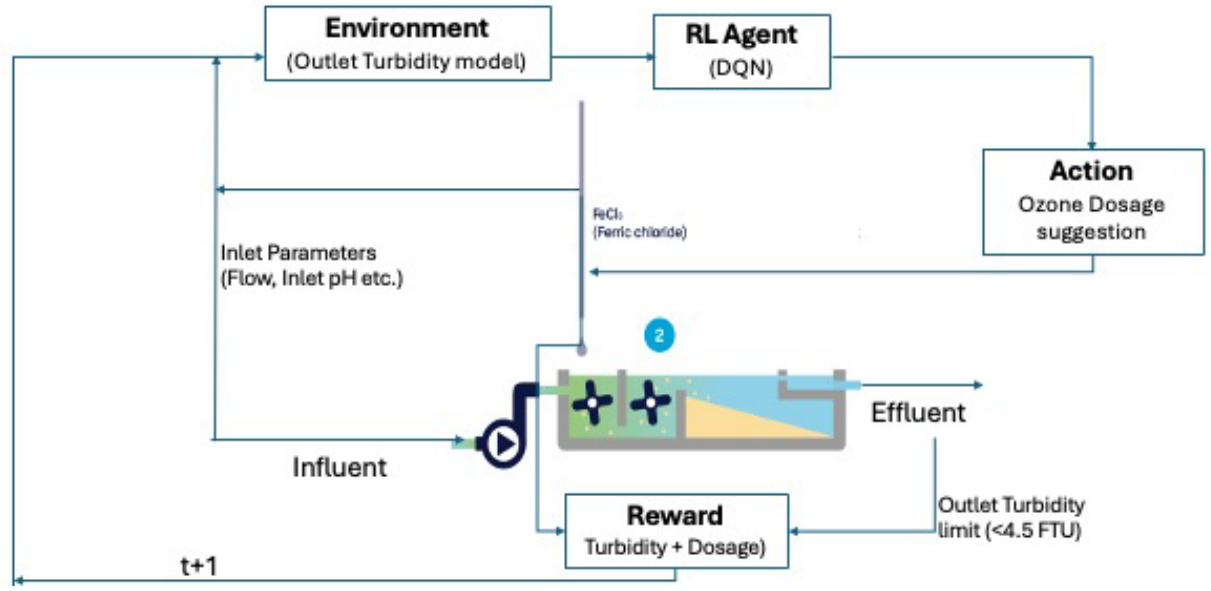

Figure 21: Flow chart of the RL model

3.2.4 Results

The RL model was trained for 100 episodes and required around 12 hours for its training. The range of the suggested dosages for each step are presented in the following Figure 22. It is clear that the range of actions that the RL used was between 2 to 4 mg/l and it completely ignored the 1.5 mg/l the actions above 4mg/l.

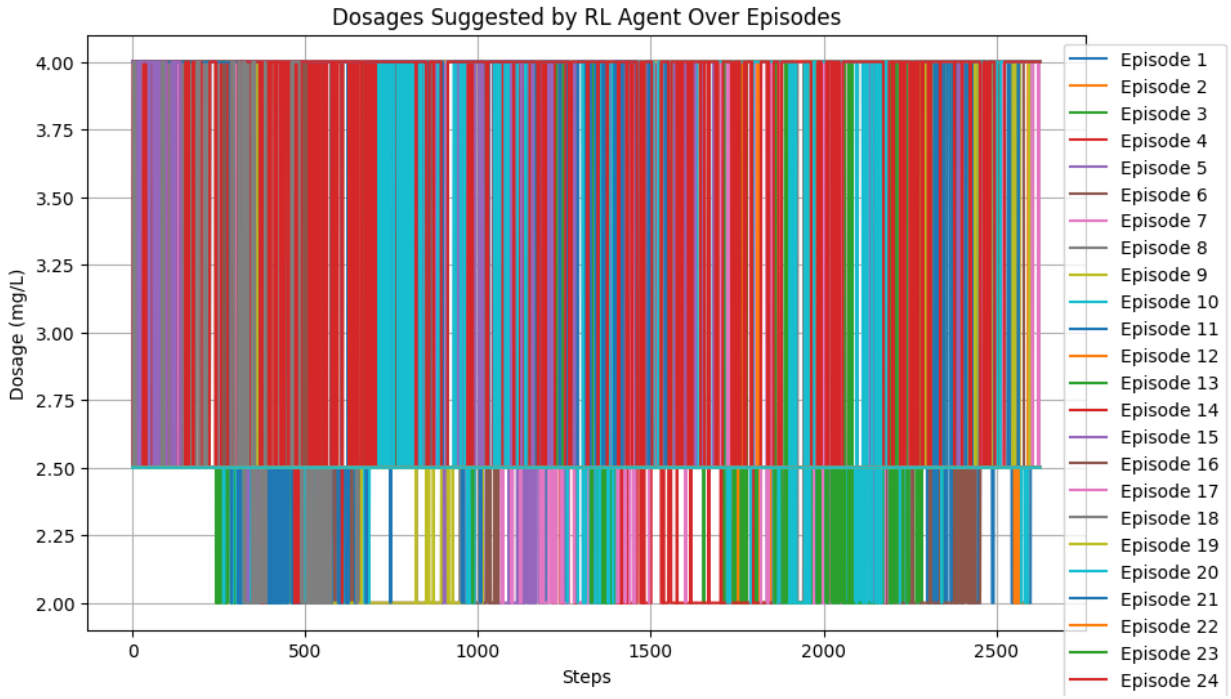

Figure 22: RL suggested dosages for each step per episode

In Figure 23, the recommended dosages from the best episode of the training period are compared to the actual coagulant dosages. The RL model generally suggests higher dosages than those measured; however, as shown by the peaks in the plot, the maximum RL-suggested dosage is lower than the actual maximum dosage. This outcome suggests that the 4.5 NTU threshold is conservatively set, ensuring reliable turbidity control.

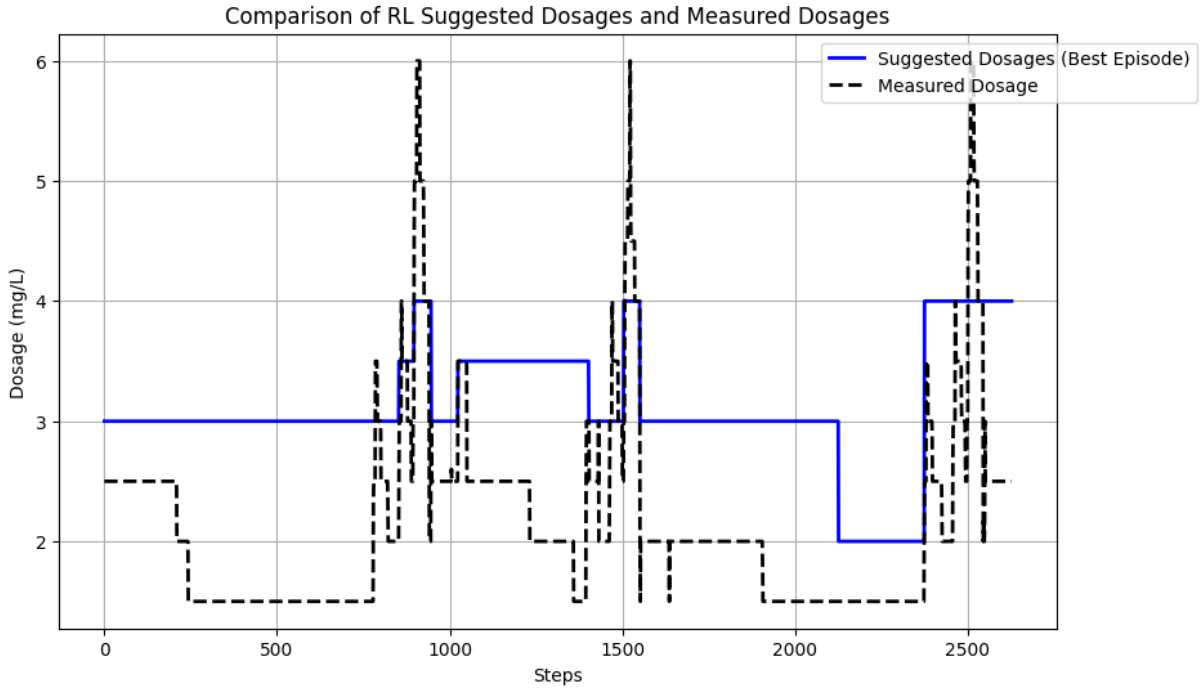

Figure 23: Best RL episode suggested coagulant dosages vs measured coagulant dosages

The subsequent plot in Figure 24 compares the predicted turbidity using the RL-suggested dosages with the actual turbidity. This comparison further supports the conclusion, as the RL dosages are expected to result in lower average turbidity in the coagulation-flocculation outlet compared to the actual turbidity.

To refine these thresholds, further discussions with WTNT's process engineers are necessary, especially to consider additional water quality parameters for setting the threshold. It's also important to note that the accuracy of the RL model's recommendations depends on the predictive model's accuracy. Further development and fine tuning of this model, in the third year of the ToDrinQ project, will enable retraining the RL model to enhance its optimization capabilities.

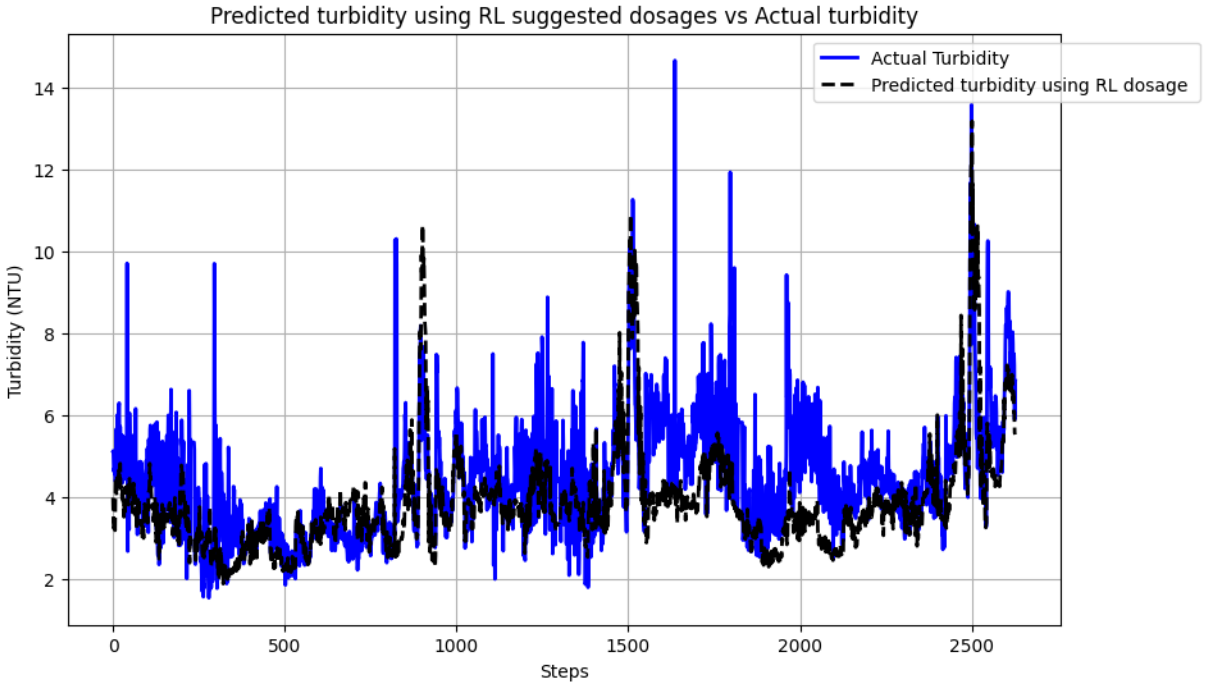

Figure 24: Actual turbidity outputs vs predicted turbidity outputs using the RL suggested dosages

3.2.5 Conclusions and next steps

An RL model is generated with the aim to optimise the $FeCl_3$ dosage coagulation-flocculation process. The RL model is using the under-development turbidity predictive model as an environment with the aim to suggest a minimum dosage that will reduce the outlet turbidity to below 4.5 NTU. The initial results indicated that this approach is a potential tool that could be used from WTNT's process engineers, however, it requires high computational power for its initial training. The next steps for the development of this RL are as follows:

- Re-run the RL model for more episodes and with a new dataset with the intention to diversify the data inputs, cover various turbidity conditions and testing a more active response from the model.
- Use further fine-tuning optimization to reduce the complexity of the model and, consequently, reduce its computational demand.
- Re-run the RL model using the improved version of the predictive model that will be fine-tuned during the 3rd year of the project. This process also includes the use as an environment of the three versions of the predictive model (physics model, deep learning model, physics and deep learning model).
- Try different reward scenarios (higher turbidity threshold, different dosage reward) and potentially introducing additional factors such as pumping dosage energy consumption or general costing of the process.
- Replace the DQN RL algorithm with a large language model (LLM) that will learn the patterns in the dataset but also will aim to learn the way that the predictive model makes its turbidity predictions.
- The final test for this RL will be implemented in the experimental setup of the Leiduin DWTP which is the goal in the 4th year of the project.

4. Upscaling and European Added Value (EAV)

Water treatment plants (WTPs) are critical components of the water supply chain, omnipresent in every urban water system in Europe and beyond, irrespective of their complexity and size. The ToDrinQ project attempts to advance the state of play in the management and optimization of WTPs, by leveraging approaches and techniques from the realm of Reinforcement Learning (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 WTP. 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.

Despite the fact that the RL models have been developed around the needs of DC#1 and DC#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. At the same time, they are modular and hence they can be deployed individually depending on the treatment process exist in a treatment unit. The successful deployment of RL models in two demanding demonstration cases (serving two large capitals, i.e., Amsterdam and Athens) provides compelling evidence of their effectiveness. 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 to mention 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.

5. 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