Safe Reinforcement Learning Control for Water Distribution Networks

Ledesma, Jorge Val

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

SAFE REINFORCEMENT LEARNING CONTROL FOR WATER DISTRIBUTION NETWORKS

BY
JORGE VAL LEDESMA

DISSERTATION SUBMITTED 2022

AALBORG UNIVERSITY
DENMARK

Safe Reinforcement Learning Control for Water Distribution Networks

Ph.D. Dissertation
Jorge Val Ledesma

Dissertation submitted May, 2022

Dissertation submitted: May, 2022

PhD supervisor: Professor Rafał Wisniewski
Aalborg University

Assistant PhD supervisor: Professor Carsten Skovmose Kallesøe
Grundfos Holding A/S / Aalborg University

PhD committee: Professor Jan Østergaard (chairman)
Aalborg University, Denmark
Associate Professor Carlos Ocampo-Martinez
Universitat Politècnica de Catalunya, Barcelona, Spain
Professor Sébastien Gros
Norwegian University of Science and Technology, Norway

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-894-6

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Stibo Complete, 2022

Abstract

This thesis is concerned with the problem of designing a safe model-free control solution that provides optimal pressure management to Water Distribution Networks (WDNs) with elevated reservoirs.
Optimal pressure management can mitigate pipe bursts. Thus, reducing the amount of lost water and the implicit repair cost. The deployment of optimal controllers in industry shows promising results. However, their implementation requires a system model that includes network dynamics and consumption demands. Building and maintaining such a model requires qualified personnel. This factor significantly increases the costs, and impedes the proliferation of advanced control solutions in small utilities.
This project's objective is to improve the accessibility of optimal control solutions to water utilities, especially to small-medium-size utilities with a limited budget. A Reinforcement Learning (RL) controller is proposed to gradually adapt to the deployed system for optimising the pressure without knowing the system dynamics, only by observing the measured data. Additionally, other management objectives are considered in the controller design, such as smoothness of the pump actuation or quality of the supplied water. The learned policy consists of a linear controller that is derived from the Q-value function. The structure of the Q-value function is built with a polynomial approximation that is suitable for a nominal linear system and a quadratic cost function. This formulation aims to simplify the learning of such function and to increase the interpretability of the control solution. The performance of data-driven methods relies on the quality of the collected data and approximation structure. This project proposes two solutions to improve the learning robustness when identification is performed under poor experimental conditions.
WDNs are critical infrastructures, and their operation is essential for society. Therefore, the management must provide a robust and continuous supply. This project develops a solution for safe exploration. The RL controller searches for the optimal solution within a predefined safe area, and if a predicted system trajectory violates the safety limits, a policy supervisor overrules the control input.

A modular laboratory setup is built that emulates, on a small scale, the behaviour of different water infrastructures. The developed solutions are validated in a testbed that reproduces a WDN with an elevated reservoir. The laboratory results emphasise the strengths and limitations of the designed methods, thus showing the importance of a safe exploration when implementing learning controllers.

Resumé

Denne afhandling beskæftiger sig med design af sikker modelfri kontrolløsninger, til optimal trykstyring af Vanddistribution Netværk (WDN'er) med højdebeholdere.
Optimal trykstyring kan afbøde rørbrud og således reducere mængden af tabt vand og de medfølgende reparationsomkostninger. Udbredelsen af optimale controllere i industrien viser lovende resultater, men implementeringen kræver systemmodeler, hvilket i vandforsyningen inkluderer netværksdynamik og forbrugskrav. Opbygning og vedligeholdelse af sådanne modeller kræver kvalificeret personale, med øgede omkostningerne til følge, hvilket vanskeliggør spredningen af avancerede kontrolløsninger i små forsyningsselskaber.
Dette projekts mål er at lette adgangen til kontrolløsninger for forsyningerne, især for de små- og mellemstore forsyninger med et reduceret budget. En *Reinforcement Learning* (RL) controller foreslås. Denne RL-controller tilpasser sig gradvist det system den er installeret i ved at observere de målte data. Herved optimeres trykket uden at kende systemdynamikken. Derudover tages andre driftsmål i betragtning i controllerdesignet, f.eks. jævnheden af pumpeaktiveringen og kvaliteten af det tilførte vand. Den lærte styringsstrategi består af en lineær controller, der er afledt af de såkaldte *Q-value* funktionen. Strukturen af *Q-value* funktionen er polynomisk. Denne struktur egner sig godt til et nominelt lineært system og en kvadratisk kostfunktion. Denne formulering har til formål at forenkle indlæringen og at tolkbarheden af kontrolløsningen. Præstationen af datadrevne metoder afhænger af kvaliteten af de indsamlede data og den underliggende modelstruktur, her polynomisk. Dette projekt præsenterer to løsninger til at forbedre robusthed under læring, når identifikation udføres under dårlig eksperimentelle betingelser.
WDN'er er kritiske infrastrukturer, og deres drift er afgørende for samfundet. Derfor skal ledelsen i vandforsyningerne sørge for en robust og kontinuerlig drift. Dette projekt udvikler en løsning til sikker modelfri kontrol også under træning. RL-controlleren søger efter den optimale løsning inden for et foruddefineret sikkert område, og hvis en prædikteret systembane overtræder sikkerhedsgrænserne, overtager en beskyttelseskontroller og kor-

rigere kontrolinputtet så sikkerhedsgrænserne overholdes.
Som en del af projektet er en modulær laboratorieopsætning blevet designet, som i lille skala kan emulere adfærden af forskellige vandinfrastrukturer. De udviklede løsninger er valideret i det designede laboratorie. Laboratorieresultater understreger styrkerne og begrænsningerne ved læringsbaserede designmetoder, hvilket viser vigtigheden af sikre læringsstrategier ved implementering af læringscontrollere som RL.

Contents

Preface

This thesis is submitted as a collection of papers in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the *Department of Electronic Systems, Automation and Control, Aalborg University, Denmark*. The work covered by this thesis has been carried out in the period from November 2018 to April 2022 under the Smart Water Infrastructures Project funded by Poul Due Jensens Fond.

The thesis is structured in two parts, the first part gives an introduction to the presented work together with a summary of the contributions, and the second part consists of six published or submitted papers.

I would like to thank my supervisors, Prof. Rafał Wisniewski and Prof. Carsten S. Kallesøe, firstly for trusting me with the development of the *Smart Water Infrastructures Laboratory* and secondly for their support throughout this research project. They have been great mentors in my career and they have guided me, with optimism, to overcome the challenges of this project. I also would like to thank Agisilaos Tsouvalas and his team for welcoming me at Grundfos during *lock-down* times when establishing external collaboration was exceptionally difficult. This collaboration provided suggestions for contextualising a research problem into real needs. I would like to thank my colleagues in the section of Control & Automation for the inspiring discussions and experiences. Additionally, I would like to thank my family for encouraging me to be curious in life, and my friends Giannis, Rubén and Diego for hosting me in Aalborg and making me feel at home after long working hours. Finally, I would like to especially thank Andrea for her relentless support and patience during all the stages of this project, becoming a trusted reviewer and ally to fighting deadlines.

Jorge Val Ledesma
Aalborg University, May 24, 2022

Part I

Introduction & Summary

Introduction

This chapter presents the background and motivation for investigating adaptive-optimal management of water distribution networks. In Section 1, a brief description of the water distribution network is given, and its operational objectives are presented in Section 2, a general outlook of the current management solutions is provided in Section 3, the project objectives are listed in Section 4 and a brief presentation of how this project contributes to the extension of the state-of-the-art is given in Section 5.

1 Motivation

This research project was initiated by Aalborg University as part of the Smart Water Infrastructures Laboratory (SWIL) project. The SWIL project is funded by Poul Due Jensens Foundation, and its objective is to create a research facility that supports the discovery of new solutions for the management of water infrastructures. The motivation for this project is built upon two main aspects, environmental and economic.

1.1 Environmental aspects

Water is a limited resource that is essential for life. In 2018, the report from the United Nations's (UN), World Water Development Report (WWDR) [1], presents an update on the current challenges regarding clean water availability and future trends.
The increasing water demand due to a growing human population and higher living standards is causing water stress. Hence, putting at risk a safe supply of drinking water. The water stress is defined as "*the ratio of total annual water withdrawals to total available annual renewable supply*"[2]. This ratio depends on two variables, supply availability and the demand for that water. In Figure 1 the measures of water stress are visualized across the world. The report presented in [1] relates the increasing water demand patterns with the population and economic growth. With the current consumption trend, significant water demand is forecasted for the next decades in three sectors: industry,

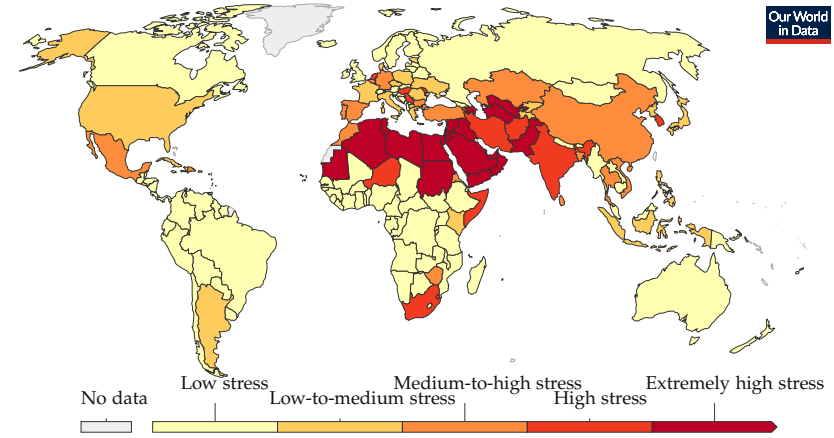
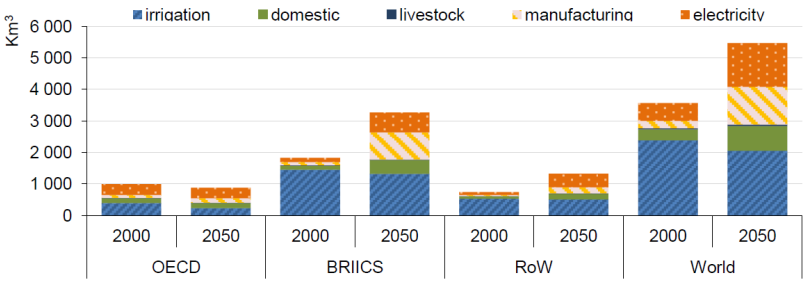


Figure 1: World map of the water stress measures in 2017. [3, 4]

domestic and agriculture. Under this baseline scenario, the water demand will exceed the resources available in many world regions.
The OECD presented a set of policies with the purpose of changing the 2050 outlook and therefore mitigating future water crises [5]. In order to make these policies effective, the proposed solution must be, apart from necessary, affordable. Therefore, an essential task is to integrate *green growth* into the global economic policies. In addition to this, another important task is to improve societal understanding and self-awareness of the upcoming challenge.



Source: OECD Environmental Outlook Baseline; output from IMAGE suite of models

Figure 2: Global water demand projection for 2050 [5]. Note: BRIICS: Brazil, Russia, India, Indonesia, China, South Africa. RoW: Rest of the world.

1.2 Economic aspects

As previously introduced, the environmental challenges cannot be addressed without considering the economic impact of the actions. Figure 2 shows that domestic water represents a small share of total water usage globally. Nevertheless, drinking water usage can still be improved by optimising its distribution. This subsection presents three operational challenges of the water distribution systems that significantly impact the utility economy. These issues are listed in [6], and they are the following:

Issue 1: *Non-Revenue Water*

Non-Revenue Water (NRW) is defined as *"the difference between the volume of water that is put into a water distribution system and the volume that is billed to the customers"* [7]. There are three main causes for having NRW: *Physical loss* due to leakages at some parts of the network or reservoir overflow, *commercial loss* due to customer meter errors or unauthorised consumption (water theft) and *not-billed authorised* consumption due to operational purposes or firefighting.

The study presented in [8] shows that the costs caused by NRW are estimated at USD 14 billion per year. The percentage of NRW caused by leakages varies from 95% to 50% depending on the case study [9].

Issue 2: *Energy consumption*

The urban water cycle consumes a considerable amount of energy; this comprises the different stages of the water supply and wastewater collection. The Environmental Protection Agency shows that water infrastructures consume around 2% of the whole nation's energy consumption [10], this is translated into an annual cost of USD 4.7 billion [11]. The pumping systems are the major energy consumers in this water cycle.

Issue 3: *Operation and maintenance*

Some major operational issues are pipe bursts and the associated costs. Besides the lost water, the repairs are particularly expensive due to their difficult accessibility, placed at a certain depth within an urban district. The repair frequency of the pipes is, to a great extent, related to the pipe burst [12].

European water operators are already committed to achieving the Sustainable Development Goal (SDG) 6 *"Ensure availability and sustainable management of water and sanitation for all"*[13]. This statement implies that the management of these utilities must steer their policies to solve the aforementioned distribution challenges.

2 Brief introduction to water distribution networks

The basis of the operation of a water distribution systems is to transport potable water from a source to multiple consumers. The water infrastructure varies in complexity, from rural areas to big cities. However, generally, most the water distribution system comprises four main components [14]: Water sources, treatment works, transmission mains and distribution network.
The sources are divided into surfaces like rivers or lakes and ground like boreholes or wells. Then, the intake facility extracts the water and delivers it to the plant for its treatment and storage. Finally, the pumping station regulates water inflow to the distribution network. The network is typically divided into main transmission pipelines that transport the water to an urban district, a distribution network that supplies within an urban area and service pipes that connect the network with consumers. Figure 3 illustrates part of the water cycle from the water source to the consumers.

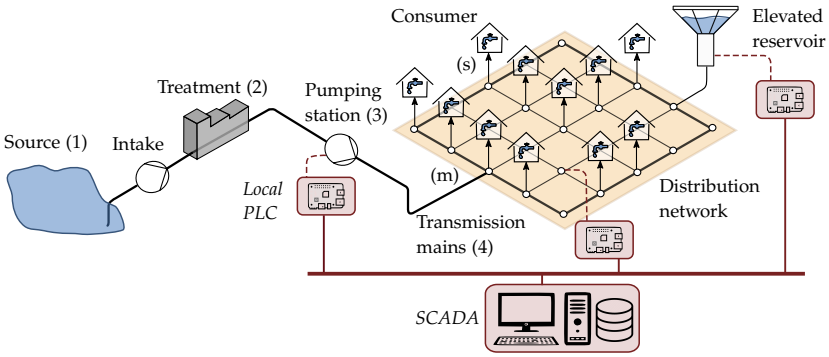


Figure 3: Scheme of a simplified water supply system where (1) represents the water source, (2) a treatment plant, (3) a pumping station and (4) the transmission system with mains (m), distribution network and service pipe connections (s). The communication network that links the local PLCs with the SCADA is depicted in garnet-red

Smart Water Infrastructures Laboratory (SWIL)

The modernisation of water distribution networks entails the installation of new technology that allows the implementation of advanced control solutions. Water infrastructures are critical infrastructures that demand a robust and continuous operation. Therefore, the future development of management solutions and the implementation of new technology in large scale infrastructures require extensive validation before deploying. However, the validation against certain management scenarios implies a high cost and risk for the system operation or the environment. Examples of these scenarios are water leakages, water contamination, wastewater overflow, or interruption of

the infrastructure service. A test centre that recreates these scenarios, along with the proposed management solution, facilitates safe validation of new technology.
The SWIL can accommodate experiments in district heating, water distribution networks and wastewater collection. The physical behaviour of these water systems is qualitatively emulated, and the real-time monitoring and control systems are replicated [Paper E]. The laboratory is designed with a modular architecture for increasing the test-beds versatility, allowing to construct a wide variety of network topologies with a reduced number of modules. Moreover, the modular architecture is accordingly applied to the Data Acquisition (DAQ) system. A picture of the laboratory is depicted in Figure 4 and the communication network structure is illustrated on the modules and the SCADA PC. A brief description of the laboratory modules compared with the real components is provided in Section 2.1, and a detailed description of the facility is given in Paper E.

2.1 Distribution network components

This project studies only a section of the water supply management, particularly the transportation of water from the pumping station to the consumers with pressurized pumping systems. In this section, a network is mainly characterized by four components and its interconnection. A brief description of each component is given below followed by a description of the laboratory emulation:

Pumping station: This component is in charge of the water inflow to the pipe network, there may be one or multiple inflow nodes, the pumping station must regulate the network pressure/flow such that the demands at the end-users are meet. See examples of a real pumping station and its SWIL analogous in Figure 5 and Figure 6 respectively. Variable speed pumps and sensors (pressure and flow) are used to emulate a pumping station in the SWIL.

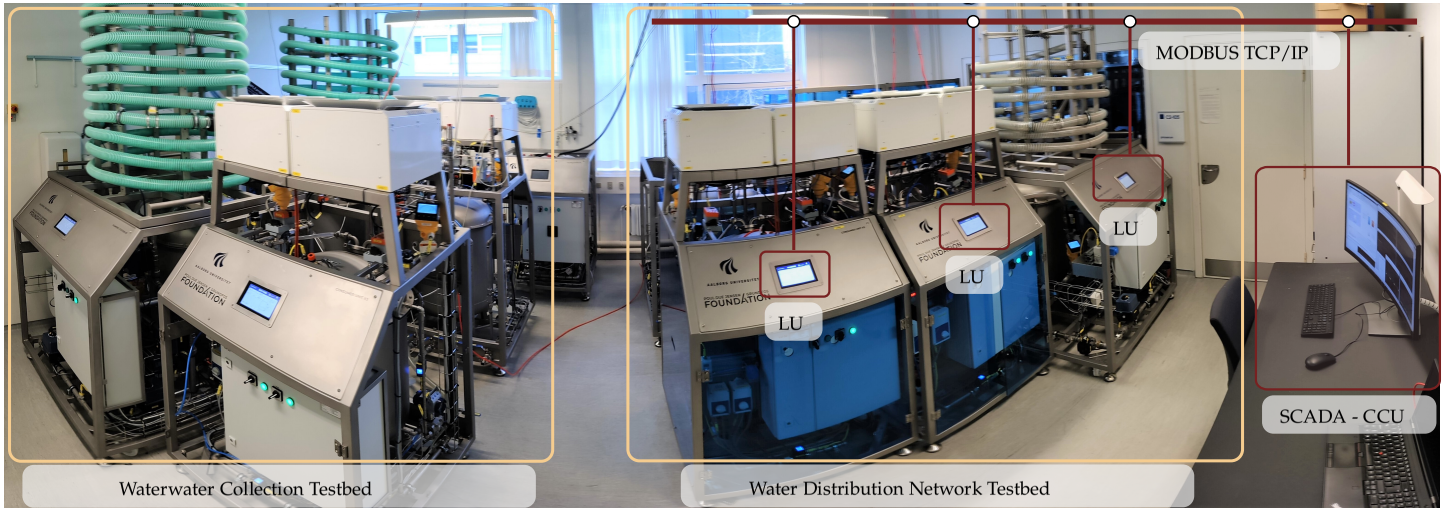


Figure 4: Picture of the SWIL with two test-beds and the SCADA-PC at Aalborg University [Paper E]: (**Left**) Wastewater collection. (**Center**) Water distribution network. (**Right**) SCADA-PC. A simplified communication architecture that connects the local PLCs - Local Units (LUs) and the SCADA PC - Central Control Unit (CCU) is represented in garnet-red.

Figure 5: Example of a pumping station with several pumps in parallel.

Figure 6: Example of a pumping station module with three pumps in parallel at the SWIL.

Pipe network: The hydraulic network distributes the water from the pumping inflow to the end-users. The structure of the networks varies depending on the location of the district since there is a qualitative difference between the topology in urban and rural areas or hill and flatland areas. A brief description of the network types is given in Section 2.2.
Pressure Reducing Valves (PRVs) are installed in the pipe network with the purpose of strategically reducing the pressure in some areas. In this project, PRVs are considered passive elements. Consequently, they are not part of the network management. See examples of a real pipe network and its SWIL analogous in Figure 7 and Figure 8 respectively. Pipelines of different lengths and diameters are folded inside a module and used to interconnect the components in the SWIL, thus emulating the physical effects of real distribution networks.

Figure 7: Example of a pipe network in an urban district. Picture source:[15]

Figure 8: Example of a pipe module at the SWIL with inlet valves, folded pipelines and sensors.

Consumers: In a city district, the end-users consume water from the pipe network. The water at the consumption nodes has to fulfil the quality stan-

dards, and the pressure must be sufficient for its consumption. See examples of a real consumer and its SWIL analogous in Figure 9 and Figure 10 respectively. In the SWIL, the end-users consumption is emulated by variating the opening degree of a controllable valve, reproducing the consumption pattern of a city district.

Figure 9: Example of a consumer where the user consumes drinking water from a tap. Picture source:[16]

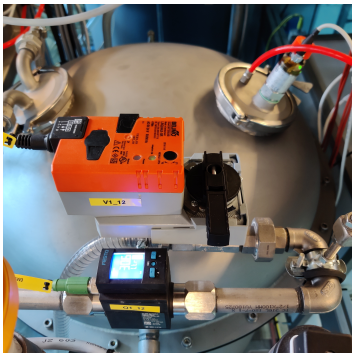


Figure 10: Example of a consumer at the SWIL with a controllable valve and a flow sensor.

Elevated reservoir: This element has three main purposes in the network: Supporting the water supply during the peak of the demand, maintaining constant pressure in the network, and guaranteeing the supply during emergencies. It is not required for all distribution networks to have an elevated reservoir. Nevertheless, the reservoir can provide alternative management strategies for reducing the pumping cost by exploiting its storage capacity. See examples of a real elevated reservoir and its laboratory equivalent in Figure 11 and Figure 12 respectively. In the SWIL, the tank stores water similarly to a real elevated reservoir, and the physical elevation is emulated by regulating the air pressure inside the tank.

Figure 11: Example of an elevated reservoir in Huesca (Spain).

Figure 12: Example of an elevated reservoir at the SWIL.

This project's scope focuses on small-medium sized pressurized networks with the following network structure: a ring or branch pipe network topology with a single pumping station and an elevated reservoir.

2.2 Monitoring and operation

The implementation of the advances in Information and Communication Technology (ICT), sensors, actuators and smart meters opens the possibility of potential improvements in the system operation. For instance, by improving the monitoring in the network, the management can be adapted for each particular scenario in real-time, such as meeting consumer demands, delivering sufficient water quality, or detecting network failures [17].
The operation of the network can be controlled in real-time via a Supervisory Control And Data Acquisition (SCADA) system, where the data is collected from different points in the network (via local PLCs or smart meters). In the SCADA system, the collected data is used to provide a control strategy that meets the utility management objectives. Figure 3 illustrates in garnet-red the communication architecture, where a simplified version of a SCADA system collects information from different points in the network and controls the operation.

One of the utility practices for operating an urban water distribution system is via *pressure management*, Pressure Reducing Valves (PRV) and Variable Speed Pumps (VSPs) are the two actuators used for this kind of operation [18]. Several studies relate the network pressure with the frequency of pipe burst [19, 20]. These show that distribution systems benefit in both environmental and economic aspects, by operating continuously at moderate pressure.

Definition 1 (Pressure Management)
The Water Loss Specialist Group from the International Water Association (IWA) defines pressure management as [21]:
"The practice of managing system pressure to optimum levels of service ensuring sufficient and efficient supply to legitimate uses and consumers, while

- *reducing unnecessary or excess pressure*
- *eliminating transients and faulty level controls*
- *reducing the impact of theft*

all of which cause the distribution system to leak unnecessarily."

In this project, the management regulates the flow of a single pumping

station with VSPs, maintaining an adequate level at the elevated reservoir. In this way, the network pressure is balanced with the reservoir nodal pressure. The chart in Figure 13 gives an overview of the advantages of adequate pressure management for different sectors:

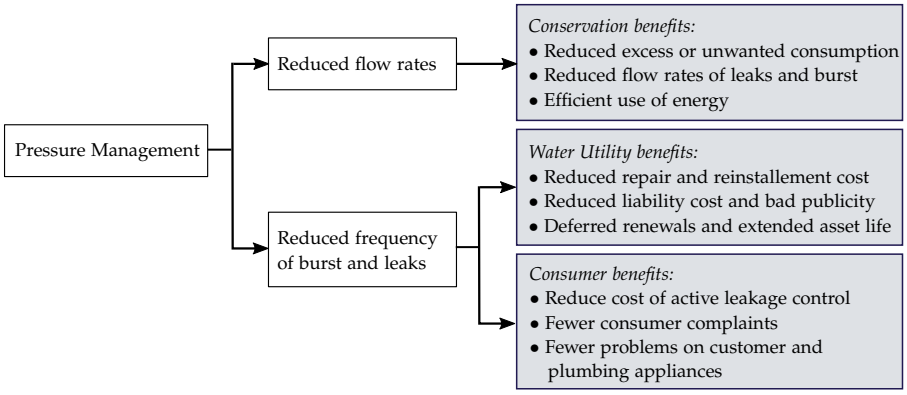


Figure 13: Chart of the optimal pressure management benefits [22]. Remark that the energy usage can be considered as a utility benefit since it also reduces the operational costs.

This overall objective must be achieved considering an operating criterion that is defined by multiple operational objectives. A list of the most common is given below [23], remark that the importance of each objective might change according to the utility needs:

1. *Safety of the supply:* To satisfy the water demands at any time. For networks with an elevated reservoir, or storage capacity, this objective implies that a volume of water must be stored in the tank as a protection mechanism against peaks in water demand, fire emergencies, service, etc.

2. *Water quality:* To reduce the deterioration of the water inside the network. This means that the management must avoid water stagnation in the pipes and tanks, and the stored water must be regularly renovated.

3. *Smoothness in actuations:* To reduce the switching in the pump actuations. This means that the management must operate the pumps continuously and smoothly, thus avoiding the pressure peaks that accelerate the deterioration of the distribution infrastructure.

4. *Minimize costs:* To reduce the economic operational costs of the distribution that are mainly associated with the energy consumption of the pumps.

3 State-of-the-Art

3.1 Management of WDNs

The vast majority of Water Distribution Networks (WDNs) with elevated reservoirs are managed by rule-based controllers. This kind of control schedules the operation of the pumps with a condition on specific network parameters, typically tank levels. Rule-based management provides an On/Off policy that is easy to maintain; however, it requires further optimisation to achieve adequate pressure management since the actuation introduces pressure peaks in the network and lowers the efficiency of the pumping system.
This project considers rule-based controllers as the baseline for improving the operation of a water network. This section gives an overview of the existing control solutions that both industry and academia have contributed to supporting and improving water infrastructure management.

Industrial Solutions

Decision support tools: As introduced in Section 2.2 the implementation of recent advances in monitoring can considerably improve the control strategies of a water utility. Some software tools like Aquis, TaKaDu and Visenti [24–26] help to understand the network operation to support management decisions. However, these tools must be configured and calibrated for each network. In Demark, DHI, Krüger (Veolia), Envidan [27–29] are examples of consultant companies that offer software services that analyse the collected data for providing an improved operation of the studied network. Nevertheless, the economic investment for implementing these support tools is limiting, especially for small-medium-size utilities.

Optimal management: Schneider Electric in collaboration with VandCenter Syd developed an optimal pressure management solution [24], Rockwell Automation offers a management solution with Model Predictive Control (MPC)[30]. However, both of them require complex system models. Similarly to the decision support case, the implementation of this technology implies high commissioning costs in addition to qualified personnel to calibrate and maintain the controller.
Other companies, like Xylem, Grundfos or I2O [31–33] offer easy to maintain (adaptive) pressure management products which do not require a recalibration.

Academic Solutions

Several research studies address the improvement of the management of water distribution via automated control and digitalisation. Most of these studies deal with the management issues that Section 2.2 states. However, their methodologies differ.
Advanced rule-based solutions are developed to provide optimal management [34, 35]. This control technique requires calibrating the commissioned controller for the specific network. Optimal control strategies are presented in [36–40] most of them use an MPC optimization framework. MPC is a model-based optimal control method that provides a policy over a control horizon. This control method is widely understood in academia; for instance, Economic MPC (EMPC) is applied to find a control policy based on an economic-oriented cost function that includes the process objectives [41]. When MPC is benchmarked against an On/Off strategy, most of the cited results conclude that the MPC reduces the operational cost and handles other operational objectives. However, the accessibility issue of this method from some utilities is rarely addressed.
Robust MPC [42] or stochastic MPC approaches are proposed in [43] to deal with the uncertainty that the network models do not capture.
Alternatively, there are several studies that use data-based approaches to improve the operation of water network: iterative learning control [44], dynamic programming [45] or automatic system identification [46]. A *plug-and-play* solution that adapts to the system automatically can reduce the commissioning and maintenance costs. The Section 3.2 presents a more detailed overview of learning controllers that could solve this issue.

3.2 Learning controllers

Introduction

Water utilities strive to build a high fidelity model for multiple reasons, such as outdated or absent system information or the requirement of qualified external personnel for commissioning and maintenance. These reasons pose a significant increase in the cost. Small-medium size utilities struggle to formulate a business case that justify such economic effort. Consequently, the implementation of optimal model-based controllers in these industrial applications is limited because its performance depends on a model.
This issue motivates the study of data-driven controllers that in the absence of a system model, the system information can be extracted from the collected data: Extremum seeking control is a model-free control method that is useful for adapting parameters when the system dynamics are unknown and the mapping of control parameters to an objective function is unknown

[47]. Iterative Learning Controller gradually adapts a control policy by minimising the error of the system's output [48]. This method is beneficial for reference tracking in periodic systems that execute a particular operation repeatedly. Moreover, Reinforcement Learning is a machine learning method that adapts the control policy based on the interaction with an environment and by collecting rewards.

Reinforcement Learning

When a system model is available, Dynamic Programming (DP) methods are applied to solve optimal control problems [49]. However, the model is not always available or easy to formulate. Similarly, Reinforcement Learning (RL) [50] algorithms are proposed as an optimisation method to find a near-optimal policy without system knowledge, well-known are the results from DeepMind in playing AlphaGo and Atari games [51, 52]. Promising results are also obtained when the RL is applied in control for robotics applications [53, 54]. In [55] and [56], RL and classical optimal controllers like Linear Quadratic Regulators (LQR) are combined to solve Riccati equation in real time without knowledge of the system dynamics.
The usefulness of RL in control is also studied in industrial applications. The domain of application and the problem formulation considerably differs, to mention a few examples in water applications: Reservoir optimisation [57], multiagent RL in combination with MPC [58], other application for tank filling control [59, 60] or detection of cyberattacks [61].

Function approximator: RL methods such as Q-learning are originally proposed to find an optimal control policy in a Markovian domain [62]. In this problem formulation, the Q-values are stored for each state-action pair. This representation becomes quickly an issue due to the course of dimensionality. *Function approximators* are proposed to find a compact representation of the Q-value space that is referred to as Q-function. There are several methods to describe this compact representation, parametric or non-parametric.
The linear parametric is a typical structure of the Q-function. The mapping of this function is characterised by a coordinate vector, or weights, and a selection of Basis Functions (BFs).
The BFs are selected to fit the system dynamics; therefore, different function structures are proposed. Some of the most commonly used are listed in [50]: polynomial, Fourier or Gaussian radial basis. In [63], fuzzy approximators are combined with Q-learning to partitioned representation of a continuous state-action space.

Disturbance rejection: Traditional control theory methods are applied together with an RL optimisation framework to address classic control chal-

lenges; for instance, a policy iteration method is presented in [64, 65] to solve the LQR problem with unknown system dynamics.
One of the biggest challenges of control in water distribution is the uncertainty of the water demand; from a control theory perspective, it can be seen as system disturbances. Therefore, it is important to verify the robustness of the RL method against disturbances. Other known control methods such as, small signal theorem [66], sliding mode [67] or H_∞ [68] are integrated with RL to provide robustness in control of non-linear systems or systems with disturbances.

Robust Learning: The parametric approximation of the Q-function is subject to issues associated with system identification. For instance, the model structure proposed for the approximation must be a suitable representation of the identified system, and adequate experimental conditions must be provided during the identification [69]. When dealing with model-free approaches, it is difficult to provide a model structure that perfectly fits the system in advance. High dimensional approximation spaces might provide an accurate description of the system. However, large approximation spaces complicate the implementation and identification. Some feature selection methods are proposed to reduce dimensionality by selecting the most relevant subsets. In [70–73] different feature selection strategies are proposed to increase the learning robustness of the RL algorithm.
Moreover, the identification requires an adequate persistence of excitation in the input signal that excites all the modes and subsequently facilitates the convergence of the identification algorithm. Some of the aforementioned work [66, 67, 74] include Persistet Excitation (PE) in the input signals of their algorithms. However, the impact of this signal also decreases the performance of the controlled system. The experience replay method is utilised in [75] to have more efficient use of the collected data and relax the PE requirements in real-time.
Alternatively, the parameter identification with Least Squares can be improved by using regularisation terms in the loss function. The identification of overdetermined systems can lead to failures due to the redundancy of some parameters; standard least squares assumes that all the parameters in the identification are equally important. However, it might be convenient to reduce the vector space with a norm-1 regularisation term to allow sparsity of the identified solution [76]. On the other hand, an underdetermined system, where there are fewer measurements than unknown parameters, might require a penalisation of the sparsity with a Tickhonov regularisation [77]. This regularisation introduces a norm-2 term in the loss function to mitigate the effect of an ill-posed identification problem.

3.3 Safety

Water supply systems are critical infrastructures [78], and their operation is essential for the functioning of society and the economy. Therefore, their physical infrastructure and operation must be resilient. The supply must be maintained, thus protected from malicious attacks and the effects of changing environments. Therefore, the system's safety should be considered a top priority in the development of new solutions for water systems and, subsequently, the technology supporting management of such infrastructures.
Some of the aforementioned RL control methods show promising results for finding optimal policies in simulated environments or where a learning agent can explore a large domain of the state-action space with no risks. When learning simple tasks, a reward function might be sufficient to determine the correct behaviour of the system. However, this is not the case for complex tasks or safe-critical systems like water infrastructures where the agent faces a changing environment while pursuing multiple operational objectives.
Completing a rich RL training implies that the system must experience diverse operation scenarios, including safe and unsafe. Therefore, learning from experiences might conflict with the operation that critical infrastructures require, where a failure might significantly impact the economy and society. Providing safety in an RL framework is an essential control challenge; this project considers safety a stepping-stone for further developing learning controllers in industrial applications.
The safety problem in RL is addressed from different perspectives, [79] presents a broad overview of safe reinforcement learning approaches where the safety methods are divided into two categories. Their control scheme is represented in Figure 14, and a brief description of the two categories and related work is given as follows:

- *Optimisation criterion:* The safety criteria are encoded together with the reward function, thus modifying the optimality criterion. For instance, by modifying the predefined weight in the objectives, [80] presents an RL control with an input constraint method using low-gain feedback. By adding barrier function terms the reward (objective) function can be modified to include safety [81–85]. In [86] model-based RL with statistical models are utilised to provide a safe optimisation framework.

- *Exploration process:* The safety criteria are not included in the learning controller. An external module (filter) is introduced between any legacy controller and the system. This filter incorporates knowledge to supervise that the applied action is safe and, if required, modify the action and ensure safety. When the legacy controller is an RL controller, this method first solves an unconstrained problem and then applies a safe filter to limit the exploration according to some safety criteria.

Some studies share this control structure, but the safety method differs. For instance, reachability analysis, invariance-based and control barrier safety filters are proposed in [87–89] respectively. However, the explicit computation of a safe set and controller can lead to an expensive computation or conservatism. Using knowledge of a linear system, [90, 91] proposes a safe exploration. A Learning-Based MPC (LBMPC) is presented in [92] to build a safety filter with support of Gaussian Process regressions, thus reducing the reliance on the system knowledge.

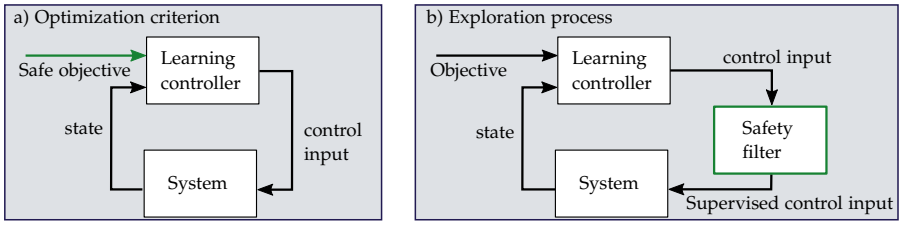


Figure 14: Control structure of the two main categories of safe learning: Optimization criterion and exploration process. The input of the safety criteria is highlighted in green.

There are other safety methods that combine learning with other classic control methods. For instance, the usage of MPC in combination with RL is studied to exploit the learning capabilities of RL and the constrained optimisation framework of MPC [93, 94]. LBMPC are also presented in [95] for iterative control task and in [96] a constrained optimisation framework from MPC is combined with a Gaussian Process regression to describe the uncertainty of imperfect system models. Alternatively to LBMPC, [97] presents differentiable predictive control, a method for learning constrained neural control policies for linear systems.

4 Research Objectives

As previously shown in the state-of-the-art in Section 3, optimal control techniques are suitable for regulating large-scale infrastructures. Nonetheless, the success of these optimization-based techniques depends, to a great extent, on the system model that must be regularly calibrated to represent the behaviour of the system. This project's scope is to give access to small and medium-sized utilities for optimal control solutions. The commissioning and maintenance of optimal controllers are not always economically feasible for small network operators. Data-driven control algorithms provide a system policy in absence of a system model, thus reducing the associated costs.
The overall objective of this project is to integrate the advances in optimal control in an easy-commissioning controller to provide both an economically

feasible controller and improved efficiency of the network. Specifically, to develop a self-learning control solution that safely finds an optimal management policy.
This project's scope is to investigate the potential benefits and implications of introducing a model-free control solution to manage the operation of water distribution networks. To this end, four research objectives are defined:

Research Objective 1: *To develop a model-free controller*

In this work, the purpose is to find a common ground between efficient and complex solutions proposed by academia and the real utility needs. For this purpose, a Reinforcement Learning control method is proposed to find optimal management that gradually adapts to the changing environment of a water distribution network with an elevated reservoir.

- *With disturbance rejection:* Water distribution networks are critical infrastructures where the drinking water supply needs to be guaranteed. Utility management is required to maintain a robust operation. Therefore, the designed controller must satisfy some operational objectives while rejecting the stochastic disturbances of the system.
- *With a robust learning:* Reinforcement Learning is a data-driven optimal control solution that does not rely on a system model. However, it depends confidently on the quality of the data. RL provides a policy based on its Q-value function. This function is initially unknown, and its parameters must be identified with collected data. This identification process is subject to issues such as poor experimental conditions.
  Formulating suitable conditions for the identification while operating the system is challenging. Therefore, the proposed solution is required to cope with situations where the quality of the collected data is low.

Research Objective 2: *To provide safety during learning and operation*

When the learning and control strategies are developed in objective 1, this project targets the safe operation of the water network. The learned policy given by the Reinforcement Learning algorithm is not necessarily safe, and it can drive the system to areas where there is a risk of failure. Therefore, a policy supervisor module is proposed to detect unsafe system trajectories and rectify the operation accordingly.

Research Objective 3: *Experimental validation of the solution*

The deployment of newly developed control solutions requires careful verification, and performing these tests in real water infrastructures puts the network operation at risk. For this reason, the construction of a laboratory is proposed to validate the performance of the developed methods. This modular setup must emulate small-scale water infrastructures and reproduce control scenarios according to a particular study case. Thus, allowing to study of the feasibility of the management solutions safely.
Furthermore, the modularity of this facility enables the laboratory tests for other control solutions or application domains that are not in the scope of this project, like leakage detection, wastewater collection or district hearing systems.

5 Contributions

The main contributions of this project are structured into three areas, similarly to the presented research objectives in Section 4. A chart, relating the research outcome and the papers, is given in Figure 15. Subsequently, a brief description of each paper contribution is given below, and the papers in full are found in Part II.

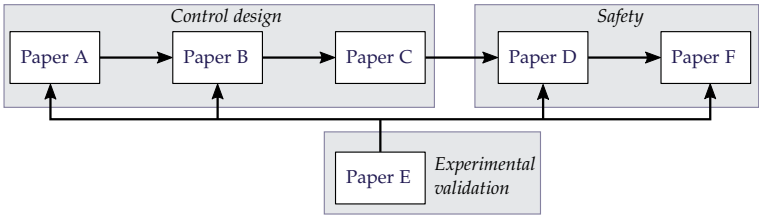


Figure 15: Simplified dependency chart where the papers are classified by research area.

Control Design

Paper A

J.V. Ledesma, R. Wisniewski and C.S. Kallesøe. *"Optimal Control for Water Distribution Networks with Unknown Dynamics."* 21^{th} IFAC World Congress, 2020 Vol. 53(2), pp.6577-6582, 2020.

Proposal of a model-free controller for reference tracking and rejection of constant disturbance in linear systems. This method uses an RL method based on a state-space augmentation that includes the reference and integral error in addition to the system state vector. Then, considering that the system is

linear, and the objective function is quadratic, a set of second-degree polynomials is built with a combination of the states and control action. These polynomials are used as basis functions to iteratively approximate the Q-value function with the Least Squares Temporal Difference algorithm (LSTD).

> **Paper B**
>
> J.V. Ledesma, R. Wisniewski and C.S. Kallesøe. *"Reinforcement Learning Control for Water Distribution Networks with Periodic Disturbances."* 2021 American Control Conference (ACC), pp. 1010-1015, 2021.

Proposal of an RL control for rejection of unknown periodic disturbances. In this paper, the periodic signal described by the total water demand is represented with a Fourier Series approximation. The corresponding Fourier harmonics are merged with the system by augmenting the state space. By having a linear model structure, the RL problem is formulated similarly to Paper A. Moreover, assuming that the system operates around an operating point, the minimization of the energy usage is included in the management objectives.

> **Paper C**
>
> J.V. Ledesma, R. Wisniewski and C.S. Kallesøe. *"Real-Time Reinforcement Learning Control in Poor Experimental Conditions."* 2021 European Control Conference (ECC).

Proposal of two control algorithms that provide numerical robustness during the learning, especially in situations with poor experimental conditions. When the measured data do not contain system information, the method might lead to failures in the approximation. The first algorithm identifies data batches with low system information using the Fisher information matrix and pauses the approximation of the Q-value function. The second algorithm sorts the data based on the amount of system information. This analysis is performed with Singular Value Decomposition; subsequently, this method discards the features or basis functions that do not provide value to the estimation.

Safety

> **Paper D**
>
> J.V. Ledesma, R. Wisniewski and C.S. Kallesøe. *"Safe Reinforcement Learning Control for Water Distribution Networks."* 2021 IEEE Conference on Control Technology and Applications (CCTA) pp. 1148-1153, 2021.

Proposal of safe exploration for an unconstrained RL algorithm. This method

provides local robustness nearby the boundaries using a safety filter. This filter assesses the operational risk of the RL control input and corrects the system trajectory in case the prediction is unsafe. Both prediction and safe control input are computed with a linear deterministic model. This method assumes that the core system dynamics are known at the boundaries. Therefore, a guess of the tank size and average water demand is required. The filter provides conservative or relaxed supervision depending on the model's accuracy.

> **Paper F**
>
> J.V. Ledesma, R. Wisniewski, C.S. Kallesøe and A. Tsouvalas. *"Water Age Control for Water Distribution Networks via Safe Reinforcement Learning."* Submitted for journal publication 2022.

Proposal of safety exploration of an unconstrained RL algorithm. This method extends the safety filter presented in Paper D by including a Gaussian Process regression to model the uncertainty between an imperfect linear model and the real system. Thus, providing a better prediction of the exploration and accurate response that ensures safety. The GP model is trained in real-time. The model's confidence interval is used as an exploration guideline, providing conservative supervision when the GP model is not trained and then more relaxed supervision to the RL when the GP model training is completed.
Furthermore, the water quality (water age in the tank) is incorporated into the control strategy by formulating it as a safety problem. Hence, the safety filter actuates as a fallback control when the water age indicator rises above a safety threshold.

Validation

> **Paper E**
>
> J.V. Ledesma, R. Wisniewski and C.S. Kallesøe. *"Smart Water Infrastructures Laboratory: Reconfigurable Test-Beds for Research in Water Infrastructures Management."* Journal of Water, Special Issue Advances in the Real-Time Monitoring and Control of Urban Water Networks Vol. 13(13), 2021.

Proposal of a modular laboratory facility that enables the experimental validation of the designed control strategies. On a small scale, this laboratory setup emulates the main features of water distribution networks, wastewater collection and district heating. The modular structure allows customizing each testbed to the desired study case. This flexibility enables its use for a wide range of control problems like leakage detection, fault-tolerant or optimal management. This paper presents a description of the facility with the

development considerations and some examples where control strategies are validated against management scenarios that cannot be safely replicated in real-scale water infrastructures.

Summary

This chapter summarises the study carried out in this project. It comprises a brief description of the experiments and laboratory configuration in Section 6, an overview of the different models used in this work in Section 7, the development of a model-free controller via Reinforcement Learning in Section 8, and a safety filter that provides safe exploration in Section 9. This summary presents the methods conceptually and only the most representative results are discussed; each section includes references to the corresponding papers where the details of the method and results are found. In this summary, the paper notation is revised to maintain a consistent description of the methods along the summary sections. This chapter closes in Section 10 with a discussion and conclusion where the stated objectives are compared with the project contributions and suggestions for future work.

6 Experimental validation

> This section presents an overview of the experimental validation, first with a brief description of the *Smart Water Infrastructures Laboratory* from Paper E, and then by describing the testbed configuration used in Paper A, Paper B, Paper D and Paper F.

This project selects Bjerringbro (Denmark), a small-size urban district, as a study case. Figure 16 shows the map and the WDN layout that consists of a single pumping station in the south, an elevated reservoir, and two Pressure Zones (*PZ1* and *PZ2*). Additionally, a second pumping station boosts the pressure in the *PZ2*.

Two modular testbeds are built with the SWIL. These testbeds emulate, on a small scale, the hydraulic configuration and communication architecture of the study case, see Figure 4 where the communication architecture is illustrated on top of the laboratory modules picture.

Testbed 1 covers *PZ 1* which is emulated with two consumer units, and a ring topology network interconnects with pipes the network components. The elevation of different consumers is emulated with pressurised air at the col-

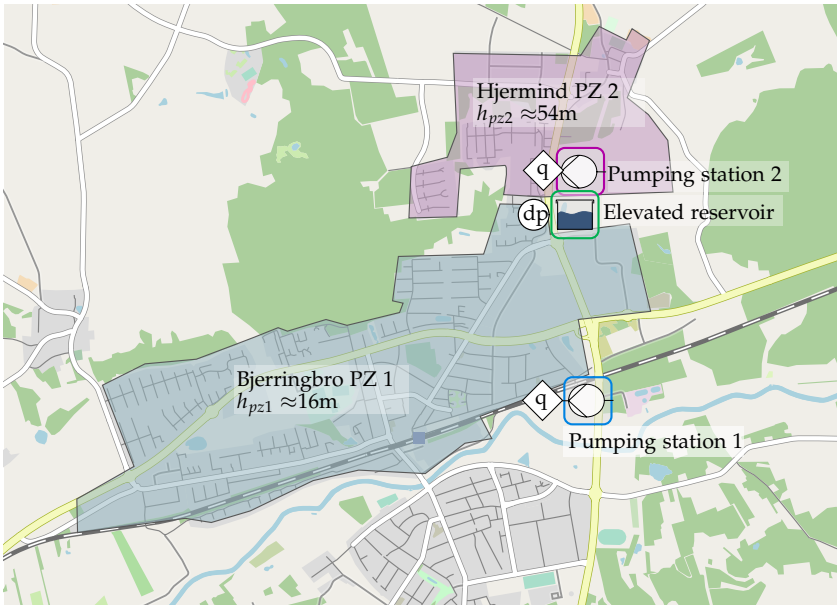


Figure 16: Map of a water distribution network with an elevated reservoir in Denmark. Pressure Zone 1 (PZ 1) covers the Bjerringbro district, and Pressure Zone 2 (PZ 2) covers Hjermind that are located at a different elevation (h_{pz}) [Paper F].

lection tanks, similarly to the elevated reservoir. The valves have local PI controllers that regulate the water outflow, reproducing Bjerringbro's consumption patterns. The pumping station is equipped with a flow sensor to regulate, with a local PI, the inflow of water to the network. A DAQ system collects the sensor data and provides a flow reference to the pumping station. To reduce the impact of these local controllers in the tests, the sampling time for local controllers is 1 second, and for supervisory control (SCADA) is 60 seconds. The data from the testbed is locally collected at the LUs with Codesys Runtime [98], and it is interfaced with the CCU via TCP/IP Modbus. The supervisory controller is implemented on the CCU with *Matlab - Simulink* where the testbed is globally monitored. A simplified pipe and instrumentation diagram of the *Testbed 1* with local controllers is illustrated in Figure 17.

Testbed 2 extends the hydraulic network of the testbed by including the *PZ2* with the booster pumping station and its consumption district. A simplified pipe and instrumentation diagram of the *Testbed 2* is depicted in Figure 18.

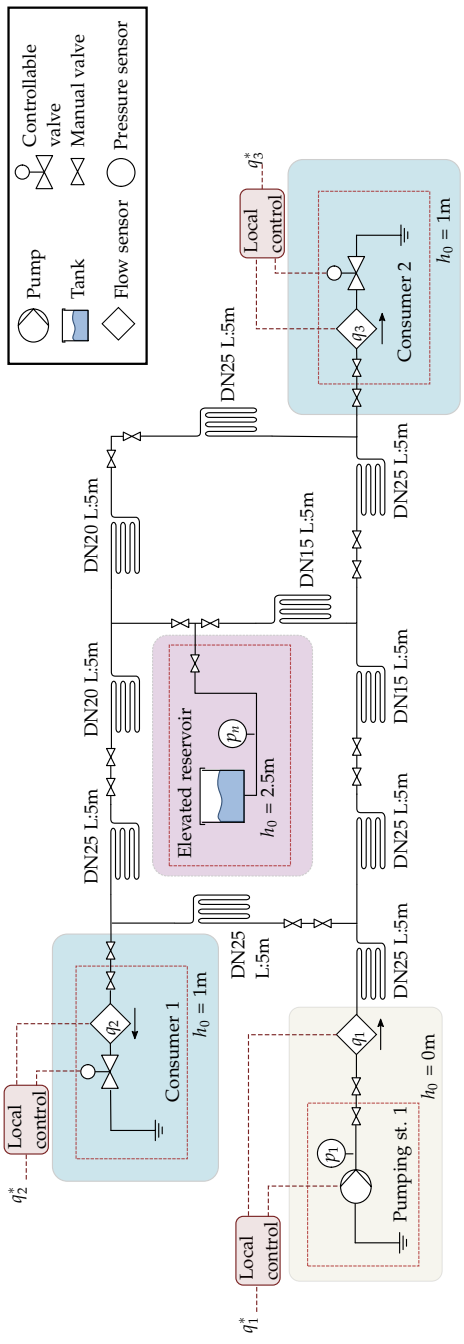


Figure 17: *Testbed 1* - Detailed topology of the laboratory setup used to validate the controllers in Paper A, Paper B, and Paper D. DN represents the pipe diameter in mm and L the pipe length, and the notation (*) represents the reference signal received from the SCADA -PC

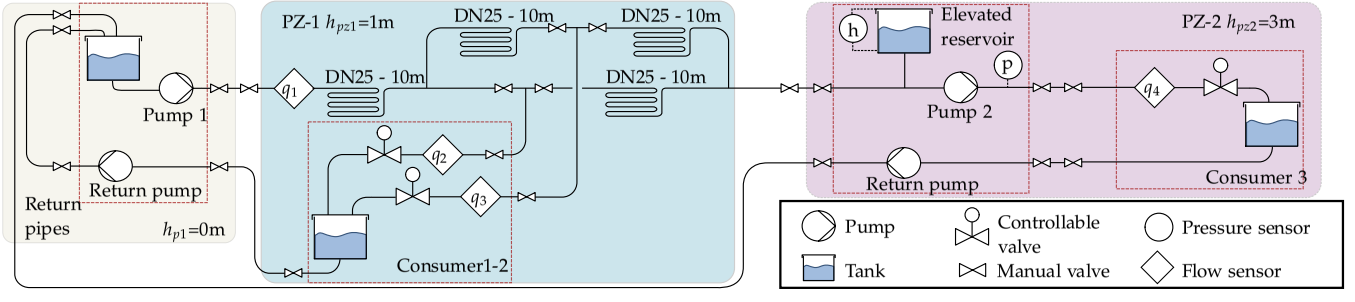


Figure 18: *Testbed 2* - Scheme of the laboratory setup used to validated the controller in Paper F. It consists of a main pumping station (grey), two Pressure Zones (PZ1-blue and PZ2-pink) and an elevated reservoir with a booster station (pump 2). The water is collected at the consumer tanks and recirculated to the supply reservoir. A legend is shown on the bottom-right side.

7 System model

> This section presents an overview of the modelling work used in this project, an extended description is provided in Paper A, Paper B and Paper F.

The main obstacle to implementing optimal control solutions is the cost of developing and maintaining a system model that accurately represents reality. The motivation of this project is built upon the need for a model-free controller that facilitates the implementation of optimal management solutions.
This project proposes a grey-box approach where partial knowledge of the system is utilized to support learning an optimal management policy. For this purpose, this section summarizes the development of a control-oriented model that incorporates this partial system knowledge.
The (partial) system knowledge is restricted to available sensor measurements and infrastructure information. This project proposes using mainly linear models that represent the essential system dynamics.

7.1 Water distribution network

Water distribution networks are systems where a complex pipeline network interconnects the main components. Some similarities are found between water distribution networks and electric circuits, network topology and component behaviour wise. For instance, voltage/current sources in electric circuits have an analogous purpose to pumping stations which regulate the supply pressure/flow in water networks. Similarly, resistors are equivalent to pipes and capacitors to tanks. Some methods for circuit analysis like Norton-Thevenin equivalences allow a representation of complex circuits in a simplified scheme. Likewise, in this work, the essential behaviour of a WDN is represented by an equivalent model which considers the following assumptions:

1. **Aggregated consumers:** The water demands from the multiple end-users in a city district are aggregated and expressed as a total water demand d_{total} [Paper A]. Recall that the total water demand is not measured in real-time.

2. **Fast dynamics are neglected:** WDNs with elevated reservoirs are stiff systems that contain fast varying components (pumping station and pipe dynamics) and slow varying (elevated reservoir). The elevated reservoir dynamics is considered dominant in the system, and therefore the pressure-flow transients at pumping stations and pipes are neglected.

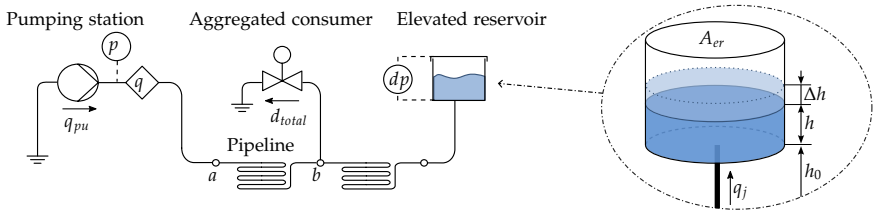


Figure 19: (Left) A simple scheme of a water distribution network where the measurements available are pressure (p) and flow (q) at the pumping station and water level at the elevated reservoir is measured with a differential pressure sensor (dp). (Right) Tank geometry and variables where A_{er} is the cross-sectional area, Δh is a level variation, h tank level, h_0 is the physical elevation of the tank and q_j is the inflow of the tank inlet j.

The scheme in Figure 19 shows a simplified WDN where the *pumping station* is an ideal flow source, the *aggregated consumer* represents the demand of an entire city district, the *elevated reservoir* is a linear tank. These three main components are interconnected with two pipelines representing the pipe network's friction. With these simplifications, the behaviour of a WDN system is described by a continuous-time linear model as follows,

$$A_{er}\dot{h}(t) = q_{pu}(t) + d_{total}(t). \tag{1}$$

where $h(t)$ is the tank level, $q_{pu}(t)$ is the controlled inflow at the pumping station and $d_{total}(t)$ is the total water consumption (disturbance). The pressure drop in an equivalent pipeline that connects nodes a and b is given by

$$\Delta p_{ab}(t) = \underbrace{r_{ab}|q_{ab}(t)|q_{ab}(t)}_{\text{pipe friction}} + \underbrace{\Delta h_{ab}}_{\text{elevation}}, \tag{2}$$

where the r_{ab} is a constant representing the surface resistance, this form of the friction losses in (2) assumes that the flow is turbulent [14], the second term represents the pressure due to the differential geodesic level between the two nodes, see Paper E for further pipe model details. Finally, a simple model of the power of the pump $P(t)$ is defined as

$$P(t) = q_{pu}(t)\Delta p_{pu}(t)/\eta, \tag{3}$$

where Δp_{pu} is the pressure across the pump and η is a constant representing the pump efficiency [Paper B].

7.2 Disturbance

The uncertainty in the water demand is a major challenge in managing water distribution networks. The individual consumption is not measured, but the utilities can reconstruct the past consumption using billing data. Although

this data cannot be directly used for real-time control, it contains information about daily consumption patterns. Typically, consumption patterns present two peaks, one in the morning and the other in the afternoon. This pattern is more evident when the individual consumptions are aggregated into a signal that represents the total demand.
This summary describes the approximation of the periodic signal with a Fourier Series of order N as follows, the complete model formulation is presented in Paper B. First, the continuous signal is given by

$$\bar{d}(t) = a_0 + \sum_{n=1}^{N} (a_n \cos(\omega_n t) + b_n \sin(\omega_n t)) + w, \tag{4}$$

where a_0, a_n and $b_n \in \mathbb{R}$ are the Fourier coefficients, $\omega_n = 2\pi n f_0$ and f_0 represents the fundamental frequency and w is Brownian noise. The f_0 represents the lowest frequency, and it is set by a period of one day. Paper B formulates

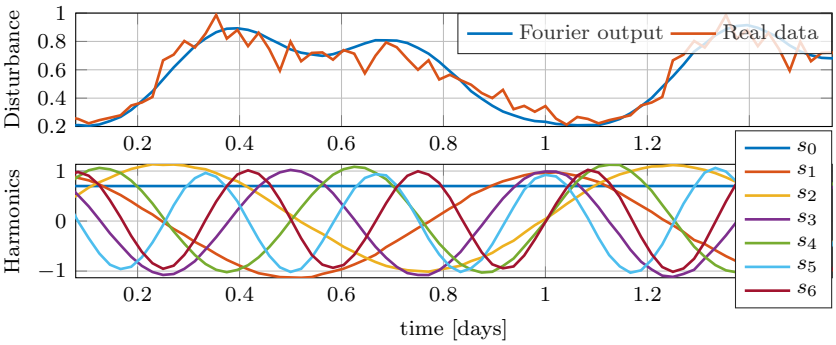


Figure 20: (Top) Output disturbance signal and consumption data for two days. (Bottom) Fourier Harmonic states for n=0,1,2,3 [Paper B].

the mean of the Fourier output (4) in its discrete version as follows,

$$\begin{aligned} s_{k+1} &= A_d s_k, \\ d_k &= C_d s_k, \end{aligned} \tag{5}$$

resulting in a linear state-space form, where the system matrix $A_d = \text{diag}(1, F_1, \ldots, F_N)$, with $F_n = \begin{bmatrix} \cos(\omega_n \Delta t) & -\sin(\omega_n \Delta t) \\ \sin(\omega_n \Delta t) & \cos(\omega_n \Delta t) \end{bmatrix}$ where Δt is the sampling time and the output matrix C_d includes the Fourier coefficients that scale the harmonics. The state vector $s_k \in \mathbb{R}^{n_d}$, with $n_d = 2N + 1$, is subject to the following initial condition

$$s_{i,t_0} = \begin{cases} c_0 & \text{if } i = 0 \\ \cos(\omega_n t_0) & \text{if } i > 0, \quad i \text{ odd} \\ \sin(\omega_n t_0) & \text{if } i > 0, \quad i \text{ even} \end{cases} \tag{6}$$

where c_0 is a constant, t_0 is the initial time value and the index vector $i \in \mathbb{Z}, [0, n_d]$.
The top graph in Figure 20 compares real consumption and the Fourier approximation signal. The Fourier output follows the measured data, describing the morning and evening water consumption peaks.
The state vector includes Fourier harmonics for a given frequency, and the scaling of the signal is performed with the output matrix C_d. This form is strategically selected for formulating a controller that learns the scaling of the signal only with measurements of normalised harmonics.

7.3 Augmented state space - Control model

This project aims to build a model-free controller that rejects system disturbances. In Section 8, the system model structure is utilised to build a learning controller. For this purpose, two control-oriented models are formulated in Paper A and Paper B. Both models are developed with a linear system and partial knowledge of its disturbances. The models differ from each other in the way that the disturbance is introduced. The structure of each model is described below:

In Paper A, the model is built for tracking a reference and for compensating constant disturbances. First, consider (1) in a continuous-time state space form,

$$\begin{aligned} \dot{h}(t) &= A_c h(t) + B_c q_{pu}(t) + W_c d_{total}(t) \\ y_c(t) &= C_c h(t), \end{aligned} \tag{7}$$

where $h(t) \in \mathbb{R}$ represents the tank level, $q_{pu}(t) \in \mathbb{R}$ the controlled inflow and $d_{total}(t) \in \mathbb{R}$ is a disturbance representing the total water consumption, with A_c,B_c and C_c constant matrices with compatible dimensions. Then, the trajectory of the reference, $\dot{r}(t) = Lr(t)$, and the integral error, $\dot{\xi}(t) = y_c(t) - r(t)$, are combined with (7) in an augmented state-space model,

$$\begin{bmatrix} \dot{h} \\ \dot{r} \\ \dot{\xi} \end{bmatrix} = \begin{bmatrix} A_c & 0 & 0 \\ 0 & L & 0 \\ C_c & -I & 0 \end{bmatrix} \begin{bmatrix} h \\ r \\ \xi \end{bmatrix} + \begin{bmatrix} B_c \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} q_{pu} \end{bmatrix} + \begin{bmatrix} W_c \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} d_{total} \end{bmatrix}. \tag{8}$$

This model is not further developed in Section 7, the details are given in Paper A.

In Paper B, the model is built to compensate periodic disturbances. First, the tank model presented in (1) is expressed as a linear discrete-time system in the state-space form,

$$h_{k+1} = Ah_k + Bu_k + Ed_k \tag{9}$$

where $h_k \in \mathbb{R}$ is the system state, $u_k \in \mathbb{R}$ is the controlled input flow and $d_k \in \mathbb{R}$ are the system disturbances, and A,B and E are constant matrices with compatible dimensions. Then, discrete-time tank model (9) and periodic signal (5) are combined as follows,

$$\begin{aligned} x_{k+1} &= \underbrace{\begin{bmatrix} A & EC_d \\ \mathbf{0} & A_d \end{bmatrix}}_{A_e} x_k + \underbrace{\begin{bmatrix} B \\ \mathbf{0} \end{bmatrix}}_{B_e} u_k, \\ y_k &= \underbrace{\begin{bmatrix} \mathbf{I} \\ C_p \end{bmatrix}}_{C_e} x_k + \underbrace{\begin{bmatrix} \mathbf{0} \\ D_p \end{bmatrix}}_{D_e} u_k, \end{aligned} \tag{10}$$

where $x_k = \begin{bmatrix} h_k & s_k \end{bmatrix}^T$, u_k is the controlled input and $y_k = \begin{bmatrix} h_k & s_k & p_k \end{bmatrix}^T$ is the measured output vector. Moreover, note that the vector y outputs an additional state, corresponding to the pressure at the pumping station, that is used for energy optimisation. The pressure model is introduced in the state-space model by linearising (2), see the complete model formulation in Paper B. Finally, (10) is represented in a compact form in discrete-time,

$$\begin{aligned} x_{k+1} &= A_e x_k + B_e u_k, \\ y_k &= C_e x_k + D_e u_k, \end{aligned} \tag{11}$$

where $x \in \mathbb{R}^{m_a}$, $u \in \mathbb{R}^{n_a}$. For simplicity, the notation in model (11) is used as reference for describing the control method in Section 8.

7.4 Tank turnover

Maintaining adequate levels of water quality is highlighted in Section 2.2 as a crucial management objective. In this project, the water sources are assumed to have sufficient quality to be distributed to the end-users without chlorine treatment. Therefore, in a WDN with elevated reservoirs, the water quality issue arises mainly when the water is stored in the tank for long periods.
In Paper F, a daily turnover signal is proposed to monitor the tank's water age and regulate the daily inflow of freshwater with respect to the stored water. First, the Average Residence Time (ART) in an elevated reservoir is defined as a discrete-time variable [100],

$$ART_k = \frac{v_{av,k}}{q_{av,k}}, \quad \text{for } q_{av,k} > 0, \tag{12}$$

where $v_{av,k}$ is the average volume, and q_{av} is the average flow entering the tank. Having WDN with flow measurements at the elevated reservoirs is unusual. Therefore, the turnover is reformulated to be computed with level

measurements. Firstly, the average volume is defined with past level measurements,

$$v_{av,k} = \frac{A_{er} \sum_{i=k-n_{av}}^{k} h_i}{n_{av}}, \tag{13}$$

where n_{av} represents the number of samples in a period. Secondly, the average inflow is denoted as,

$$q_{av,k} = \frac{A_{er} \sum_{i=k-n_{av}}^{k} \delta_i}{n_{av}}, \tag{14}$$

where δ represents only the positive level variations,

$$\delta_{i+1} = \max\{(h_{i+1} - h_i), 0\} \tag{15a}$$

$$= \max\{(A_{er}^{-1} \sum_{j=1}^{n_{er}} q_{j,i} \Delta t), 0\} \tag{15b}$$

where $q_{j,i}$ is the flow at tank inlet j at time i. Note that (15) has two expressions depending on the available system knowledge and tank configuration. For notational simplicity, $\delta(q_k)$ is represented as a function of the tank flows. Figure 19 (right) illustrates a simple elevated reservoir configuration with one inlet in the bottom of the tank, Δh shows both the level variations, positive and negative, for a sampling time Δt. Subsequently, the daily volume turnover [%] is denoted as a function of the tank level and network flows,

$$\tau_k = g(h_k, q_k) = 100 n_{av} \frac{q_{av,k}}{v_{av,k}}, \tag{16}$$

where n_{av} represents the number of samples in a period.

Incremental average approximation

The turnover model (16) is a non-linear function that requires data storage of past measurements to be computed. Then, to facilitate the real-time computation of the signal, an approximation is proposed in Paper F. First, this approximation calculates the mean of the positive level variations,

$$m_k(q_k) = \frac{1}{n_{av}} \sum_{i=k-n_{av}}^{k} \delta_i(q_k) \tag{17}$$

Then, the new mean is computed by extending (17) and including the new input (15) as follows,

$$\hat{m}_{k+1} = \frac{1}{n_m}(\delta_{k+1}(q_k) + (n_m - 1)\hat{m}_k), \tag{18}$$

where n_m is a constant representing the number of data-points in the moving average filter. Subsequently, the daily turnover output is approximated as follows,

$$\hat{\tau}_k = \hat{g}(q_k) = 100\frac{\hat{m}_k}{h^*}, \tag{19}$$

where h^* is a constant representing the average level during steady-state operation, and q_k is the sum of tank inflows. The graph in Figure 21 compares the turnover signal (16) with different approximations of signals (19), the daily turnover signal τ and its approximation $\hat{\tau}(h)$ converge when steady state is reached. A deviation of the $\hat{\tau}(\tilde{h})$ and $\hat{\tau}(q)$ is observed with respect to the others since (15) is evaluated with $\tilde{h}$ that is computed with a poorly calibrated model or with the sum of flows q that does not include a variable total demand d_{total} but a constant average demand d_{av}.
Further details of the water age model are presented in Paper F.

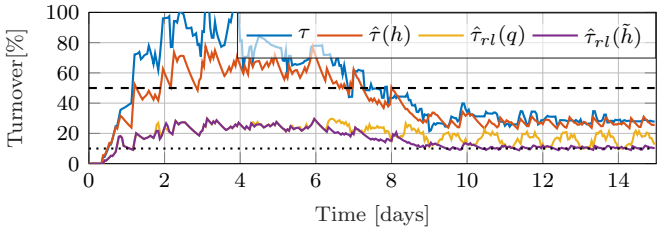


Figure 21: Simulation of the daily turnover signal and several approximations

7.5 Safety model

Nominal model

The safe exploration in Paper D is evaluated with the system trajectory of a nominal model which consists of a linear model of the form,

$$\hat{h}_{k+1} = \hat{f}(h_k, u_k, d_{av}) = \hat{A}h_k + \hat{B}u_k + \hat{E}d_{av}, \tag{20}$$

where d_{av} is a constant representing the average of the total demand and $\hat{A}$, $\hat{B}$ and $\hat{E}$ are system matrices of the nominal model. This model is built with accessible system information. For instance, d_{av} is a guess of the total average water consumption, the matrices $\hat{A}, \hat{B}, \hat{E}$ are naively calibrated with the cross-sectional area A_{er}.
This model is a first principle model, or knowledge-based, where the accuracy of the prediction is subject to the calibration. Furthermore, the linear model cannot capture some non-linearities of the system, these model differences are more patent when the model is compared with the real system.

Combined model

Inspired by the model formulation in [96], Paper F proposes an extension of the nominal model (20) that combines a nominal model and a Gaussian Process model that represents the system uncertainty,

$$\tilde{h}_{k+1} = \underbrace{\hat{f}(h_k, u_k, d_{av})}_{\text{nominal model}} + \underbrace{B_r \hat{r}(x_k, u_k)}_{\text{GP model}}, \tag{21}$$

where B_r is an index matrix, and $\hat{r}$ is the approximated residual error between the nominal prediction and the actual measurement. The GP term aims to relax the model calibration reliance by supplementing the nominal linear prediction with a model of the system uncertainty, which comprises the calibration error of the model (20) as well as the non-linear behaviour of real systems that the nominal model cannot describe.
A GP regression is formulated to approximate the residual error $r(x_k, u_k)$. The GP regression is built as follows. First, the actual residual error is expressed as,

$$y_k = r(x_k, u_k) + w_k = B_r^{\dagger}(\underbrace{h_{k+1}}_{measure} - \underbrace{\hat{f}(x_k, u_k)}_{nominal}), \tag{22}$$

where $B_r^{\dagger}$ is the Moore-Penrose pseudo-inverse of B_r. Then, consider a training data set $\mathcal{D}$ that consists of M observations,

$$\begin{aligned} \mathcal{D} = \{ \mathbf{y} &= [y_1, \dots, y_M]^T \in \mathbb{R}^M \\ \mathbf{z} &= [z_1, \dots, z_M]^T \in \mathbb{R}^{M \times n_z} \} \end{aligned} \tag{23}$$

where $z = [x^T, u^T]^T$ denotes an input vector and y a scalar output (target). This definition is performed by assuming that each of the elements y of the output vector is independent of a given input data z_k. Then, by giving a GP prior on r with kernel $k(\cdot, \cdot)$ and prior zero-mean,

$$y \sim \mathcal{N}(0, K_{\mathbf{zz}} + I\sigma^2). \tag{24}$$

The result is a normally distributed measurement where $K_{\mathbf{zz}}$ is the covariance (or Gram) matrix of the data points such that $K_{ij} = k(z_i, z_j)$, with ij denoting the elements of the matrix, the selection of the kernel k structure and its parameterisation determines the distribution of the predicted output [96]. A squared exponential function is selected as the kernel-based on domain knowledge since the residual uncertainty is expected to show a continuous and smooth behaviour,

$$k(z_i, z_j) = \sigma_f^2 \exp(-\frac{1}{2}(z_i - z_j)^T L^{-1}(z_i - z_j)), \tag{25}$$

where L is a positive diagonal length scale matrix and σ_f^2 the signal variance. The joint distribution of the training data $\mathbf{z}$ and the test data z_* is

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K_{\mathbf{zz}} + I\sigma^2 & K_{\mathbf{z}z_*} \\ K_{z_*\mathbf{z}} & K_{z_*z_*} \end{bmatrix}\right), \tag{26}$$

where $[K_{\mathbf{z}z_*}]_j = k(\mathbf{z}_j, z_*)$, $K_{z_*\mathbf{z}} = K_{\mathbf{z}z_*}^T$, and similarly $K_{z_*z_*} = k(z_*, z_*)$. The resulting conditional distribution of the uncertainty residual is Gaussian [101]. Finally, its conditional distribution describes the GP model of the uncertainty [102],

$$\hat{r}(y_*|\mathbf{y}) = \mathcal{N}(\mu^r(z_*), \Sigma^r(z_*)), \tag{27a}$$

$$\mu^r(z_*) = K_{z_*\mathbf{z}}(K_{\mathbf{zz}} + I\sigma^2)^{-1}\mathbf{y}, \tag{27b}$$

$$\Sigma^r(z_*) = K_{z_*z_*} - K_{z_*\mathbf{z}}(K_{\mathbf{zz}} + I\sigma^2)^{-1}K_{\mathbf{z}z_*} \tag{27c}$$

where $\mu^p(z_*)$ and $\Sigma^p(z_*)$ are mean and variances of the GP. A detailed description of the GP model is given in the appendix of Paper F.

GP model training

The training of the GP model is performed in real-time by executing the Algorithm 1. The input z_k and output y_k data are stored each iteration in the Last In First Out stacks (LIFO), $\mathbf{z}$ and $\mathbf{y}$. The stacks have a fixed size of M samples, and they are used for updating the parameters of the GP model in `line 9`. The Algorithm 1 is initialised with n_{gp} that represents the number of new samples each model update, a threshold e^* that indicates an acceptable residual error, and random data for z_0 and y_0.

Algorithm 1 Training of the GP model. [Paper F]

1: **Input:** n_{gp}, e^*
2: Initialisation: $[\sigma_{f0}, L_0, \sigma_0] \leftarrow$ `fitrgp` $(\mathbf{z}_0, \mathbf{y}_0)$
3: **repeat** at every iteration k = 1,2, ...
4: collect data $\hat{h}_k$, z_k and y_k
5: e_k = `RMSE`$(h_k$ - $\tilde{h}_k)$
6: **if** $e_k \geq e^*$ **then** ▷ Collect data
7: save z_k and y_k in stack $\mathbf{z}_j$ and $\mathbf{y}_j$
8: **if** $k = (j+1)n_{gp}$ **then** ▷ GP update
9: $[\sigma_f, L, \sigma] \leftarrow$ `fitrgp` $(\mathbf{z}, \mathbf{y})$
10: j = j+1
11: **end if**
12: **end if**
13: **until**

8 Reinforcement Learning

> This chapter presents a brief introduction to Reinforcement Learning (RL) applied to the control of linear systems, the function approximators proposed in Paper A and Paper B, as well as the robust learning presented in Paper C.

8.1 Reinforcement Learning in control

Reinforcement Learning (RL) is a framework for teaching a controller (agent) to interact with a process (environment) from experience, similarly to humans that empirically learn from experience in a trial-error sequence. This interaction is defined by three signals, a control input u_k which modifies the process, a state x_k which describes the state of the process and a scalar reward r_{k+1} which provides feedback of the recent performance.

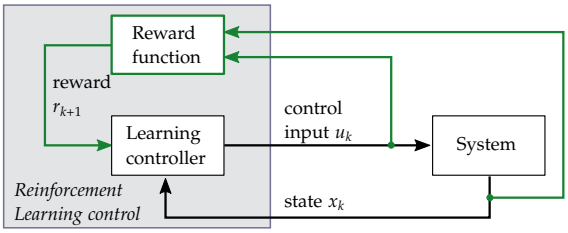


Figure 22: Scheme of the interaction between a learning controller and a system process. A control input is applied and state measurements and rewards are collected for improving the future policies.

Figure 22 presents a diagram of the interaction controller-process (agent-environment) which works as follows: the controller applies a control input to an unknown process based on a policy $u_k = \pi(x_k)$, the system reacts by changing the state $x_{k+1} = f(x_k, u_k)$, subsequently, the performance of this action is evaluated with a reward function, and a reward is obtained $r_{k+1} = \rho(x_k, u_k)$. Finally, the controller's policy is improved based on the accumulated rewards.

Bellman equation

RL problems are often formulated as maximisation problem where the aim is to maximise the accumulated rewards. However, this work formulates a minimisation problem where the objective of this method is to find a policy that minimises the cost. A V-value function that represents the collected cost

for a given policy is defined as

$$V^{\pi}(x_k) = \mathbb{E}\left[\sum_{i=k}^{\infty} \gamma^{i-k} \rho(x_i, u_i)\right], \quad \text{with } k \geq 0, \tag{28}$$

where $\gamma \in [0, 1)$ is a discount factor. From an algorithmic perspective, in an infinite horizon setting, γ ensures that the accumulated rewards are bounded, and from the application perspective, it reduces the importance of past rewards. Therefore, the selection of the discount factor poses a trade-off between high values that favour the quality of the solution and small values that favour the convergence rate [63].
RL is based on Bellman's optimality principle which is formulated as follows: first, by expanding the accumulated cost (28) for a deterministic environment,

$$V^{\pi}(x_k) = \rho(x_k, u_k) + \sum_{i=k+1}^{\infty} \gamma^{i-(k+1)} \rho(x_i, u_i). \tag{29}$$

Subsequently, the infinite sum is replaced by its value using a current policy $u_k = \pi(x_k)$, the equivalent difference equation is given by,

$$V^{\pi}(x_k) = \rho(x_k, \pi(x_k)) + \gamma V^{\pi}(x_{k+1}). \tag{30}$$

Finally, the optimal value is calculated using the *Bellman equation*

$$V^{*}(x_k) = \min_{\pi} \left[\rho(x_k, \pi(x_k)) + \gamma V^{\pi}(x_{k+1})\right], \tag{31}$$

where $(\cdot^{*})$ represents the optimal value. Alternatively, an equivalent expression to (30) is given where the value is expressed as a function of the state and control input explicitly (Q-value). Consider that $Q^{\pi}(x_k, \pi(x_k)) = V^{\pi}(x_k)$, then Q-value function is given by

$$Q^{\pi}(x, u) = \rho(x_k, u_k) + \gamma Q^{\pi}(x_{k+1}, \pi(x_{k+1})). \tag{32}$$

Subsequently, the optimal policy for a given Q-value function (32) is defined as,

$$\pi^{*}(x, u) = \underset{u}{\operatorname{argmin}}\, Q(x, u). \tag{33}$$

For convenience in the formulation, this project presents RL in a deterministic environment, and the accumulated rewards are given in an infinite horizon with a discounted return. Nevertheless, when the methods are applied to real systems, the approach becomes stochastic since the application domain and the corresponding collected data are stochastic.
In this project, the formulation of the Q-value function with Bellman equation is first introduced in Paper A where the optimal control problem is presented as a minimisation problem that aims to reduce the costs.

Q-value function for LQR case

A model-based formulation of the Q-value function is formulated to motivate the selection of suitable approximation structure for the Q-value function (32). Inspired by the formulation of the LQR case decribed in [56], this work adapts a Q-value function to the system application and objectives in Paper A and Paper B. The formulation is developed as follows. First, a quadratic cost function, or instant reward is defined

$$\rho(x_k, u_k) = (y_i - y^*)^T Q_1 (y_i - y^*) + u_k^T R u_k, \tag{34}$$

where the first term penalises the deviation of the tank level with respect to a given reference, and the second term penalises high control actions. y^* is a vector with constant reference values, $Q_1 > 0$ and $R > 0$ are weight matrices that penalise the different terms. For simplicity in the notation, this formulation includes only two objectives. The complete formulation with energy cost is provided in Paper B. For the discrete-time LQR, the Q-value function is [56],

$$Q(x_k, u_k) = (y_i - y^*)^T Q_1 (y_i - y^*) + u_k^T R u_k + \gamma V(x_{k+1}), \tag{35}$$

Assuming that there exists a candidate solution to the value function (28), of the form

$$V(x_k) = x_k^T P x_k + G x_k + c, \tag{36}$$

the solution (36) is combined with (32),

$$Q(x_k, u_k) = (y_i - y^*)^T Q_1 (y_i - y^*) + u_k^T R u_k + \gamma (x_{k+1}^T P x_{k+1} + G x_{k+1} + c). \tag{37}$$

Additionally, the augmented system model (11) is introduced

$$\begin{aligned} Q(x_k, u_k) = (C_e x_k + D_e u_k - y^*)^T Q_1 (C_e x_k + D_e u_k - y^*) + u_k^T R u_k \\ + \gamma [(A_e x_k + B_e u_k)^T P (A_e x_k + B_e u_k) + G(A_e x_k + B_e u_k) + c]. \end{aligned} \tag{38}$$

Then, the expression (38) can be reformulated in a matrix form

$$Q(x_k, u_k) = \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} M_{xx} & M_{xu} \\ M_{ux} & M_{uu} \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} N_x \\ N_u \end{bmatrix} + \begin{bmatrix} N_x \\ N_u \end{bmatrix}^T \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \tilde{c}. \tag{39}$$

By expressing (39) in a compact form, with $z_k = [x_k, u_k]^T$,

$$Q(z_k) = z_k^T M z_k + 2 N^T z_k + \tilde{c}, \tag{40}$$

a quadratic, linear and constant components are identified, equivalent to the proposed solution (36). Then, the optimal control policy for (40) can be calculated as

$$u_k^* = \underset{u}{\operatorname{argmin}}\, Q(x_k, u_k) = M_{uu}^{-1} (M_{ux} x_k + N_u). \tag{41}$$

Note that, the resulting control law (41) in is affine, the offset in the control input represents the system regulation around a non-zero equilibrium point [Paper B].

Function approximators

The RL method is originally developed for Markov Decision Process, where the Q-value functions are described by a lookup table. This method is not applicable to continuous systems since infinite-dimensional values cannot be stored in a memory. Consequently, function approximators are proposed to map the Q-value space without storing the individual values. Then, the Q-value function is approximated with a linear parametric approximation that consists of a set of Basis Functions (BFs) $\phi(x, u)$ and a coordinate vector θ [50],

$$\hat{Q}(x_k, u_k) = \phi^T(x_k, u_k)\theta, \tag{42}$$

where $\phi \in \mathbb{R}^{n_b}$ is a column vector and $\theta \in \mathbb{R}^{n_b}$ with the number of bases $n_b = (m_a + n_a + 1)(m_a + n_a)/2$, m_a and n_a represent the number of states and actions respectively [Paper A].
The BFs are chosen according to the system dynamics and reward function, considering the model-based case developed in Section 8.1 as reference. Remark that, although the exact system model is unknown, the underlying behaviour is heuristically described by (1). Therefore, considering the linear structure of the system and the quadratic form of the cost function, polynomial bases are suitable for describing this type of system [56]. The BFs consist of a finite set of second-degree monomial bases that are built as a combination of states and control actions,

$$\phi(x_k, u_k) = [x_{1,k}^2, x_{1,k}x_{2,k}, \ldots, x_{m_a,k}^2, x_{m_a,k}u_k, u_k^2]^T. \tag{43}$$

Considering that the Q-value function approximation is built upon a linear system dynamics and a quadratic reward function structure (40), a state augmentation technique is proposed to incorporate system knowledge into the learning framework strategically. In Paper A, the state-space model includes a reference for tracking and the integral error for rejection of constant disturbance. In Paper B, N harmonics from the Fourier Series (4) are reformulated in a discrete-time form and incorporated into the state vector. In this way, the learning includes partial knowledge of the periodic disturbance.

Least-Square Temporal Difference

Multiple model-free methods for finding an optimal policy are proposed in [49, 50], some of the most important methods differ in the way of collecting the rewards, such as Monte Carlo, Temporal Difference - TD(0) or TD(λ).
The RL method used in this project is based on TD(0) and Q-learning, which evaluates the reward one step ahead and iteratively updates the new Q-values

with the following update law,

$$Q^{new}(x_k, u_k) \leftarrow Q^{old}(x_k, u_k) + \alpha(\overbrace{r_{k+1} + \gamma \min_{u'} Q^{old}(x_{k+1}, u')}^{\text{Q target estimate}} - Q^{old}(x_k, u_k)) \quad (44)$$

$$\text{TD error}$$

where $\alpha \in (0, 1]$ is the learning rate and the difference between the target estimate of the optimal Q-value (updated with the observed data r_{k+1} and x_{k+1}) and the current Q-value is the TD error. This function describes a contraction map that converges to the optimal Q-value when the number of iterations tends to infinity. The convergence of the Q-learning method is proved in [62] for finite Markov Decision Processes under certain conditions such as the all state-action pairs are visited infinitely often and the selection of an adequate learning rate. The learning rate α modifies the number of iterations required to obtain a satisfactory solution. Note that accommodating these conditions in a continuous space-action domain becomes challenging since the number of state-action pairs is infinite.

The update algorithms used in Paper A and Paper B are based on Least-Squares Policy Iteration (LSPI) [103, 104]. This method combines the policy iteration with the data efficiency of Least Squares (LS). The algorithm is implemented as follows. First, the Q-value is formulated for a continuous state-action space by replacing the approximated Q-value (42) into (44),

$$\phi^T(x_k, u_k)\theta_{k+1} = (1 - \alpha)\phi^T(x_k, u_k)\theta_k + \alpha \left[\rho(x_k, u_k) + \gamma \phi^T(x_{k+1}, u'_k)\theta_k\right]. \quad (45)$$

Then, by collecting n_s samples of the BF vector (43), the expression becomes

$$\Phi_l^T \theta_{l+1} = (1 - \alpha)\Phi_l^T \theta_l + \alpha \left[J_l + \gamma {\Phi'}_l^T \theta_l\right], \quad (46)$$

$\Phi_l = [\phi_l, \ldots, \phi_{l+n_s}]$ and $J_l = [\rho_l, \ldots, \rho_{l+n_s}]^T$ are a matrix and a vector respectively and l is the iteration number. The optimal solution for θ_{l+1} is calculated by applying Least Squares

$$\theta_{l+1} = (1 - \alpha)\theta_l + \alpha(\Phi_l \Phi_l^T)^{-1} \Phi_l \left[J_l + \gamma {\Phi'}_l^T \theta_l\right]. \quad (47)$$

This process is repeated until the coordinate vector θ convergence is considered satisfactory.

The polynomial BFs allow solving the expression that defines the optimal policy in closed-form. By computing the root of the polynomials derivative with respect to the control u, a linear control policy is obtained.

$$\pi^*(x_k) = K^*(\theta)x_k = \underset{u}{\operatorname{argmin}}\, \phi(x, u)^T \theta^*. \quad (48)$$

8.2 Results

This section selects the results obtained in Paper B to illustrate the usability and applicability of the control method described in Section 8. The results are collected in a simulation environment and a testbed that emulates a WDN with a single pumping station and an elevated reservoir. The simulation is scaled to the laboratory dimensions, more details of the test and laboratory configuration are given in Paper B and Paper E.
A simulation is performed to adjust the learning hyper-parameters α and γ and the weights in the cost function. The calibration criteria aim firstly to learn an optimal policy for the reference tracking and then the smoothness of the control action and energy usage.
The graph in Figure 23 shows the system variables from simulation results. The simulation is initialised with an arbitrary policy, and during the first 15 days, the tank level falls until the minimum level (top). After day 15, a better policy is learned, and the system gradually approaches the given reference. In Figure 24, the learning variables are shown where a smooth convergence of the linear policy is observed (top) while the difference between $\hat{Q}_{target}$ estimate and $\hat{Q}$, TD error (44), is minimised (bottom). Remark that the tested control solution only regulates safety by penalising the tracking error. Therefore, this objective is favoured. Subsequently, the gain (blue) corresponding to the tracking error stands out with respect to the other gains in the vector.
The parameters learned in a simulation environment are used to initialise the laboratory tests since finding a balance between learning hyper-parameters and cost function weights is laborious and impractical due to the laboratory's physical limitations. The graphs in Figure 25 show the system states in the laboratory tests. In this case, the learning transient is not as noticeable as the simulation. However, the tank level oscillates between the boundary and the reference first. Then the level is regulated around the reference. In Figure 26 the policy slightly adapts to the new environment.

Discussion: Based on the collected results, a brief discussion of the strengths and weaknesses of the presented method is given. The simulation and experimental results show that this control strategy can learn a good policy that regulates the tank filling and rejects the periodic disturbances. However, this method is designed for linear systems. Thus, the convergence of this method relies on the operation around an operating point. This operation is a limitation when deploying this controller in real systems where non-linearities are patent in actuators, hydraulic systems and communication delays. Furthermore, the lack of state constraints in the optimisation compromises the learning of a policy with multiple objectives. In this approach, the reference tracking objective is significantly prioritised with respect to the other objectives since the safety of the operation only relies on following the given

reference. Moreover, this objective must be learned sufficiently fast to avoid operations outside the system limits.

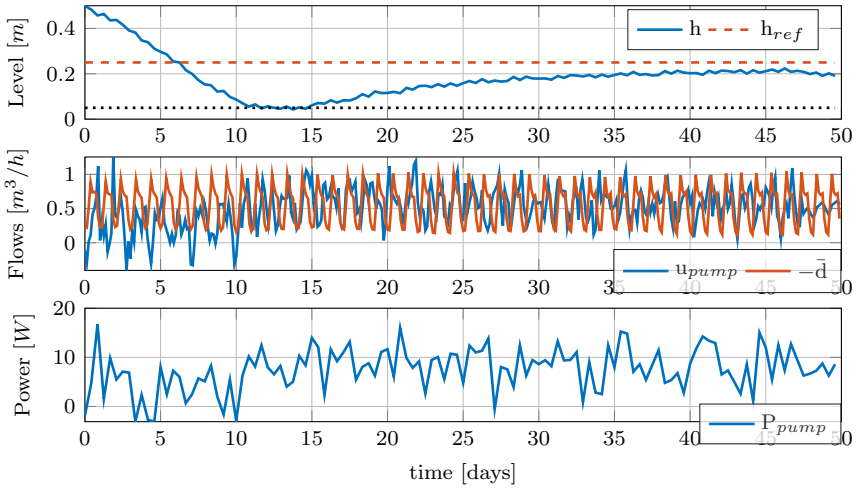


Figure 23: Simulation results. Top: Tank level and level reference Middle: Network flows control input and disturbance Bottom: Pump power consumption [Paper B]

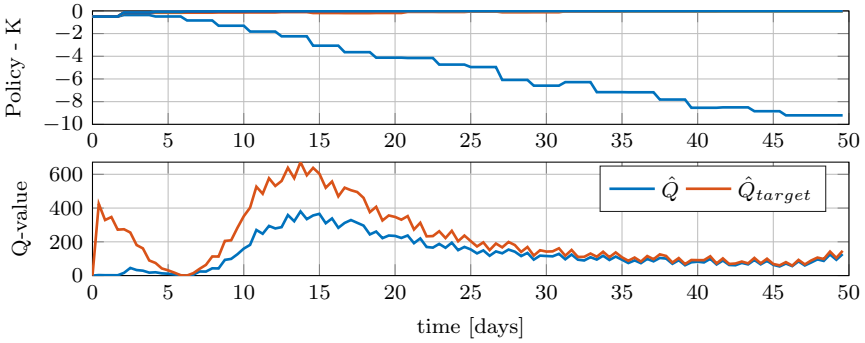


Figure 24: Simulation results. Top: Coordinate vector approximation. Middle: Control policy gain. Bottom: Q-values target and current [Paper B]

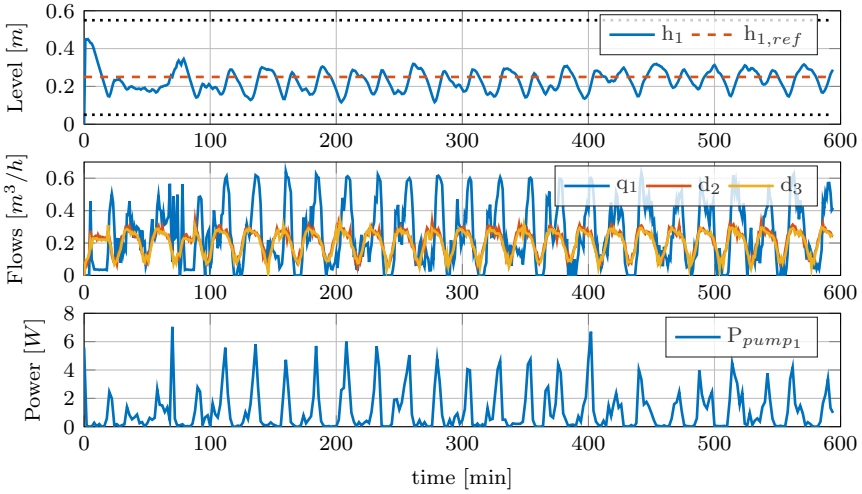


Figure 25: Experimental Results Top: Tank level and level reference Middle: Network flows: control input q_1 and disturbances d_2, d_3. Bottom: Pump power consumption [Paper B]

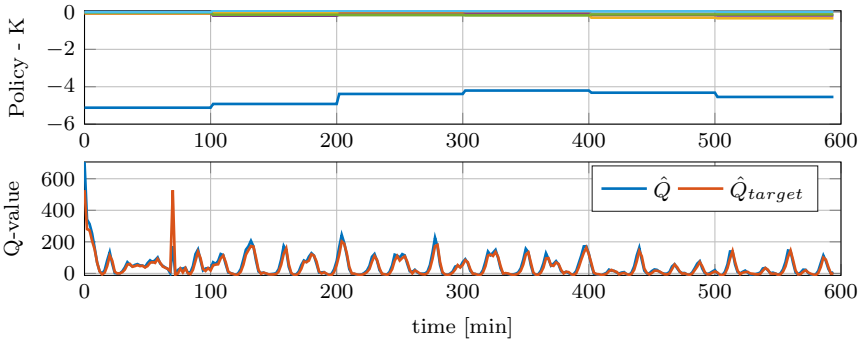


Figure 26: Experimental Results. Top: Control policy gain. Bottom: Q-values target and current [Paper B]

8.3 Robust learning

Ordinary Least Squares (LS) is a simple and efficient estimation method, and its use in combination with Reinforcement Learning algorithms shows satisfactory results. However, the performance is subject to the quality of the data. In Paper A and Paper B, the data collected for the LS estimation is generated by the BFs evaluated with the measured data of states and control action. As previously mentioned, this set is built considering the inclusion of a system structure. However, the design of this set does not consider over-fitting or

under-fitting issues since giving a suitable selection of BF in advance is difficult when their system information is limited. These issues might occur when the BFs approximation structure is not suitable for the identified system, or the data is collected under poor experimental conditions. Paper C presents two algorithms that give numerical robustness when parameter identification issues arise in real-time.

Learning Efficiency analysis

The Fisher information matrix I indicates the amount of information that a batch of data contains about a parametric approximation [69]. Paper C uses the E-optimality criteria, which consist of maximising the eigenvalues of the Fisher matrix to obtain the maximum information from the data.
The minimum eigenvalue of the matrix I is considered an index of the worst-case scenario. This value is low when the collected data has low variations. Consequently, the batch cannot be used for proper parameter estimation. In this work, the rank of the matrix $\Phi_l \Phi_l^T$ sets the threshold for low information I_{low} when it is singular or close to singular.

$$\lambda_{low} = \min \lambda(I_{low}) \tag{49}$$

where λ_{low} is a threshold for indicating low estimation efficiency.
The results in Figure 27 show a simulation where the learning controller proposed in Paper A, for compensating constant disturbances, is combined with the Fisher Information solution proposed in Paper C. After day 5, the system reaches a steady-state, and several state signals become nearly constant. In this same period, Figure 28 (bottom) shows the amount of information drops drastically, leading to ill-conditioning of the data matrices. The proposed solution skips the parameter (policy) update in the algorithm until the amount of collected information is sufficient to perform an update. When a change in the system disturbance is introduced the learning is resumed until it adapts to the new system conditions. The formulation, simulation and Algorithm 5 are further described in Paper C.

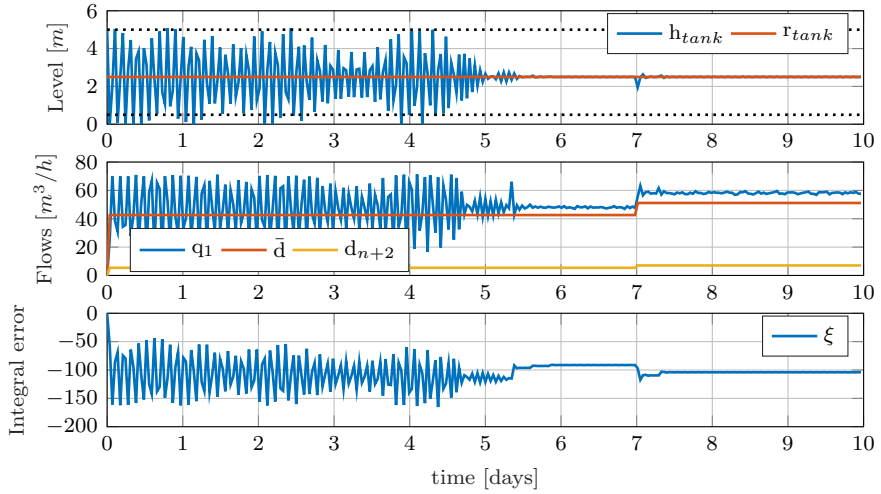


Figure 27: Simulation of a WDN with constant disturbances. Top: shows the tank level and the reference level. Middle: Controlled input flow and demand flows. Bottom: Integral error state [Paper C].

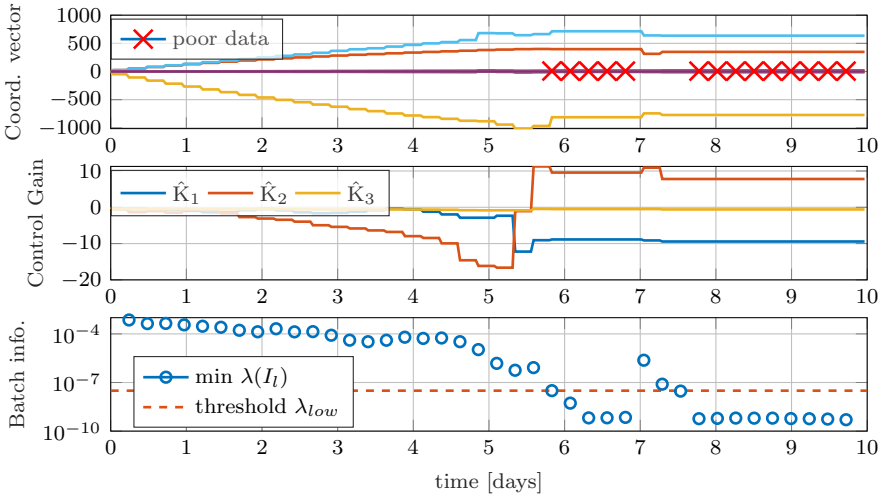


Figure 28: Simulation of a WDN with constant disturbances. Top: Coordinate vector of the Q-value function approximator. Middle: Control policy. Bottom: Batch information measured with the minimum eigenvalue of the Fisher matrix [Paper C].

Singular Value Decomposition and BFs selection

This approach is designed with a similar motivation to the previous, protecting the learning against poor experimental conditions. In this case, the esti-

mation is only performed on the parameters whose associated data contains relevant information. This selective update is executed as follows [Paper C]: first, the collected data is approximated using Singular Value Decomposition (SVD) in its compact form. Therefore, the collected data is divided into three matrices that are sorted by the amount of the system's information

$$\Phi_l = U\Sigma V^T, \tag{50}$$

where $\Phi_l \in \mathbb{R}^{n_b \times n_s}$ where $n_b \leq n_s$ is a matrix with the collected data, with n_b the number of features (BFs) and n_s the number of collected samples, $U \in \mathbb{R}^{n_b \times n_b}$ and $V \in \mathbb{R}^{n_s \times n_b}$ are unitary left and right singular matrices and $\Sigma \in \mathbb{R}^{n_b \times n_b}$ is a diagonal matrix with weights ordered by importance. Having such hierarchic sorting allows the partitioning of each approximation matrix into two parts,

$$U = [\bar{U}, \underline{U}], \quad \Sigma = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \underline{\Sigma} \end{bmatrix}, \quad V^T = \begin{bmatrix} \bar{V}^T \\ \underline{V}^T \end{bmatrix}, \tag{51}$$

where $\bar{U} \in \mathbb{R}^{n_s \times p}$, $\bar{\Sigma} \in \mathbb{R}^{p \times p}$ and $\bar{V} \in \mathbb{R}^{p \times n_s}$. The sub-index notations $(\bar{\cdot})$ and $(\underline{\cdot})$ represent high and low amount of system information respectively. The

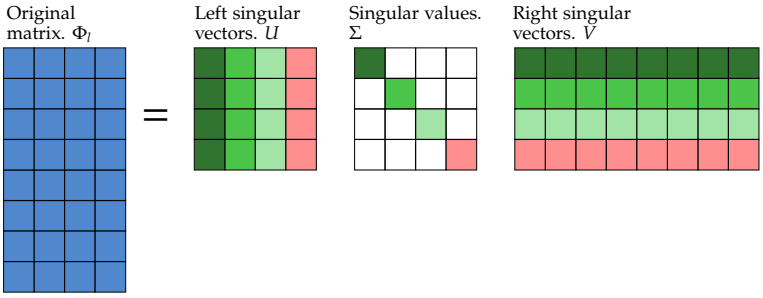


Figure 29: The SVD of the collected data is illustrated where the segregation between high and low amount of information is represented with green and red respectively

partition size p is based on the collected data's rank $p = Rank(\Phi_l \Phi_l^T)$ since a deficient rank might be related to an ill-conditioning of the analysed data. Figure 29 represents the approximation of the collected data with the SVD matrices and the partition according to the quality of the data. Finally, by replacing the SVD approximation into (46), the following linear transformation is deducted.

$$V\Sigma U^T \theta_{l+1} = (1-\alpha) V\Sigma U^T \theta_l + \alpha \Upsilon_l, \tag{52}$$

where $\Upsilon_l = J_l + \gamma \Phi_l'^T \theta_l$. By rearranging (52) with the partitioned matrices (51), the update law is expressed in two SVD sub-spaces,

$$\begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \underline{\Sigma} \end{bmatrix} \begin{bmatrix} \bar{\theta}_{l+1} \\ \underline{\theta}_{l+1} \end{bmatrix} = (1-\alpha) \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \underline{\Sigma} \end{bmatrix} \begin{bmatrix} \bar{\theta}_l \\ \underline{\theta}_l \end{bmatrix} + \alpha \begin{bmatrix} \bar{V}^T \\ \underline{V}^T \end{bmatrix} \Upsilon_l \tag{53}$$

where $U^T\theta = \begin{bmatrix}\bar{\theta} & \underline{\theta}\end{bmatrix}^T$. In this way the parameter identification is computed separately, the upper partition, with $\bar{\theta}$, is updated with standard Temporal Difference (47) while the lower partition, with $\underline{\theta}$, is discarded and its parameters are not updated. Simulation results with Algorithm 6 are presented in Paper C.

9 Safety

> This chapter summarizes the methods used in Paper D and Paper F to provide safety to a learning controller.

One of the main challenges of deploying learning controllers in critical infrastructures, like WDNs, is the uncertainty of their behaviour, especially during the initial learning transients where exploring a broad region of the state-action space is necessary.
This need motivates an important part of this project that addresses the learning controller's safe operation. As briefly introduced in the State-of-the-Art, Section 3, the safety approaches are divided into two, see Figure 14: in one approach, the safety is part of the learning method and changes the optimality criteria, and in the other approach, the safety is part of an external module.
This project uses the second safety structure and designs a reactive method that corrects the system's exploration trajectory when the system is predicted to violate the safety boundaries. The control loop with an external safety filter is represented in Figure 30.

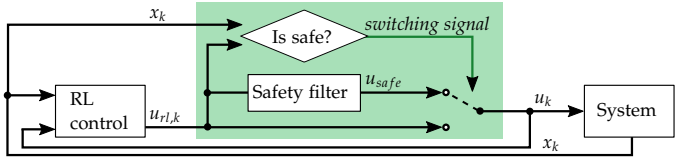


Figure 30: Block diagram of the control architecture where the RL control is connected in series with a policy supervisor (green). The policy supervisor switches the control action based on a 1-step ahead prediction [Paper F].

Safety as constraint satisfaction. A safe set represents an operational zone with low risk of system failure, while an unsafe set represents operation zones where the operation involves a high risk of failure. The constraint set encloses the safe set, but note that the safe set $\mathcal{S}$ and the constraint set $\mathcal{X}$ are not necessarily the same.
In this work, safety is built around constraint satisfaction. This means that the condition for not being safe is to violate the constraint boundaries. When

applied to a deterministic dynamical system, a state is considered safe with respect to the constraints $\mathcal{X}$ if for all $x_0 \in \mathcal{S}$ there is $x_k \in \mathcal{X}$.
Figure 31 illustrates the trajectory of a system around a safe set; starting at time 0 until time k. At time k multiple trajectories are projected depending on the applied control action. The figure shows three different predicted scenarios based on policies A, B and C: in the first scenario, the action u_A keeps the system inside the safe zone. In the second, the action u_B drives the system to a state that does not violate the constraint box, but the predicted state is inside the unsafe set. This scenario means that, at this point, any controller action in the future will drive the system out of the constraint box. For instance, this scenario might be observed in dynamical systems with coupled states like velocity or stochastic disturbances. In the third, the action u_C directly drives the system outside the constraint box.

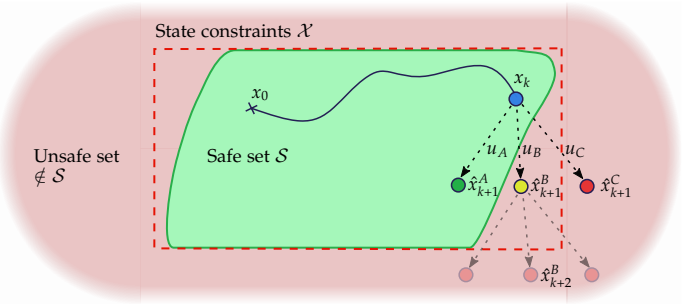


Figure 31: Example of a dynamical system operating inside a safe set from time 0 to k. At time k, three scenarios are contemplated according to the control actions: u_A - safe prediction (green). u_B - unsafe prediction inside constraint box (yellow), subsequently, any the prediction for time $k+2$ drives the system out of the constraint set. u_C - unsafe prediction, outside constraint box (red).

For simplicity in the computation, this work delimits the safe sets around the system constraints since the capacity of the actuator is assumed to be sufficient to rectify an unsafe trajectory 1-step ahead. Then, the tank level state h is considered safe if it belongs to the compact set $\mathcal{H}$ and the pump actuation u is a feasible control if it belongs to the compact set $\mathcal{U}$,

$$h_k \in \mathcal{H} \triangleq \{h_k \in \mathbb{R}^{m_a} | h_{lb} \leq \hat{h}_k \leq h_{ub}\}, \quad \forall k \tag{54a}$$

$$u_k \in \mathcal{U} \triangleq \{u_k \in \mathbb{R}^{n_a} | u_{lb} \leq u_k \leq u_{ub}\}, \quad \forall k, \tag{54b}$$

where the notation $(\cdot_{lb})$ and $(\cdot_{ub})$ define lower and upper bounds respectively. The system prediction and the safe control action are computed with a system model. Two safety methods are developed in this project, each of the methods use a different model: an imperfect deterministic system model in Paper D and a stochastic system model in Paper F.

The overall objective of these safety approaches is to give complete freedom to explore and find optimality inside the safe region and rectify only the unsafe exploration trajectories, hence providing local robustness at the boundaries.

9.1 Nominal model

This safety approach modifies the exploration process as follows. The learning controller independently solves its unconstrained optimisation problem to update the policy $u_{rl} = \hat{K}(\theta_l)x_k$, then, the potential violatation of the boundaries using control action u_{rl} is predicted with the nominal linear model (20). Although the predictions of this model are expected to be imprecise, the trajectory modification provides partial safety to a learning controller policy that otherwise is unable to differentiate between safe and unsafe, see the comparison in Figure 32.

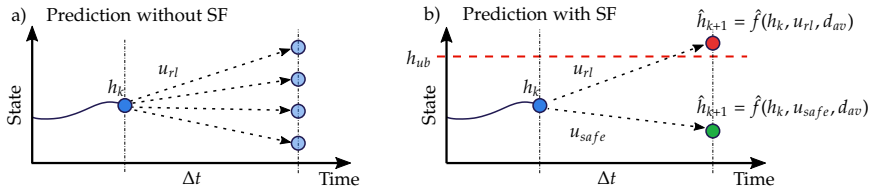


Figure 32: 1-step ahead prediction. a) Illustrates the uncertainty of a learning controller without a Safety Filter (SF). b) Illustrates the detection of potential violations of the boundaries applying u_{rl} and the correction with u_{safe}.

The safety filter works as follows, it first predicts the state with a control input u_{rl}, the potential trajectory of this system is evaluated with respect to the safety boundaries, and, if necessary, modifies the exploration trajectory with u_{safe}. The safe action is the result of the following constrained optimisation problem

$$u_{safe} \in \underset{u}{\operatorname{argmin}} \quad \hat{Q}(x_k, u_k) \tag{55a}$$

$$\text{s.t.} \quad \hat{h}_{k+1} = \hat{f}(h_k, u_k, d_{av}) \tag{55b}$$

$$h_{lb} \leq \hat{h}_{k+1} \leq h_{ub} \tag{55c}$$

$$u_{lb} \leq u_k \leq u_{ub} \tag{55d}$$

The control sequence with safety supervision is described in Algorithm 2.

Algorithm 2 RL with deterministic safety supervision [Paper F].

Input: $\hat{f}(x,u,d)$,γ, α, n_s,
Initialisation: $l \leftarrow 0, x_0, \theta_0$ where $\hat{\pi}(\theta_0)$ must be an admissible policy.
repeat at every iteration k = 0,1,2, ...
    apply u_k and measure x_{k+1}
    $Y_{l_s} \leftarrow \rho(x_k, u_k) + \gamma\hat{Q}(x_{k+1}, \hat{K}_l x_{k+1})$
    if $k = (l+1)n_s$ **then** ▷ Policy update
        $\theta_{l+1} \leftarrow (1-\alpha)\theta_l + \alpha(\Phi_l\Phi_l^T)^{-1}\Phi_l Y_l$
        $\hat{\pi}(\theta_{l+1}, x) \leftarrow \text{argmin}_u\, \phi(x,u)^T\theta_{l+1}$
        $l \leftarrow l+1$
    end if
    if $\hat{h}_{k+1} \in \mathcal{H}$ **then** ▷ Policy supervisor
        $u_k = \hat{K}(\theta_l)x_k + \epsilon_k$ ▷ *RL action*
    else
        $u_k = u_{safe} + \epsilon_k$ ▷ *Safe action*
    end if
until

9.2 Combined model

The nominal model approach shows promising results in correcting unsafe trajectories, see results in Paper D. However, the performance of this supervision depends on the calibration of a nominal model. This approach uses the same control structure with an external safety filter, and the system model is a combination of a nominal model and a Gaussian Process regression (GPR) (21).

Paper F considers the worst-case scenario to evaluate the risk of failure - operating in an unsafe zone; this means that the predicted state $\tilde{h}_{k+1}$ and its corresponding Confidence Interval (CI) must be inside the safe zone. In Figure 33 two scenarios are shown: Scenario *a)* shows the prediction 1-step ahead with the RL controller input u_{rl}, the filter evaluates if the predicted variables (red) and its CI (light red) are inside the safety boundary h_{ub}. Scenario *b)* shows the modification of the predicted trajectory, a safe control input u_{safe} aims to rectify the exploration trajectory (green) and CI (light green) keeping the variable inside the safe area.

When introducing a stochastic element in the system dynamics, a reformulation of the safe-optimal control problem is performed, such that the state constraints are represented as chance constraints. In this work, the chance constraints are expressed with respect to two states, tank level h and turnover

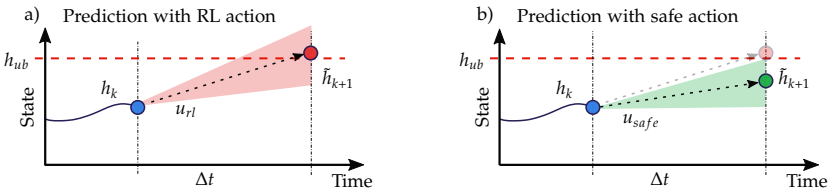


Figure 33: Safety filter process, where the sequence detection-correction is illustrated for state h at time k. a) shows the detection, an unsafe prediction where the control input u_{rl} drives the CI (light red) of $\tilde{h}_{k+1}$ out of the boundary h_{ub}. b) shows the correction, a safe prediction where the control input u_{safe} drives the CI (light green) of $\tilde{h}_{k+1}$ inside the boundary.

τ,

$$Pr\{h \in \mathcal{H}\} \geq p_h, \quad \forall k \tag{56a}$$
$$Pr\{\tau \in \mathcal{T}\} \geq p_\tau, \quad \forall k \tag{56b}$$

where p_h and p_τ are the satisfaction probabilities. Finding an algebraic solution to the problem is difficult when working with chance constraints. Therefore, a transformation of the chance constraint (56) into deterministic equivalents (57d), (57e), and (57f) is performed.

A safe control input u_{safe} is computed by solving an optimisation problem with constraints of the form,

$$u_{safe} \in \underset{u_k}{\text{argmin}} \quad ||u_{rl,k} - u_k||^2_{Q_1} + ||\tau^* - \hat{\tau}_k||^2_{Q_2} \tag{57a}$$
$$\text{s.t.} \quad \tilde{h}_{k+1} = \hat{f}(h_k, u_k) + B_r \mu^r_k(x_k, u_k) \tag{57b}$$
$$\hat{\tau}_k = \hat{g}(q_k) \tag{57c}$$
$$\tilde{h}_{k+1} \geq h_{lb} + K_c \sigma^r(x_k, u_k) \tag{57d}$$
$$\tilde{h}_{k+1} \leq h_{ub} - K_c \sigma^r(x_k, u_k) \tag{57e}$$
$$\hat{\tau}_k \geq \tau_{lb} \tag{57f}$$
$$u_{lb} \leq u_k \leq u_{ub} \tag{57g}$$

where the standard deviation $\sigma^r(z_k) = \sqrt{\Sigma^r}(z_k)$ is computed with the variance model (27c), K_c represents the confidence gain. As previously introduced, the safety filter is a safe-exploration guideline which rectifies the trajectory based on the safe criterion.

The first term in the cost function (57a) penalises differences between the learning control input (free-exploration) and safe control (safe-exploration) to reduce the impact of safe interruptions in the learning. The second term penalises the tracking error between the approximated turnover and a reference. Note that the turnover is also constrained by (57f) that sets the lower level limit. The problem described in (57) is a simplified version of the problem presented in Paper F. The complete problem includes slack variables that

ensure the feasibility of the control problem.
The filter's behaviour can be modified with the confidence gain K_c; a high gain provides conservative supervision while a low gain relaxes the safety criterium. The selection of suitable value for K_c comprises a balance between the application requirements and the constrained optimisation problem.
Figure 34 shows two examples of normal distributions, where the area of the CI is adjusted around its mean value μ^r with the confidence gain K_c.

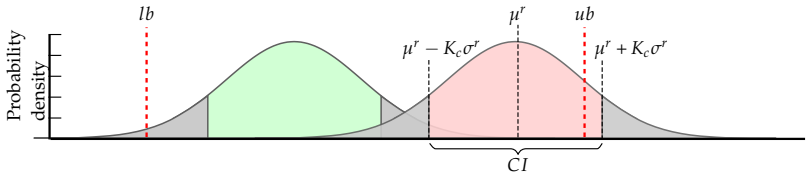


Figure 34: Scheme of two examples of normal distributions where the coloured area represents the CI of the residual function, red represents that the interval violates the safety boundary (red dash-lines *lb* or *ub*), green represents a safe CI inside the boundaries, the grey tails are neglected in the safety evaluation, the size of these areas is adjusted with K_c.

Algorithm 2 is executed in real-time with the combined safety filter, and in `line 11`, the prediction safety is assessed with the following criteria,

$$h_{lb} + K_c\sigma^r(z'_k) \leq \tilde{h}_{k+1}(z'_k) \leq h_{ub} - K_c\sigma^r(z'_k) \tag{58a}$$

$$\tau_{lb} \leq \hat{\tau}(\tilde{h}_{k+1}), \tag{58b}$$

where the observed input vector z'_k is built with the augmented system states and the RL control action, $z'_k = [x_k^T, u_{rl,k}^T]^T$. Remark that the mean $\mu_k^r(z_k)$ and variance function $\Sigma_k^r(z_k)$ include the decision variable u_k in its input vector z_k. Additionally, the standard deviation σ^r is time-variant since the regression model is trained online. This makes the optimisation problem (57) harder to solve and the selection of a suitable K_c difficult. The following condition is stated for facilitating the computation

$$|K_c\sigma^r(z_k)| \leq |h_{ub} - h_{lb}|/2, \tag{59}$$

The condition (59) limits the width of the CI based on the distance between bounds. If the condition is not met, the method uses a *naive* prediction which neglects the CI [105] by setting $K_c = 0$.
Similarly as Figure 33 illustrates, the graphs in Figure 35 show several scenarios where the safety filter is intermittently active for correcting the system trajectory. The top graph shows different states: the blue dots represent real level measurements. The free-exploration prediction is represented in red $\tilde{h}_{k+1}^{rl}$ with its corresponding CI^{rl}, they are computed at time k with the learning control input u_{rl}, this signal is used for detecting unsafe trajectories. The

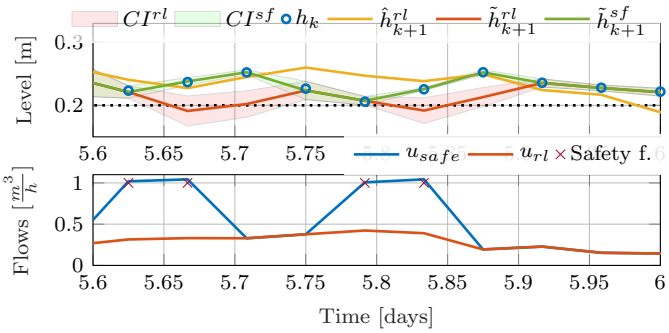


Figure 35: Simulation of the WDN with a RL controller and a combined filter. (Top) Blue dots represent the real level measurement and the dotted line represents the lower boundary. The nominal model signal $\hat{h}^{rl}_{k+1} = \hat{f}(h_k, u_{rl})$ and the combined model $\tilde{h}^{rl}_{k+1} = \hat{f}(h_k, u_{rl}) + B_r\hat{r}(z'_k)$ are 1-step ahead predictions, and $\tilde{h}^{sf}_{k+1} = \hat{f}(h_k, u_k) + B_r\hat{r}(z_k)$ is the safe prediction. (Bottom) The RL control input u_{rl} and the control input for safe exploration u_{safe} [Paper F].

safe-exploration prediction is represented in green $\tilde{h}^{sf}_{k+1}$ with its corresponding CI^{sf}, they are computed with the safe control input u_{safe}. Consequently, both signals are aligned when safe trajectories are predicted. The bottom graph shows the control signals and how the safety filter corrects the control action u_{rl} when the prediction crosses the boundaries. Furthermore, the nominal-exploration prediction, in yellow, $\hat{h}^{rl}_{k+1}$ shows a clear deviation with respect to $\tilde{h}^{rl}_{k+1}$, this is due to the poor calibration of the nominal model. The confidence in the combined model is increased since the GP regression model is trained online. Consequently, the graph shows that CI is gradually narrowing.

9.3 Results

This section selects the results obtained in Paper F to describe the usability and applicability of the stochastic safety filter technique in combination with the learning controller, described in Section 9 and Section 8 respectively. The results are collected in a testbed that emulates a WDN with two pressure zones, an elevated reservoir and a single pumping station. A brief description of this testbed configuration is given in Section 6 - Figure 18, the details are given in Paper F.

The experimental results are initialised with an arbitrary policy in the RL controller, and the safety filter uses a combined model. The nominal model is naively calibrated with a guess of the tank dimensions and average demand, and the kernel of the GP regression is initially built with random data.

In this test, two algorithms are running in real-time, Algorithm 2 that updates the controller policy with safety guidelines and Algorithm 1 that updates the

GP model.
In the graph Figure 36, the learning transients of both algorithms are compared; the top graph shows the convergence of the RL policy after hour 8, while the learning of the GP model, in the bottom, is nearly settled after hour 3. The learning of the GP model is also observed in Figure 37, during the first 3 hours, the CI indicates a lack of confidence in the GP prediction. This model causes a conservative behaviour of the safety filter that is frequently active since the safe set defined by the constraints is reduced accordingly. After 3 hours, the predictions are improved, and the system trajectory modifications occur accurately near the boundaries; the filter actions do not interrupt the learning of the controller policy. Note that, at this point, the CI is unnoticeable in the graph.
The objective of the safety filter is to assist the exploration of the learning controller. After hour 6, the top graph in Figure 37 shows a gradual regulation of the tank level towards the given reference. Thereafter, the tank level always operates within the safe boundaries, and the safety filter is only activated to correct the turnover when the prediction falls below the lower boundary. Furthermore, a scaling error is observed between the approximated and actual turnover signal $\hat{\tau}$ and τ. However, this error is considered a minor safety issue since the approximated signal represents a worst-case scenario from a model calibration.

Discussion: Based on the collected results, a brief discussion of the strengths and weaknesses of the presented method is given.
This safety approach solves one of the main challenges of learning controllers, which is safe exploration. Solving this challenge is particularly important when the learning controller is deployed in industrial applications like water infrastructures, in which robust and continuous operation is essential. Recall that in the previous RL approaches shown in Section 8 the safety relies on learning a policy that follows a reference before crossing the safety boundaries.
The experimental results show the importance of the safety filter in the learning transient. In this case, the safe exploration is externally assisted by the safety filter. This fact relaxes the tedious calibration of the learning hyperparameters and the cost function, thus facilitating the implementation of the control solution in different study cases.
The confidence interval is directly computed from the variance of the GP regression; it has a key role in early detecting potential unsafe trajectories. However, the importance of the variance in the filtering must be adapted to the application requirements. For instance, in this work, the confidence gain K_c is mainly used to limit the magnitude of the CI. Large confidence intervals increase the conservativeness of the control actions, thus limiting the explo-

ration and compromising the learning of an optimal policy. Furthermore, the variance $\sigma(x, u)$ depends on the decision variable, this increases the complexity of the safe-optimal control problem (57). The condition (59) is introduced in the algorithm to facilitate the computation of safe control input.
A control structure with a safety filter allows the inclusion of application objectives like water quality that otherwise cannot be represented by the proposed learning architecture. However, the learning controller and the filter are independent blocks that could have conflicting objectives. Therefore, the formulation of these objectives must be balanced to avoid the safety filter constantly overruling the main policy. The results show that learning an optimal policy while using a safety filter is possible. However, the safety actuations must include persistent variations to avoid that, for instance, a saturated safe action compromises the identification of a Q-value function. A more detailed discussion of the results is given in Paper F.

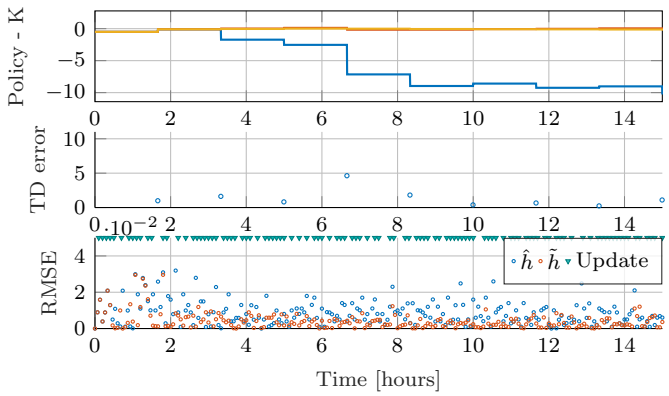


Figure 36: Experimental results with RL control and combined safety filter. Top: Transient of the learned policy. Middle: The temporal difference computed as $|\hat{K}_l - \hat{K}_{l+1}|$. Bottom: The Root-Mean Square Error (RMSE) of the residual generated by the estimated $\hat{h}$ and $\tilde{h}$ compared with the measured level h. The triangles show the GP model updates during the execution of Algorithm 9 [Paper F].

10 Concluding remarks

The work presented in this thesis addresses the development of a model-free controller for a water distribution network with an elevated reservoir. This work is motivated by the need to implement optimal pressure management in small-medium-sized water utilities that cannot afford the commissioning cost of implementing advanced control solutions requiring initial model calibration and continuous maintenance.

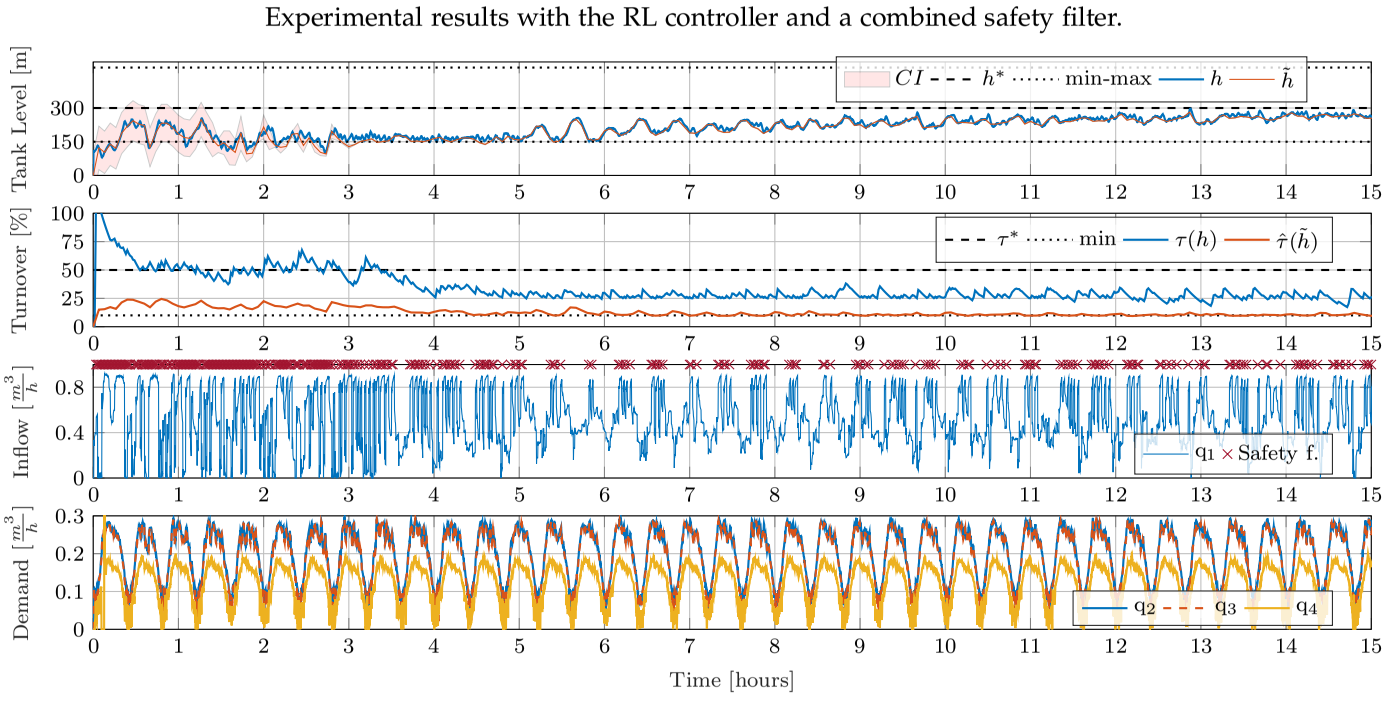


Figure 37: Tank level regulation. Daily turnover of the tank, real signal τ and its approximation $\hat{\tau}$. Inflow of the pumping station q_1 where the red marks point the actuations that are corrected by the safety filter. Demands from PZ1 (q_2 and q_3) and PZ2 (q_4) [Paper F].

The details of the project motivation and summary of the existing control solutions are given in Part I, and the results of this work are presented as a collection of papers enclosed in Part II. First, a conclusion of the project is presented in Section 10.1, this section also summarises the author's reflections on learning controllers in industrial applications. Then, recommendations for future work are presented in Section 10.2.

10.1 Conclusions

The conclusions of this project analyse the stated research objectives and contributions. Accordingly, the analysis is divided into three categories:

Model-free controller. The first objective states the development of an adaptive-optimal controller that gradually learns an optimal management policy despite the system disturbances. The controller design is addressed in Section 8 and by the contributions Paper A and Paper B. In this work, a Reinforcement Learning controller for linear systems is formulated. The function approximator of the Q-value function is constructed with 2^{nd} degree polynomials to reproduce the structure of the linear system and a quadratic cost function. These works present how a state-space augmentation technique is utilised to introduce additional system knowledge into a linear system, for example, disturbance information. The results show that the algorithm satisfactorily converges in a simulation environment where the dynamics are linear and in a test-bed where some non-linearities are introduced in the control loop. However, in the latter case, the operation is restricted to a fixed operating point to reach an optimal solution.
Moreover, in Paper C, the learning performance, which is highly sensitive to the quality of the data, is analysed. This work proposes two approaches to analyse and robustify the learning of the Q-value function in real-time: The first solution uses the Fisher information matrix to measure the data quality and pause the parameter update strategically. The second solution uses an SVD approximation to perform a parameter update only on relevant features. The results show that both approaches help cope with scenarios with poor experimental conditions in real-time.

Safe operation. As described in the introduction, Water Distribution Networks are critical infrastructures, and their operation is essential for society and the economy. This requirement poses safety as one of the backbones when developing control solutions for these infrastructures. The results in contributions Paper A and Paper B show the need for a safety guideline when deploying learning controllers. The operation during the learning transient repeatedly crosses the safety boundaries until an optimal policy is learned.

The penalisation of the tracking error in the cost function can partially mitigate this issue. However, the controller reaction time is subject to a policy update which is typically slower than the system dynamics. Providing safety by soft constraints might also conflict with other objectives, and the calibration of the cost function might become tedious when balancing multiple objectives.

This work proposes two safety methods for maintaining a robust and continuous operation at all times, even during exploration transients. A summary of the safety method is presented in Section 9 and in the contributions Paper D and Paper F.

This method does not modify the optimality criterium, and it consists of a safety filter that is inserted between the RL controller and the system plant. The filter overrides the control input provided by the RL controller in case of potential unsafe predictions. In Paper D, the prediction and safety rectification are based on a nominal linear model, and the results show that the performance of this method is subject to the structure of the nominal model and its calibration. To reduce the dependency of this model, in Paper F the proposed model is a combination of a nominal model with a GP regression, allowing for nominal model imperfections. The results show that the GP model can successfully capture the system uncertainty. The confidence intervals are utilised to safely guide the learning by first providing conservative supervision when the uncertainty is high and then relaxing the supervision when the GP performs accurate predictions. There must be a trade-off between conservative and reckless operation, such that the controller actions allow the collection of rich data while the system trajectories are safe-critical. The confidence gain, K_c, and the learning rate of the GP are factors that regulate the balance between conservative and relaxed supervision.

This approach consists of a modular control architecture; in this project, the modularity is a clear advantage for the easy implementation of a supervisor module that leverages the safety of an existing learning controller. The legacy controller and the safety filter are independent modules in this control loop. The operation of the main controller is only interrupted in case of safety. Therefore, this kind of control architecture provides the opportunity to simplify the computation of optimal controllers that require solving constrained optimisation problems in real-time. This simplification consists of delegating some constraints to the safety filter, the conservativism of the main policy is reduced as well as its computational effort since the expensive computation is only triggered when potential unsafe trajectories are predicted.

This practice is utilised to formulate the water quality as a safety problem in Paper F. The results show that the safety filter can also limit the water age in a tank based on the average turnover. In this way, the pump actuation for increasing the turnover is only triggered when necessary for safety.

Experimental validation. This project also includes the construction of a laboratory facility that emulates, on a small scale, the behaviour of water infrastructures. The laboratory is a modular system that allows for the reproduction of different topologies and scenarios, a brief comparison of the laboratory and real water distribution systems is given in Section 2, and a detailed description of the laboratory is presented in Paper E. This laboratory has supported this work by validating the usability and robustness of the solutions proposed in the contributions Paper A, Paper D and Paper F. By testing the proposed solutions in this laboratory facility, different factors that impact the control loop are introduced, such as communication delays, process noise, actuator dynamics or non-linearities of the hydraulic system. All these factors are typically difficult to include in a simulation environment. Therefore, the experimental results are especially necessary for identifying weaknesses in each method and, in most cases, pointing to the next step for future development.
Finally, another issue to discuss is the benchmarking problem. This work considers that comparing the performance of a learning control with a model-based control approach is challenging since the lack of prior system information or maintenance cost of the solution are subjective factors that differ between each case. Moreover, performing a fair comparison between different model-free approaches is complicated in industrial applications since the learning performance depends on multiple factors such as manual calibration hyper-parameters for a specific application domain.

Personal outlook. As previously introduced, the overall objective of this work is to facilitate the implementation of optimal pressure management solutions on a broad number of utilities in a cost-efficient manner. Therefore, the design criteria prioritise tractable solutions that are easy to interpret versus complex solutions that might provide better performance but require qualified personnel for operation.
To design control solutions closer to real utility needs, the author conducted interviews with two experts in the field. These interviews have supported the author in prioritising the list of management objectives and in developing system models that describe realistic scenarios.
The first interview [106] focused on the current management strategies for regulating water quality and the use of elevated reservoirs in small size utilities. From the discussion, the author concludes that:

- In Denmark, a majority of water utilities use groundwater sources where the water quality is sufficient. Hence, its distribution is chlorine-free. The water quality degradation depends on factors that cannot be controlled during the distribution like the quality of the source or water temperature. Nevertheless, the ambient temperature is moderate

most time of the year. Hence, this factor is neglected when developing dynamic models that describe the system's behaviour. Additionally, the supply is continuous, with a low risk of pipe stagnation. Thus, transportation has a minor impact on the deterioration of water quality. Therefore, the most critical elements in a distribution network are reservoirs where the water is stored over a while.

- There is no active regulation of the water quality in the tank and the water age is regulated by periodic flushing of the tank. The frequency of this flushing cycle is manually scheduled.

The second interview [107] focused on the business perspective of learning controllers in the industry, as well as the automatisation of certain operator tasks at water utilities. From the discussion, the author concludes that:

- Some water utilities can bear the initial expense of new technology. Moreover, they might not demand a short-time investment payback compared to other industries. However, the personnel cost related to the maintenance of some advanced control solutions might obstacle their implementation since small-medium-size water utilities might lack qualified personnel to operate them. In this scenario, learning controllers could play an important role in the industry, considering that a long training period is not an issue as long as a safe water supply is guaranteed.

- Technology that involves learning or adaptation to a specific water network could reduce or simplify some manual operator tasks. Having many manual operations creates a dependency on knowledgeable and experienced personnel for managing critical infrastructures. This issue highlights the need for digitalised infrastructures that permit a transition to automated management. Hence, generational knowledge can be gradually replaced by intelligent controllers.

The use of elevated reservoirs in water distribution networks differs depending on the location and size of the network or governmental rules [108]. As briefly described in Section 2, the main function of elevated reservoirs is to balance the pressure in the network and support the pumping stations during peaks in demand. Recently, some reservoirs have been decommissioned due to the infrastructure's age, and smart pumping systems are replacing their functionality. However, elevated reservoirs are still necessary since they are passive control elements that provide management alternatives for balancing the pressure in the network. For instance, in some cases, like in New York, buildings of 6-stories or more are required to install a rooftop water tank to balance the pressure or fire fighting locally.
Based on the interviews, discussions and project results, the author concludes

that safety considerations are essential when designing any control solution for critical infrastructures.
In conclusion, the design of new control solutions (smart pumping systems) must exploit all the network elements, including elevated reservoirs, first to guarantee a safe water supply and then to improve the overall system efficiency. Despite their uncertain behaviour, learning controllers have a great potential to solve issues that challenge water infrastructures. However, implementing some model-free solutions can become as costly as other model-based approaches since some model-free controllers have to be configured in advance to learn a specific application.

10.2 Future work

Pure model-free approaches like Deep Reinforcement Learning are black-box methods capable of learning complex problems, but they are known for their lack of interpretability. Nevertheless, having a simple learning structure comes at a price. The design of this project's controller aims at increasing the interpretability of the resulting policy by providing a model structure. In this way, the tuning of the RL hyper-parameters is facilitated. The project results show that the performance of this RL controller is limited by the polynomial basis when applied to a real WDN. These bases cannot describe the non-linear behaviour of the system. Some non-linear effects could be captured by increasing the approximation vector space and selecting higher degree polynomials. However, this solution could easily lead to numerical issues. Orthogonal polynomials like Chebyshev polynomials can provide a better-conditioned parameter identification with respect to plain polynomials [63]. A potential extension of this controller toward non-linear applications could be studied by using radial bases. This kind of basis also allows the formulation of a reward (cost) function that is not restricted to a quadratic form.
Including safety in this work has been essential for easily deploying the learning controller, especially when tested in the laboratory test-bed. The learning algorithm freely explores the safe area, and the system trajectory is modified otherwise. Paper F extends the safety filter and includes confidence intervals for assisting the exploration based on the accuracy of the predicted GP model. In this approach, the GP model is naively trained with the same data as the RL algorithm. In the future, a better selection of the input data could be required. Alternatively, the author highlights two methods that have the potential to solve this control problem in water distribution networks:

- Learning-based predictive controllers: MPC is a well-known optimal control method where safety is implicit in its constrained optimisation framework. [96] combines a GP regression model in an MPC to deliver a data-driven MPC, [93] proposes a control scheme that combines Eco-

nomic Non-Linear MPC with RL to generate an optimal policy despite the underlying system model, by adjusting the stage and terminal cost alone. However, the control structures of these two approaches differ from the control approach proposed in this study.

- Barrier functions: The safety of the proposed control solution can be increased by modifying the RL optimality criteria such that the unsafe operations are learned and penalised accordingly. This extension does not necessarily imply the modification of the proposed control structure. An equilibrium between optimality and safety can be achieved by introducing barrier functions in the cost function. However, the resulting policy is expected to be more conservative. Extending this project approach with barrier functions would also imply a selection of more suitable bases for the Q-value function approximation. The use of non-linear optimisation frameworks or new bases might reduce the tractability of the solution. On the other hand, they facilitate the inclusion of management objectives in the optimality criteria, such as water quality or energy consumption.

References

[1] "The united nations world water development report 2018 : nature-based solutions for water," p. 139 p. :, 2018.

[2] F. Gassert, P. Reig, T. Luo, and A. Maddocks, "A weighted aggregation of spatially distinct hydrological indicators," *World Resources Institute: Washington, DC*, 2013.

[3] FAO, "Aquastat database. food and agriculture organization of the united nations (fao)," 2013.

[4] "Our world in data,." OurWorldInData.org/water-access-resources-sanitation/. Accessed: 2022-03-01.

[5] K. Kitamori, T. Manders, R. Dellink, and A. Tabeau, "Oecd environmental outlook to 2050: the consequences of inaction," tech. rep., OECD, 2012.

[6] M. Fontozzi, A. Lambert, C. Kallesøe, A.-S. Hassan, D. Stærk, S. Lieknins Neve, and M. Riis, "Whitepaper. intelligent pressure management," tech. rep., Grundfos, 2012.

[7] R. Frauendorfer and R. Liemberger, *The issues and challenges of reducing non-revenue water*. Asian Development Bank, 2010.

[8] P. Marin, "The challenge of reducing non-revenue water in developing countries—how the private sector can help: A look at performance-based service

contracting," *World Bank Water Supply and Sanitation Sector Board Discussion Paper Series*, no. 8, 2006.

[9] S. Hamilton and R. McKenzie, *Water management and water loss*. IWA Publishing, 2014.

[10] J. Horne, J. Turgeon, and E. Boyus, "Energy self-assessment tools and energy audits for water and wastewater utilities," *Webinar. Washington, DC: US Environmental Protection Agency*, 2014.

[11] L. Reekie and L. et al Reekie, *Electricity use and management in the municipal water supply and wastewater industries*. Water Research Foundation, 2013.

[12] J. Thornton and A. Lambert, "Progress in practical prediction of pressure: leakage, pressure: burst frequency and pressure: consumption relationships," in *Proceedings of IWA Special Conference'Leakage*, pp. 12–14, 2005.

[13] "Drinking water directive: Eureau position," tech. rep., EurEau, 2018.

[14] P. K. Swamee and A. K. Sharma, *Design of Water Supply Pipe Networks*. Wiley Interscience, 2008.

[15] "Nyeri Water & Sanitation Company." http://www.nyewasco.co.ke/. Accessed: 2022-03-01.

[16] "Photo by jacek dylag on unsplash." https://unsplash.com/photos/Vve7XkiUq_Y. Accessed: 2022-03-01.

[17] V. Puig, C. Ocampo-Martínez, R. Pérez, G. Cembrano, J. Quevedo, and T. Escobet, *Real-time monitoring and operational control of drinking-water systems*. Springer, 2017.

[18] H. Monsef, M. Naghashzadegan, R. Farmani, and A. Jamali, "Pressure management in water distribution systems in order to reduce energy consumption and background leakage," *Journal of Water Supply: Research and Technology—AQUA*, vol. 67, no. 4, pp. 397–403, 2018.

[19] J. Thornton and A. Lambert, "Pressure management extends infrastructure life and reduces unnecessary energy costs," in *IWA Conference'Water Loss*, 2007.

[20] A. Lambert, M. Fantozzi, and J. Thornton, "Practical approaches to modeling leakage and pressure management in distribution systems–progress since 2005," in *Proceedings of the 12th Int. Conference on Computing and Control for the Water Industry-CCWI2013*, 2013.

[21] A. Lambert, S. Trow, C. Merks, B. Charalambous, A. Donnelly, S. Galea, M. Fantozzi, A. Hulsmann, J. Koelbl, J. Kovac, *et al.*, "EU reference document good practices on leakage management WFD CIS WG PoM," *European Commission: Brussels, Belgium*, 2015.

[22] A. Lambert and M. Fantozzi, "Recent developments in pressure management," in *IWA Conference" Water loss 2010*, 2010.

[23] C. Ocampo-Martínez, V. Puig, G. Cembrano, R. Creus, and M. Minoves, "Improving water management efficiency by using optimization-based control strategies: the barcelona case study," *Water science and technology: water supply*, vol. 9, no. 5, pp. 565–575, 2009.

[24] "Schneider electric, water & waste water." https://www.se.com/ww/en/work/solutions/for-business/water/. Accessed: 2022-03-01.

[25] "Takadu." https://www.takadu.com/. Accessed: 2022-03-01.

[26] "Visenti, a xilem brand." https://www.xylem.com/en-nz/brands/visenti/industries--applications/. Accessed: 2022-03-01.

[27] "DHI group,." https://www.dhigroup.com/. Accessed: 2022-03-01.

[28] "Krüger, veolia water technology." https://www.kruger.dk/veolia-water-technologies. Accessed: 2022-03-01.

[29] "Envidan,." https://www.envidan.dk/. Accessed: 2022-03-01.

[30] "Rockwell automation, pavilion 8." https://literature.rockwellautomation.com/idc/groups/literature/documents/br/rsbrp8-br001_-en-p.pdf. Accessed: 2022-03-01.

[31] "Smart water solutions, by xylem." https://www.xylem.com/siteassets/industries--applications/resources/xylem_smart_water_white_paper_low.pdf. Accessed: 2022-03-01.

[32] "Demand driven distribution, pressure management whitepaper." https://www.grundfos.com. Accessed: 2022-03-01.

[33] "I2o water,." https://en.i2owater.com/solutions/advanced-pressure-management/#devices. Accessed: 2022-03-01.

[34] M. Kazantzis, A. Simpson, D. Kwong, and S. Tan, "A new methodology for optimizing the daily operations of a pumping plant," in *Proceedings of 2002 Conference on Water Resources Planning, Roanoke, USA. ASCE*, 2002.

[35] A. Marchi, A. R. Simpson, and M. F. Lambert, "Pump operation optimization using rule-based controls," *Procedia Engineering*, vol. 186, pp. 210–217, 2017.

[36] C. Biscos, M. Mulholland, M. Le Lann, C. Buckley, and C. Brouckaert, "Optimal operation of water distribution networks by predictive control using minlp," *Water Sa*, vol. 29, no. 4, pp. 393–404, 2003.

[37] C. L. Celi, P. L. Iglesias-Rey, and F. M. Solano, "Energy optimization of supplied flows from multiple pumping stations in water distributions networks," *Procedia Engineering*, vol. 186, pp. 93–100, 2017.

[38] J. Pascual, J. Romera, V. Puig, G. Cembrano, R. Creus, and M. Minoves, "Operational predictive optimal control of barcelona water transport network," *Control Engineering Practice*, vol. 21, no. 8, pp. 1020–1034, 2013.

[39] C. Ocampo-Martinez, V. Puig, G. Cembrano, and J. Quevedo, "Application of predictive control strategies to the management of complex networks in the urban water cycle [applications of control]," *IEEE Control Systems Magazine*, vol. 33, no. 1, pp. 15–41, 2013.

[40] S. Leirens, C. Zamora, R. Negenborn, and B. De Schutter, "Coordination in urban water supply networks using distributed model predictive control," in *Proceedings of the 2010 American Control Conference*, pp. 3957–3962, IEEE, 2010.

[41] M. Ellis, J. Liu, and P. D. Christofides, "Economic model predictive control,"

[42] A. Cimiński, "Optimized robust model predictive control—application to drinking water distribution systems hydraulics," pp. 395–400, 07 2010.

[43] J. M. Grosso, P. Velarde, C. Ocampo-Martinez, J. M. Maestre, and V. Puig, "Stochastic model predictive control approaches applied to drinking water networks," *Optimal Control Applications and Methods*, vol. 38, no. 4, pp. 541–558, 2017.

[44] T. N. Jensen, C. S. Kallesøe, J. D. Bendtsen, and R. Wisniewski, "Iterative learning pressure control in water distribution networks," in *2018 IEEE Conference on Control Technology and Applications (CCTA)*, pp. 583–588, IEEE, 2018.

[45] E. Ertin, A. N. Dean, M. L. Moore, and K. L. Priddy, "Dynamic optimization for optimal control of water distribution systems," in *Applications and Science of Computational Intelligence IV*, vol. 4390, pp. 142–149, SPIE, 2001.

[46] C. Kallesøe, T. Jensen, and J. Bendtsen, "Plug-and-play model predictive control for water supply networks with storage," vol. "50", pp. "6582–6587", "Elsevier", 2017.

[47] K. B. Ariyur and M. Krstic, *Real-time optimization by extremum-seeking control*. John Wiley & Sons, 2003.

[48] D. A. Bristow, M. Tharayil, and A. G. Alleyne, "A survey of iterative learning control," *IEEE control systems magazine*, vol. 26, no. 3, pp. 96–114, 2006.

[49] D. Bertsekas, *Dynamic programming and optimal control: Volume I*, vol. 1. Athena scientific, 2012.

[50] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[51] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[52] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[53] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Proceedings of The 2nd Conference on Robot Learning* (A. Billard, A. Dragan, J. Peters, and J. Morimoto, eds.), vol. 87 of *Proceedings of Machine Learning Research*, pp. 651–673, PMLR, 29–31 Oct 2018.

[54] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396, IEEE, 2017.

[55] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE circuits and systems magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[56] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[57] A. Castelletti, S. Galelli, M. Restelli, and R. Soncini-Sessa, "Tree-based reinforcement learning for optimal water reservoir operation," *Water Resources Research*, vol. 46, no. 9, 2010.

[58] V. Javalera, B. Morcego, and V. Puig, "Negotiation and learning in distributed mpc of large scale systems," in *Proceedings of the 2010 American Control Conference*, pp. 3168–3173, IEEE, 2010.

[59] M. Mahootchi, H. R. Tizhoosh, and K. P. Ponnambalam, "Reservoir operation optimization by reinforcement learning," *Journal of Water Management Modeling*, 2007.

[60] D. Ochoa, G. Riano-Briceno, N. Quijano, and C. Ocampo-Martinez, "Control of urban drainage systems: Optimal flow control and deep learning in action," in *2019 American Control Conference (ACC)*, pp. 4826–4831, IEEE, 2019.

[61] R. Taormina and S. Galelli, "Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems," *Journal of Water Resources Planning and Management*, vol. 144, no. 10, p. 04018065, 2018.

[62] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[63] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement learning and dynamic programming using function approximators*. CRC press, 2017.

[64] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[65] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[66] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, 2014.

[67] Q.-Y. Fan and G.-H. Yang, "Adaptive actor–critic design-based integral sliding-mode control for partially unknown nonlinear systems with input disturbances," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 1, pp. 165–177, 2016.

[68] S. Xue, B. Luo, D. Liu, and Y. Yang, "Constrained event-triggered $\mathcal{H}_\infty$ control based on adaptive dynamic programming with concurrent learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 1, pp. 357–369, 2022.

[69] T. Söderström and P. Stoica, "System identification," 1989.

[70] P. W. Keller, S. Mannor, and D. Precup, "Automatic basis function construction for approximate dynamic programming and reinforcement learning," in *Proceedings of the 23rd international conference on Machine learning*, pp. 449–456, 2006.

[71] H. Hachiya and M. Sugiyama, "Feature selection for reinforcement learning: Evaluating implicit state-reward dependency via conditional mutual information," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 474–489, Springer, 2010.

[72] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[73] B. Behzadian, *Feature selection by singular value decomposition for reinforcement learning*. PhD thesis, University of New Hampshire, 2019.

[74] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Transactions on neural networks and learning systems*, vol. 24, no. 10, pp. 1513–1525, 2013.

[75] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[76] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[77] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[78] R. Setola, E. Luiijf, and M. Theocharidou, "Critical infrastructures, protection and resilience," in *Managing the complexity of critical infrastructures*, pp. 1–18, Springer, Cham, 2016.

[79] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[80] S. A. A. Rizvi and Z. Lin, "Adaptive dynamic programming for model-free global stabilization of control constrained continuous-time systems," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 1048–1060, 2022.

[81] Y. Yang, Y. Yin, W. He, K. G. Vamvoudakis, H. Modares, and D. C. Wunsch, "Safety-aware reinforcement learning framework with an actor-critic-barrier structure," in *2019 American Control Conference (ACC)*, pp. 2352–2358, IEEE, 2019.

[82] M. Mazouchi, S. Nageshrao, and H. Modares, "Conflict-aware safe reinforcement learning: A meta-cognitive learning framework," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 466–481, 2021.

[83] Z. Marvi and B. Kiumarsi, "Safe reinforcement learning: A control barrier function optimization approach," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 6, pp. 1923–1940, 2021.

[84] Z. Marvi and B. Kiumarsi, "Barrier-certified learning-enabled safe control design for systems operating in uncertain environments," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 437–449, 2021.

[85] Y. Luo and T. Ma, "Learning barrier certificates: Towards safe reinforcement learning with zero training-time violations," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[86] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *Advances in neural information processing systems*, vol. 30, 2017.

[87] J. H. Gillula and C. J. Tomlin, "Guaranteed safe online learning via reachability: tracking a ground target using a quadrotor," in *2012 IEEE International Conference on Robotics and Automation*, pp. 2723–2730, IEEE, 2012.

[88] K. P. Wabersich and M. N. Zeilinger, "Linear model predictive safety certification for learning-based control," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 7130–7135, IEEE, 2018.

[89] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European control conference (ECC)*, pp. 3420–3431, IEEE, 2019.

[90] Z. Li, U. Kalabić, and T. Chu, "Safe reinforcement learning: Learning with supervision using a constraint-admissible set," in *2018 Annual American Control Conference (ACC)*, pp. 6390–6395, IEEE, 2018.

[91] Y. Okawa, T. Sasaki, and H. Iwane, "Control approach combining reinforcement learning and model-based control," in *2019 12th Asian Control Conference (ASCC)*, pp. 1419–1424, IEEE, 2019.

[92] K. P. Wabersich and M. N. Zeilinger, "A predictive safety filter for learning-based control of constrained nonlinear dynamical systems," *Automatica*, vol. 129, p. 109597, 2021.

[93] S. Gros and M. Zanon, "Data-driven economic nmpc using reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 65, no. 2, pp. 636–648, 2019.

[94] M. Zanon and S. Gros, "Safe reinforcement learning using robust mpc," *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3638–3652, 2020.

[95] U. Rosolia and F. Borrelli, "Learning model predictive control for iterative tasks: A computationally efficient approach for linear system," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 3142–3147, 2017.

[96] L. Hewing, J. Kabzan, and M. N. Zeilinger, "Cautious model predictive control using gaussian process regression," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, pp. 2736–2743, 2020.

[97] J. Drgona, A. Tuor, and D. Vrabie, "Learning constrained adaptive differentiable predictive control policies with guarantees," *arXiv preprint arXiv:2004.11184*, 2020.

[98] 3S-Smart Software Solutions GmbH, "Codesys."

[99] T. Maschler and D. A. Savic, "Simplification of water supply network models through linearisation," Tech. Rep. 99/01, University of Exeter, 1999.

[100] V. Gauthier, M.-C. Besner, B. Barbeau, R. Millette, and M. Prévost, "Storage tank management to improve drinking water quality: case study," *Journal of water resources planning and management*, vol. 126, no. 4, pp. 221–228, 2000.

[101] A. Girard, C. E. Rasmussen, J. Q. n. Candela, and R. Murray-Smith, "Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, (Cambridge, MA, USA), p. 545–552, MIT Press, 2002.

[102] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA, 2006.

[103] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine learning*, vol. 22, no. 1, pp. 33–57, 1996.

[104] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *The Journal of Machine Learning Research*, vol. 4, pp. 1107–1149, 2003.

[105] J. Kocijan, *Modelling and control of dynamic systems using Gaussian process models*. Springer, 2016.

[106] C. Macleod, "Interview with application manager (Grundfos)," Jun 2021. Topic: Water ageing of gravity tanks.

[107] S. L. Neve, "Interview with chief bussiness development water& wastewater (Grundfos)," Nov 2021. Topic: Data-driven controllers in industry.

[108] "Finished water storage facilities," tech. rep., U.S. Environmental Protection Agency (EPA), 2012.

Part II

Papers

Paper A

Optimal Control for Water Distribution Networks with Unknown Dynamics

Jorge Val, Rafał Wisniewski and Carsten S. Kallesøe.

The paper has been published in the
21st IFAC World Congress, 2020 Vol. 53(2), pp.6577-6582, 2020.

The layout has been revised.

Abstract

Optimal control for Water Distribution Networks (WDN) is subject to complex system models. Typically, detailed models are not available or the implementation is too expensive for small utilities. Reinforcement Learning (RL) methods are well known techniques for model-free control. This paper proposes a model-free controller for WDNs based on RL methods and presents experimental evidence of the practicality of the design.

1 Introduction

Water Supply Systems (WSS) are critical infrastructures which deliver water from a source to a number of end-users. These systems consist of the following main parts: water sources, treatment plant and storage, transmission stations and distribution network. The WSS studied in this paper consists of the infrastructure after the water treatment plant, where drinking water is transported long distances through a distribution network to the consumer districts. The system overview is illustrated in Figure A.1. The elevated

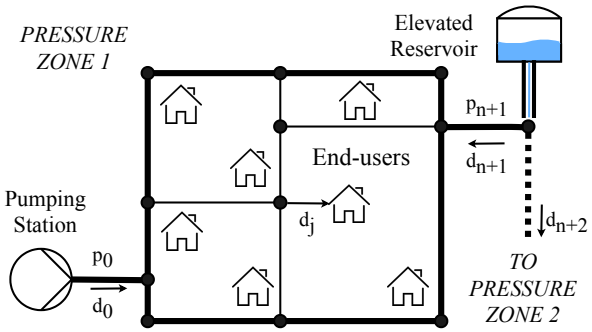


Figure A.1: Illustration of a simplified water distribution network with a pumping station and a storage tank where a city district is supplied through a ring topology network.

reservoirs (ER) play an important role in a water distribution network. The ER contribute to the pressure regulation of the network, additionally these storage units provide extra water capacity to meet demands in different scenarios such as peak demand periods, service works or emergency situations. Having certain storage capacity combined with proper control strategies, provides the system a suitable framework for energy efficient management as shown in many studies [8], [15], most of them in the Model Predictive Control (MPC) framework.

Efficient management of these infrastructures requires complex control algorithms and detailed network models. This requirement increases the commissioning cost of these controllers and makes these strategies unaffordable for most of small utilities. Therefore, plug & play techniques are proposed to give a control solution which adapts to the network complexity, [5], [4].
Reinforcement learning (RL) is a type of machine learning used in multiple disciplines including control of systems. RL methods are employed to find optimal control policies despite of model uncertainties [14], [1]. Hence, control RL (model-free) approaches can provide a great advantage when implementing a control solution in large-scale systems. Promising results are presented in [3], [2] and [13] using RL methods as hierarchical control strategy for other water systems applications.
When dealing with large-scale continuous systems, the amount of state-action pairs required to map values of the system must be considered. RL techniques where the values are stored can become computationally expensive. Instead, function approximation methods evaluate at every step the state-action pair, leading to a compact representation and efficient use of the data samples [7].
[9] and [10] present Q-learning algorithms that converge to an optimal controller by using function approximations. These methods find an approximate value function which replaces the complete mapping of the enormous state-action space.

This paper presents an online control solution that uses a Q-Learning algorithm for a system with unknown dynamics. Additionally, this paper presents a novel reformulation of the state space for including an integral control action on the controller response. Part of the RL algorithm is based on the Linear Quadratic Tracking (LQT) controller presented in [6]. This approach assumes that a full state feedback is available and the reference signal is given by a linear function. In order to validate that this optimal control solution is able to adapt to different network structures and scenarios, the algorithm is tested in a laboratory testbed which emulates a reference WDN. This reference model is based on a realistic network structure which can be typically found in small utilities like Bjerringbro in Denmark. It consists of a single pumping station, a storage tank and the different consumers are interconnected in a ring topology network. Numerical results are obtained in a simulation of Bjerringbro's WDN. Subsequently, experimental results are obtained at the Smart Water Infrastructure (SWI) laboratory at Aalborg University. This modular testbed allows to replicate real infrastructures in a smaller scale. The laboratory is adapted to qualitatively emulate the particular study case.

The rest of this paper is organised as follows. Section 2 recapitulates LQR formulation using Bellman equation. Section 3 describes the model of the WDN. Section 4 reviews the control algorithm design. Section 5 presents the

simulation and experimental results as well as an overview of the testbed used. Section 6 sums up the contributions of the work and relevant ideas for future work.

2 Preliminaries

The work presented in the following section is based on the contribution of [9] and [10] on optimal control and RL. First, a LQR problem is reformulated with the Bellman function. Then, a Q-learning approach is considered to address a LQR problem without knowledge of the system dynamics. Although the following control approach is considered model-free, the problem structure developed in Section 2.1 is used as reference.

2.1 Bellman function based LQR problem

Consider the following linear discrete-time system in the state-space form

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k, \end{aligned} \tag{A.1}$$

where $x_k \in \mathbb{R}^{n_a}$ are the system states, $u_k \in \mathbb{R}^{m_a}$ are the control inputs, and $y_k \in \mathbb{R}^{p_a}$ are the system outputs and A,B and C are constant matrices with compatible dimensions. The reward function is formulated as a quadratic function of the states as follows

$$V(x_k) = \frac{1}{2}\sum_{i=k}^{\infty} \gamma^{i-k} \rho(x_i, u_i) = \frac{1}{2}\sum_{i=k}^{\infty} \gamma^{i-k} \left[x_i^T Q x_i + u_i^T R u_i \right], \tag{A.2}$$

where Q>0, R>0 are weights of the cost function $\rho(x, u)$ and $0 < \gamma < 1$ represents a discount factor that reduces the weight of the cost obtained further in the future. Then, the feedback control policy is given by the linear controller

$$u_k = \pi(x_k) = -Kx_k \tag{A.3}$$

The optimal control policy is found by solving the Linear Quadratic Regulator (LQR) problem by minimising (A.2) over infinite horizon

$$V^*(x_k) = \frac{1}{2}\min_u \sum_{i=k}^{\infty} \gamma^{i-k} \left[x_i^T Q x_i + u_i^T R u_i \right], \tag{A.4}$$

using the given state feedback policy u_k, the solution to the Algebraic Riccati Equation (ARE) gives the matrix P such that

$$V^*(x_k) = \frac{1}{2} x_k^T P x_k, \qquad P = P^T > 0 \tag{A.5}$$

Alternatively, a formulation of this problem can be described by the Bellman equation

$$V(x_k) = \frac{1}{2}\rho_k(x_k, Kx_k) + \gamma V(x_{k+1}), \tag{A.6}$$

where $V(x_{k+1})$ is the cost of the policy K evaluated at the next time step. This paper uses a similar version of (A.6), a q-function where the state x_k and control action u_k are explicitly expressed:

$$q(x_k, u_k) = \frac{1}{2}\rho_k(x_k, u_k) + \gamma V(x_{k+1}) \tag{A.7}$$

By introducing the associated cost function from the LQR problem and (A.5), the q-function can be expressed as

$$\begin{aligned} q(x_k, u_k) &= \frac{1}{2}(x_k^T Q x_k + u_k^T R u_k) + \gamma x_{k+1}^T P x_{k+1} \\ &= x_k^T Q x_k + u_k^T R u_k + \gamma (Ax_k + Bu_k)^T P (Ax_k + Bu_k) \end{aligned} \tag{A.8}$$

Then, (A.8) can be expressed in a matrix form as follows

$$q(x_k, u_k) = \frac{1}{2}\begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} \gamma A^T P A + Q & \gamma A^T P B \\ \gamma B^T P A & \gamma B^T P B + R \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \tag{A.9}$$

Rearranging (A.9) in a compact form yields

$$q(x_k, u_k) = \frac{1}{2}\begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \triangleq \frac{1}{2} z_k^T H z_k \tag{A.10}$$

where $z(x_k, u_k) = \begin{bmatrix} x_k, u_k \end{bmatrix}^T$. Subsequently, the optimal control policy is given by

$$u_k^* = \underset{u}{\operatorname{argmin}}\, q(x_k, u_k) = -H_{uu}^{-1} H_{ux} x_k \tag{A.11}$$

This is the optimal control action when the system dynamics is completely known and full state feedback x_k is available.

2.2 Q-learning for LQR

In this section, the system dynamics is unknown. Then, the Bellman optimality principle is applied to formulate the q-function (A.7) in a recursive form. First, by introducing the Bellman optimality equation $V_k^*(x_k) = \min_u q_k(x_k, u_k)$ into the q-function (A.7) leads

$$q_{k+1}(x_k, u_k) = \rho_k(x_k, u_k) + \gamma q_k(x_{k+1}, K^* x_{k+1}), \tag{A.12}$$

where K^* is the optimal policy. In the future the next state is denoted as $x' = x_{k+1}$.

Then, the q-function expression (A.12) is rearranged based on the RL Temporal Difference (TD) method for prediction proposed in [14]

$$\begin{aligned} q_{k+1}(x_k, u_k) =& q_k(x_k, u_k) \\ &+\alpha \left[\rho(x_k, u_k) + \gamma \min_u q_k(x', u) - q_k(x_k, u_k)\right], \end{aligned} \tag{A.13}$$

where α represents the learning rate. Finally, the expression (A.13) is reformulated to obtain the update law which gives the q value.

$$\begin{aligned} q_{k+1}(x_k, u_k) = \\ (1-\alpha) q_k(x_k, u_k) + \alpha \left[\rho(x_k, u_k) + \gamma q_k(x', u')\right] \end{aligned} \tag{A.14}$$

where u' represents the optimal control action with $u' = \pi^*(x)$.

3 System model

A WDN consists of a pipe network with different elements such as valves, pumps and elevated reservoirs. The distribution network is divided into several districts - Pressure Zones (PZ), see Figure A.1. The end-users water consumption (demands) are generally an unknown input or disturbance to the system.

3.1 Network Model

The studied network model is restricted to a ring topology which is a structure typically found in small water utilities. This model can be simplified by unifying the end-users (nodes) that are geographically close because the pressure loss due to pipe resistance is relatively low between them [11]. Figure A.1 shows a standard ring network where the multiple end-user demands are represented by aggregated demands from the main pipes d_j, the controlled inflow from the pumping station is denoted by d_0 and the tank inflow by d_{n+1}. Due to mass conservation in the network, the relation between supply flow d_0, the reservoir flow d_{n+1} and the end-user water consumption d_j can be denoted as

$$d_0 + d_{n+1} = -\sum_{j=1}^{n} d_j, \tag{A.15}$$

where $d_j \leq 0$ and n is the number of end-user demands. Then, by assuming that the distribution of daily water consumption between the end-users is alike, the demand profile for all the consumers can be described by

$$d_j = \beta_j \bar{d} \qquad \forall j = 1, \ldots, n \tag{A.16}$$

where β_j is a constant describing the distribution, $\sum_{j=1}^{n} \beta_j = 1$ and $\bar{d}$ is the total district demand in a PZ. The pressure at the reservoir node p_{n+1} is given by the level h in the reservoir and the geodesic level h_0.

$$p_{n+1} = \mu(h + h_0) \tag{A.17}$$

where μ is a constant scaling the water level and pressure unit and h is the tank level that belongs to an interval restricted by the height of the reservoir. The reservoir level rate depends on the flows leaving the reservoir (d_{n+1} and d_{n+2})

$$A_t \dot{h} = -d_{n+1} - d_{n+2}, \tag{A.18}$$

where A_t is the constant cross sectional area of the elevated reservoir and the outflow to other PZs d_{n+2}. For simplicity, in the laboratory test this outflow is not further considered.

4 Control

The management of WDNs must ensure the supply of water to the end-users with sufficient pressure head and quality, this task must be performed while considering multiple objectives during the daily operation. Some studies performed in [12] state some control objectives: economic, safety, smoothness and water quality.
In this paper only safety is considered in the control strategy, this means that the operational goal is to guarantee the water supply to the end-users. This control task is challenging due to the uncertainty of the water consumption. Therefore, storage tanks must contain enough water to meet future stochastic demands.

4.1 Internal Model Principle

One of the contributions of [6] is the solution to the LQT problem and quadratic form of the LQT value function where the problem is formulated as a quadratic form in terms of the system states x and trajectory reference r. In this paper, an additional extension of the state space is proposed for introducing an integral action ξ which rejects the constant disturbances - demands. The augmented system model is built as follows.
First, the physical model above (A.18) is expressed in a state space form for the control design

$$\begin{aligned} \dot{h} &= A_c h + B_c u + W_c d \\ y_c &= C_c h, \end{aligned} \tag{A.19}$$

where $h \in \mathbb{R}$ represents the tank level, $u \in \mathbb{R}$ the controlled inflow d_0 and $d \in \mathbb{R}$ the end-user demand, with A_c,B_c and C_c constant matrices with

compatible dimensions. Then, a reference trajectory r is defined by a linear function

$$\dot{r} = Lr, \tag{A.20}$$

where $r \in \mathbb{R}$, then defining the integral error

$$\dot{\xi} = y_c - r \tag{A.21}$$

Equations (A.19), (A.20) and (A.21) are combined to build the following augmented state space model

$$\begin{bmatrix} \dot{h} \\ \dot{r} \\ \dot{\xi} \end{bmatrix} = \begin{bmatrix} A_c & 0 & 0 \\ 0 & L & 0 \\ C_c & -I & 0 \end{bmatrix} \begin{bmatrix} h \\ r \\ \xi \end{bmatrix} + \begin{bmatrix} B_c \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} u \end{bmatrix} + \begin{bmatrix} W_c \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} d \end{bmatrix} \tag{A.22}$$

Finally, expressing the state space representation (A.22) in a more compact form for discrete time

$$\begin{aligned} x_{k+1} &= A_e x_k + B_e u_k + W_e d_k \\ y_k &= C_e x_k, \end{aligned} \tag{A.23}$$

where $x = \begin{bmatrix} h, & r, & \xi \end{bmatrix}^T$ is the augmented state vector. A cost (reward) function similar to the previously stated in (A.2) is built by using the augmented system output from (A.23)

$$V(x_k) = \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} \left[y_i^T Q y_i + u_i^T R u_i \right]. \tag{A.24}$$

By reformulating (A.24) with $C_e = \begin{bmatrix} C_c & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, the cost function includes the tracking error in terms of x and u.

$$\begin{aligned} V(x_k) &= \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} [(C_c h_i - r_i)^T Q_1 (C_c h_i - r_i) + \xi_i^T Q_2 \xi_i + u_i^T R u_i] \\ &= \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} (x_i^T Q_e x_i + u_i^T R u_i) = \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} \rho(x_i, u_i) \end{aligned} \tag{A.25}$$

with $Q_1 > 0$, $Q_2 > 0$ and R>0 and

$$Q_e = \begin{bmatrix} C_c^T Q_1 C_c & -C_c^T Q_1 & 0 \\ -Q_1 C_c & Q_1 & 0 \\ 0 & 0 & Q_2 \end{bmatrix}. \tag{A.26}$$

4.2 q-function Approximation using linear architectures

A linear architecture is selected for the approximation over other black-box methods such as Neural Networks. Although the latter methods can provide a more generalised solution, a linear architecture is easier to implement since its behaviour is more transparent, facilitating the troubleshooting task when the algorithm fails.
The q-function proposed in (A.10) is linearly approximated by a set of Basis Functions (BF) ϕ and the corresponding coordinate vector θ or weights. The BFs are a combination of monomial basis. Thus, learning upon the state vector structure from (A.10) which is quadratic, a finite set of monomial basis of 2^{nd} degree polynomials, formed with x and u, is chosen as follows. For a multi-index $a \in \mathbb{Z}^{n_a} \geq 0$, with $\mid a \mid = a_1 + \cdots + a_{n_a}$,

$$\hat{q}(x,u) = \sum_{|b|=2} \theta_{(b,0)} x^b + \sum_{|a|=1} \theta_{(a,1)} x^a u + \theta_{(0,2)} u^2. \tag{A.27}$$

Then, by representing (A.27) in a vector form

$$\hat{q}(x,u) = \phi^T(x,u)\theta, \tag{A.28}$$

where ϕ is an n_b-dimensional column vector of BFs and θ is an n_b-dimensional coordinate vector and $n_b = m_a n_a + p_a n_a + m_a$

$$\phi = [x_1^2, x_1 x_2, \ldots, x_{n_a}^2, x_{n_a} u, u^2]^T \tag{A.29}$$

Subsequently, the approximated control law can be described as $u = \hat{\pi}(\theta, x)$, where $\hat{\pi}(\theta, x)$ can be computed by

$$u_k' = \underset{u}{\operatorname{argmin}}\, \hat{q}(x_k, u_k) = \underset{u}{\operatorname{argmin}}\, \phi^T(x_k, u_k)\theta \tag{A.30}$$

This yields to the feedback control policy given by the linear controller

$$u_k' = \hat{K}(\theta) x_k \tag{A.31}$$

Alternatively, since (A.27) is quadratic with respect to x and u, a moment matrix $\hat{H}$ can be formed with the coordinates of the BFs such that

$$\hat{q}(x_k, u_k) = z_k^T \hat{H}(\theta) z_k, \tag{A.32}$$

where $z_k = \left[x_k, u_k\right]^T$ and $\hat{H}$ matrix is a symmetric matrix parametrised with the coordinate vector θ as follows $\hat{H} = \begin{bmatrix} \theta_1 & \frac{\theta_2}{2} & \cdots \\ \frac{\theta_2}{2} & \theta_3 & \cdots \\ \vdots & \vdots & \theta_l \end{bmatrix}$ where $\hat{H} \in \mathbb{R}^{n_b(n_b+1)/2}$
Note that q-function (A.10) and approximated q-function (A.32) have the same quadratic structure.

4.3 Parameter Update

For the following method, a sample is organised as a tuple of (x_k, u_k, ρ_k, x') and a data batch as a set of collected samples $(\overline{x}_{l_s}, \overline{u}_{l_s}, \overline{\rho}_{l_s}, \overline{x}'_{l_s} \mid s = 1, \ldots, n_l)$ where n_l is the batch size and the index l is the batch iteration number.
The coordinate vector θ is initially unknown, therefore the parameters must be recursively learned. For this, the q-value approximation (A.28) is introduced into the update law (A.13)

$$\phi^T(x_k, u_k)\theta_{k+1} = (1-\alpha)\phi^T(x_k, u_k)\theta_k + \alpha\left[\rho(x_k, u_k) + \gamma\phi^T(x', u')\theta_k\right] \tag{A.33}$$

Then, by evaluating (A.33) recursively, a batch of samples is obtained. The update law for a batch is denoted as

$$\Phi_l^T\theta_{l+1} = (1-\alpha)\Phi_l^T\theta_l + \alpha\left[J_l + \gamma\Phi_l^T(x', u')\theta_l\right] \tag{A.34}$$

where $\Phi \in \mathbb{R}^{n_b \times n_l}$ is a matrix of BFs ϕ, $J \in \mathbb{R}^{n_l}$ is the vector of rewards ρ collected on a batch iteration l. In order to solve the expression (A.34), a linear Least-Squares Temporal Difference (LSTD) method, similar to [7], is followed to solve the q-function

$$\theta_{l+1} = (1-\alpha)\theta_l + \alpha G_l^{-1}\Phi_l\left[J_l + \gamma\Phi_l^T\theta_l\right] \tag{A.35}$$

Note that a persistent excitation must be added to the control signal such that the term $G_l = \Phi_l\Phi_l^T$ is invertible. The equation (A.35) is solved by recursively executing the steps described in Algorithm 3.

Algorithm 3 LSTD for Q-function.

Input: γ, α, n_s,
Approximation mapping of the BFs,
Initialisation: $l \leftarrow 0, x_0, \theta_0$ where $\hat{\pi}(\theta_0)$ must be an admissible policy.
repeat at every iteration k = 0,1,2, ...
  apply $u_k = K_l x_k$ and measure x_{k+1}
  $Y_{l_s} \leftarrow \rho(x_k, u_k) + \gamma\hat{q}(x_{k+1}, K_l x_{k+1})$
  if $k = (l+1)n_s$ **then**
    $\theta_{l+1} \leftarrow (1-\alpha)\theta_l + \alpha G_l^{-1}\Phi_l Y_l$
    $\hat{\pi}(\theta_{l+1}, x) \leftarrow \text{argmin}_u\, \Phi_l^T\theta_{l+1}$
    $l \leftarrow l+1$
  end if
until $\|\theta_{l+1} - \theta_l\| < \epsilon$

5 Results

To validate the practicality of the proposed control strategy, Algorithm 3 is tested in a computer simulation, then deployed in the Smart Water Laboratory. In this application example the pressure in the network is regulated by controlling the level in the tank. The pressure at the node p_{n+1} is set conservatively enough to meet the flow demands. The weights of the reward function (A.24) are set to prioritise the minimisation of the tracking error over control action.
The discount factor γ is set close to 1 nearly to the optimal solution, while the learning rate α is sufficiently small such that the old information prevails over new information collected.

5.1 Numerical Results

A simulation environment is developed with the purpose of verifying the proposed control algorithm and training for further implementation. This computer simulation reproduces the water network model from Bjerringbro, a simplified version of the aforementioned network is illustrated in Figure A.1.
As shown in Figure A.2, the tank level has an oscillatory transient where the system dynamics are controlled with a non-optimal policy. Once the learning is considered satisfactory, the persistent excitation on the control action is no longer applied and the tank level stabilises at the reference target despite of the demands $\bar{d}$ and d_{n+2}. This excitation consists of a sum of sines and cosines of different frequencies. In Figure A.3, the coordinate vector parameters θ converge to a satisfactory policy.

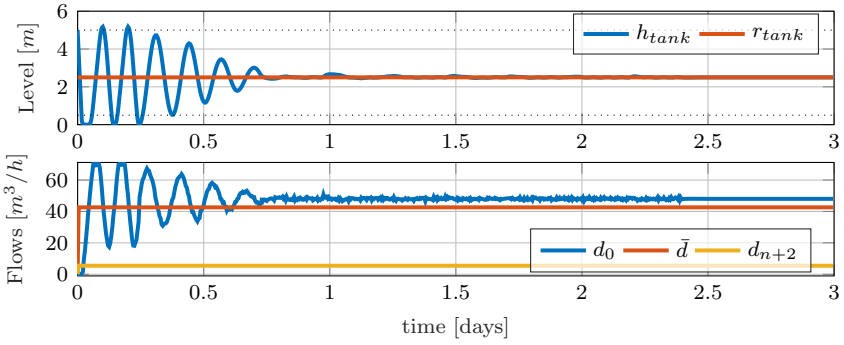


Figure A.2: Simulation Results. Top: Tank Level (blue), reference level(red) Bottom: Controlled input flow (blue), Water Demands (red) (yellow)

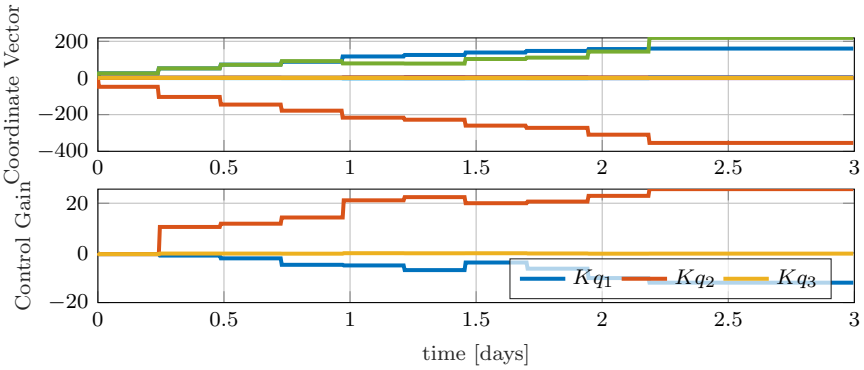


Figure A.3: Simulation Results. Top: Coordinate vector. Bottom: Control Policy

5.2 Experimental Results

The testbed scheme consists of a set of Laboratory Units (LU) that can be interconnected to reproduce the desired network. As mentioned earlier, data from Bjerringbro WDN is used to emulate a real water utility. This WDN consists of a single pumping station and storage units, see Figure A.4 and Figure A.5.

Figure A.4: Photo of the SWI laboratory.

The WDN is built in the laboratory by two aggregated consumers in the City Districts (CD1 and CD2), a pumping station (Pu1), an Elevated Reservoir (ER) and multiple pipe units to reproduce the network structure. A local

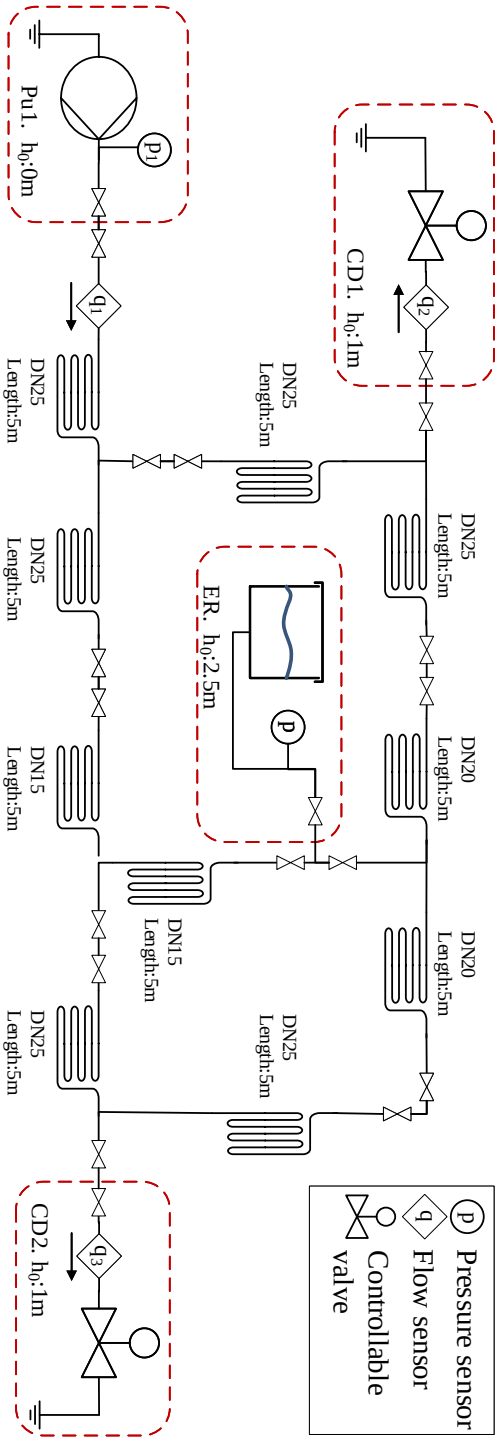


Figure A.5: Detailed topology of the laboratory setup

controller ensuring fast flow control is implemented at Pu1. CD1 and CD2 are equipped with a valve regulating the water consumption. Different geodesic levels h_0 at each critical node (Pu1, ER, CD1, CD2) are simulated by air-pressurising the collecting containers with the equivalent head pressure. The LUs are equipped with multiple sensors and actuators. Each of them has a soft-PLC in charge of the data acquisition, local control and communication. The soft-PLCs at the LUs are interfaced with *CODESYS Control*. Furthermore, the LUs are interconnected to a Central Control Unit (CCU) that can be used for central management of the modules.
The control Algorithm 3 for optimal level control is tested in the described laboratory setup. An admissible initial policy is given based on simulation training. As shown in Figure A.6, the tank level is regulated around the reference after some adaptation period. A small error is observed in steady state due to the different accuracy of the flow sensors. Figure A.7 shows the update of the q-function parameters based on the new data, adapting the optimal policy to the new system.

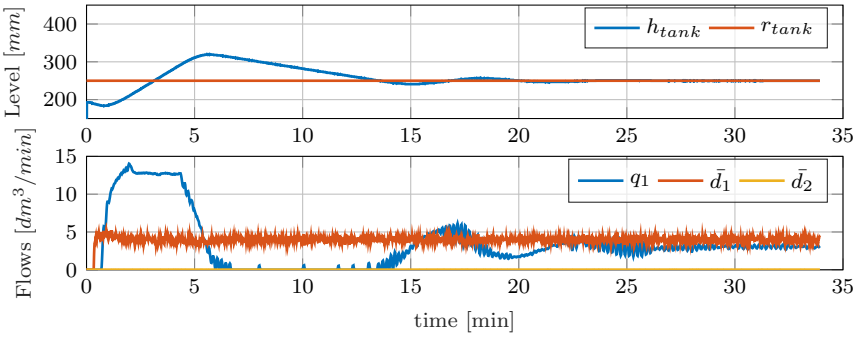


Figure A.6: Experimental Results. Top: Tank Level (blue), reference level(red) Bottom: Controlled input flow (blue) Water Demand (red)

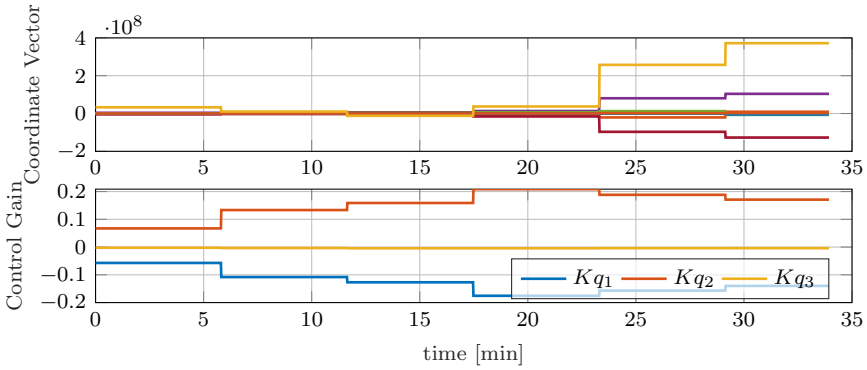


Figure A.7: Experimental Results. Top: Coordinate vector. Bottom: Control Policy

6 Discussion and Future Work

The q-learning algorithm succeeded in finding an approximated optimal policy. However, the learning process in a real system is uncertain. This exploration typically leads to saturation of the control actuators and violation of the safety boundaries on the testbed. This factor is a limitation when implementing the controller on systems that have physical boundaries compared with other solutions such as MPC.
The integral action successfully rejects disturbances when the demand profiles are constant. In real scenarios, stochastic disturbances occur, which must be considered in the control design. Due to the real system non-linearity and stochastic disturbances, which are not considered in this control approach, the algorithm does not reach a smooth convergence of the parameters. However, it can be observed that the variation of the controller gains remains to a stable value during the learning, see Figure A.7.

In the future, in order to improve the applicability to a high-dimensional system, this control approach can be improved by considering periodic disturbances in the control design. Moreover, a controller for WDNs must include input and output constraints that set the safe operation boundaries.

7 Conclusion

A model-free solution is proposed to regulate the level in the ER in a WDN. This adaptive-optimal control is successfully implemented on a small-scale WDN since the tank level is regulated despite not having the network model. Furthermore, a novel approach is presented, an integral action in the control

policy that compensates steady-state constant disturbances. This solution offers an easy-commissioning tool which can reduce the implementation costs.

Acknowledgement

We would like to thank *Poul Due Jensen Fond* for funding this project and *Bjerringbro vandforsyning* for providing the network data.

References

[1] D. P. Bertsekas, *Dynamic programming and optimal control. Vol. 2*, 3rd ed., ser. Athena scientific optimization and computation series. Athena Scientific, 2007.

[2] A. Castelletti, G. Corani, A.-E. Rizzoli, R. Soncini-Sessa, and E. Weber, "Reinforcement learning in the operational management of a water system," in *Modelling and Control in Environmental Issues*. Pergamon Press, 2002.

[3] E. Ertin, A. Dean, M. Moore, and K. Priddy, "Dynamic optimization for optimal control of water distribution systems," *Proceedings of SPIE - The International Society for Optical Engineering*, 2001.

[4] T. N. Jensen, C. Kallesøe, J. D. Bendtsen, and R. Wisniewsk, "Plug-and-play Commissionable Models for Water Networks with Multiple Inlets*," in *2018 European Control Conference (ECC)*, 2018.

[5] C. S. Kallesøe, T. N. Jensen, and J. D. Bendtsen, "Plug-and-Play Model Predictive Control for Water Supply Networks with Storage," *IFAC-PapersOnLine*, vol. 50, no. 1, 2017.

[6] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, 2014.

[7] M. G. Lagoudakis and R. Parr, "Least-Squares Policy Iteration," *Journal of Machine Learning Research*, 2003.

[8] S. Leirens, C. Zamora, R. R. Negenborn, and B. De Schutter, "Coordination in urban water supply networks using distributed model predictive control," in *American Control Conference*, 2010.

[9] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement Learning for Partially Observable Dynamic Processes: Adaptive Dynamic Programming Using Measured Output Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, no. 1, 2011.

[10] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement Learning and Feedback Control: Using Natural Decision Methods to Design Optimal Adaptive Controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, 2012.

[11] T. Maschler and D. A. Savic, "Simplification of water supply network models through linearisation," University of Exeter, Tech. Rep. 99/01, 1999.

[12] C. Ocampo-Martinez, V. Puig, G. Cembrano, and J. Quevedo, "Application of Predictive Control Strategies to the Management of Complex Networks in the Urban Water Cycle," *Control Systems, IEEE*, 2013.

[13] D. Ochoa, G. Riaño-Briceño, N. Quijano, and C. Ocampo-Martinez, "Control of Urban Drainage Systems: Optimal Flow Control and Deep Learning in Action," in *American Control Conference (ACC)*, 2019.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, 2nd ed., ser. Adaptive computation and machine learning series. The MIT Press, 2018.

[15] Y. Wang, V. Puig, and G. Cembrano, "Non-linear economic model predictive control of water distribution networks," *Journal of Process Control*, vol. 56, 2017.

Paper B

Reinforcement Learning Control for Water Distribution Networks with Periodic Disturbances

Jorge Val, Rafał Wisniewski and Carsten S. Kallesøe.

The paper has been published in the
2021 American Control Conference (ACC), pp. 1010-1015, 2021.

The layout has been revised.

Abstract

Cost efficient management of Water Distribution Networks with storage units requires of extensive knowledge of the water network. However, the network models are not always available or the calibration costs are too high for most of small water utilities. This paper proposes a model-free control solution based on Q-learning methods that provides a policy for the operation of the network. This supervisory controller must guarantee the water supply despite of the uncertainty of the daily water consumption and reduce the operation cost. The function approximation proposed for the Q-learning controller uses Fourier Basis Functions which provide an accurate approximation of the periodic disturbances. This paper presents results of the control validation in a simulation framework as well as experimental evidence of the advantages and limitations of the proposed design.

1 Introduction

Water Distribution Networks (WDNs) are large-scale systems requiring a considerable amount of energy, this consumed energy is in most cases produced using fossil fuels. WDNs with elevated reservoirs have some storage capabilities that can be utilised to save energy. Therefore, an efficient operation of the WDN can reduce the carbon footprint. Additionally, the management at the water utilities must guarantee a robust water supply, this operation becomes difficult since the water consumption in a urban district is uncertain. In order to achieve an optimal management of these infrastructures where all the objectives are satisfied, Model Predictive Control (MPC) management strategies are implemented where models of the network are required [4], [16]. However, these models are not always available or the maintenance costs are too high. In order to facilitate that modern control techniques are implemented by a greater number of utilities, easy commissioning control tools are developed [15],[1]. Reinforcement Learning (RL) is a machine learning technique which has been successfully implemented in various control applications [9], [2]. RL is a model-free optimisation method that can be combined with function approximators to extend the application of these methods to continuous state-spaces.
The periodic disturbance signal described by the consumed water can be approximated using Fourier Series. Some studies argue that Fourier Basis functions can provide a better approximation of the system when using RL methods [7] compared to popular fixed bases. This paper proposes an extension of the Q-learning control strategy proposed in [14] that includes Fourier Basis for the approximation of the periodic disturbances. Additionally, the control objectives are defined to minimise the energy usage during the operation. Linearising a smooth non-linear system via a small-signal analysis is

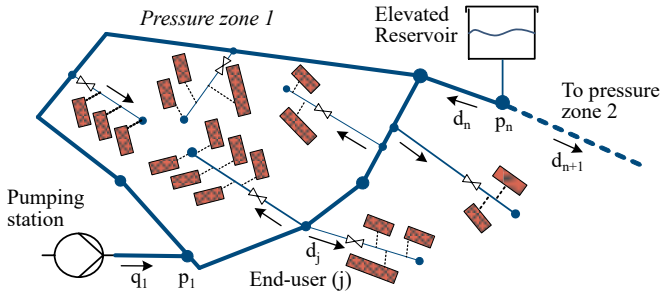


Figure B.1: Illustration of a simplified WDN with a pumping station, an elevated reservoir and multiple end-users located in a city district.

a widely used technique in control design. This method allows to use linear controllers in non-linear systems around an operating point [5]. However, this operating point is often unknown in real systems. In this paper, a structure of a linear system is assumed, this model is adaptatively updated in the Q-function identification scheme.

The rest of this paper is organised as follows. Section 2 develops a system model of a water network with an elevated reservoir. The structure of this model is reduced for convenience in the control design, the periodic disturbances are considered in the control strategy and approximated via Fourier series. Section 3 defines the control strategy for a WDN with unknown dynamics. Then, the learning algorithm is described. Section 4 presents numerical and experimental results where the control strategy is validated. Section 5 summarises the contribution of the work and gathers some ideas for future development of the presented method.

2 System model

A WDN is generally divided into several districts or pressure zones. The main network elements of the studied district are illustrated in Figure B.1. The scope of this system is to supply drinking water from the pumping station to the end-users where the water is consumed. The water consumption in an urban district is typically unknown, and it is considered as disturbance to the system. The uncertainty in the demand hampers the good operation of the system. However, these demands typically follow a daily pattern and therefore they can be approximated. This section also contains the assumptions made to represent a WDN system in a linear form.

2.1 Water network model

This study proposes a heuristic model of a WDN where the only dynamic element is the elevated reservoir. The reduced network structure is represented in Figure B.2 and it consists of four main elements: a pump station, a consumer district, an equivalent pipe and an elevated reservoir. The purpose of such simplification is to represent the bottom line dynamics of a WDN. This simple structure is later used for the approximation scheme in the model-free controller in Section 3. This model assumes that the pump is locally

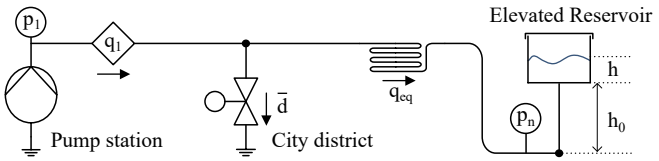


Figure B.2: Diagram of an equivalent WDN.

controlled to deliver the input flow $q(t)$. A simple model of the power of a pump $P(t)$ is defined as

$$P(t) = q(t)\Delta p(t)/\eta, \tag{B.1}$$

where η is a constant representing the performance of the pump. In this network model a set of end-users, which are geographically close, are represented by a single nodal demand [11]. The relation between the demand profile for each end-user $d_j(t)$ and total water demand is described by

$$d_j(t) = v_j \bar{d}(t) \tag{B.2}$$

where $\bar{d}$ is the total water demand, v_j is a constant describing the distribution between the demands, j is the consumer node index. A district pipe network is reduced to a single pipe which has an equivalent pipe resistance. The pressure drop in the equivalent pipe is given by

$$\Delta p_{eq}(t) = \lambda(q_{eq}(t)) + \Delta z_{eq}, \tag{B.3}$$

where λ is the pressure loss due to pipe friction and Δz_{eq} is the differential geodesic level between two nodes. The equivalent flow q_{eq} can be calculated assuming mass conservation in the pipe node $q(t) = \bar{d}(t) + q_{eq}(t)$.

Assumption 1. *The hydraulic resistance λ takes the form $r_{eq}|q_{eq}|q_{eq}$ and $r_{eq} > 0$ is a constant parameter. This form describes the friction losses in a pipe with turbulent flow [13].*

The elevated reservoir dynamics is given by the following equation

$$A_{er}\dot{h} = q(t) + \bar{d}(t) \tag{B.4}$$

where A_{er} is the cross sectional area of the elevated reservoir. Then, the pressure at the elevated reservoir node p_n is given by

$$p_n(t) = \mu(h(t) + h_0) \tag{B.5}$$

where μ is a constant scaling the water level and pressure unit and h_0 is the elevation of the tank. Then, for the reduced network topology, the pressure at the pumping station is given by

$$p(t) = p_n(t) + \Delta p_{eq}(t) \tag{B.6}$$

The model presented in (B.4) can be expressed as a linear discrete-time system in the state-space form,

$$h_{k+1} = Ah_k + Bq_k + E\bar{d}_k \tag{B.7}$$

where $h_k \in \mathbb{R}$ is the system state, $q_k \in \mathbb{R}$ is the controlled input flow and $d_k \in \mathbb{R}$ are the system disturbances, and A,B and E are constant matrices with compatible dimensions.

2.2 Inclusion of periodic disturbances

A real water demand can be described as a stochastic Wienner process with a period of one day. Thus, it follows a similar pattern from day to day. It is assumed that this signal, with known periodicity, can be approximated using Fourier Series (FS) of certain order N. The signal approximation is developed as follows, consider a FS continuous signal of the form

$$\bar{d}(t) = a_0 + \sum_{n=1}^{N}(a_n \cos(\omega_n t) + b_n \sin(\omega_n t)) + w, \tag{B.8}$$

where a_0, a_n and $b_n \in \mathbb{R}$ are the Fourier coefficients, $\omega_n = 2\pi n f_0$ and f_0 represents the fundamental frequency and w is Brownian noise. Note that for the studied case the f_0 is determined by a period of one day. The mean of the periodic function (B.8) can be reformulated using a discrete-time state space representation as follows

$$\begin{aligned} s_{k+1} &= A_d s_k, \\ d_k &= C_d s_k, \end{aligned} \tag{B.9}$$

where the system matrix $A_d = \text{diag}(1, F_1, \ldots, F_N)$, with $F_n = \begin{bmatrix} \cos(\omega_n \Delta t) & -\sin(\omega_n \Delta t) \\ \sin(\omega_n \Delta t) & \cos(\omega_n \Delta t) \end{bmatrix}$ where Δt is the sampling time and the output matrix C_d includes the Fourier

coefficients. The state vector $s_k \in \mathbb{R}^{n_d}$, with $n_d = 2N + 1$, is subject to the following initial condition

$$s_{i,t_0} = \begin{cases} c_0 & \text{if } i = 0 \\ \cos(\omega_n t_0) & \text{if } i > 0, \quad i \text{ odd} \\ \sin(\omega_n t_0) & \text{if } i > 0, \quad i \text{ even} \end{cases} \tag{B.10}$$

where c_0 is a constant, t_0 is the initial time value and the index vector $i \in \mathbb{Z}, [0, n_d]$. Figure B.3 (top) depicts the resulting output signal compared with real water consumption data. The time series data of the district's water consumption is provided by *Bjerringbro vandforsyning*, a small water utility in Denmark. Note that for N=3, the output signal describes the water consumption pattern with high and low demand period.

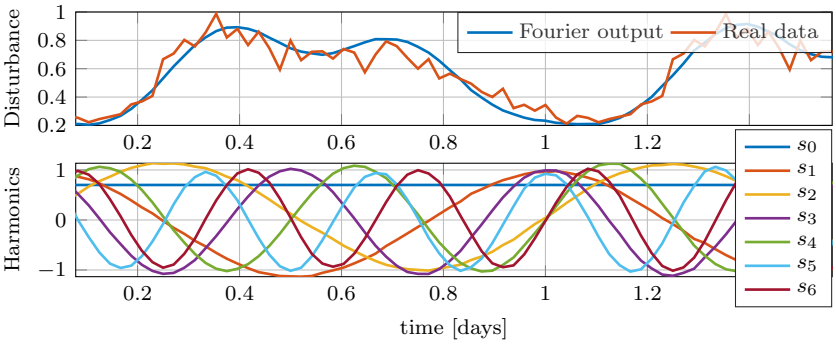


Figure B.3: Top: Output disturbance signal and consumption data for two days. Bottom: Fourier Harmonic states for n=0,1,2,3.

2.3 Augmented state space model

This subsection presents a reformulation of the WDN model such that the disturbance term is included in the state vector of the system. For this, the elevated reservoir model (B.7) and the periodic disturbance function (B.9) are combined in an augmented state space of the form

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} A & EC_d \\ \mathbf{0} & A_d \end{bmatrix} x_k + \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix} u_k, \\ y_k &= \begin{bmatrix} \mathbf{I} \\ C_p \end{bmatrix} x_k + \begin{bmatrix} \mathbf{0} \\ D_p \end{bmatrix} u_k, \end{aligned} \tag{B.11}$$

where $x_k = \begin{bmatrix} h_k & s_k \end{bmatrix}^T$, u_k is the controlled input and $y_k = \begin{bmatrix} h_k & s_k & p_k \end{bmatrix}^T$ is the measured output vector. Note that the pressure at the pumping station p is introduced in the output signals as linear variable by considering assumption 2.

Assumption 2. *The pressure losses in the equivalent network are given by* (B.3). *Considering that the system operates around an arbitrary operating point* $\bar{q}_{eq}$*, the term* $\lambda(q_{eq})$ *can be linearised by using a first order Taylor series, and therefore the pressure at the pumping station* (B.6) *can be approximated with the linear structure* $p_k \approx C_p x_k + D_p u_k$.

Note that, the approximation of p_k uses the states from the Fourier Series vector to represent the pressure offset introduced by Δz_{eq} and h_0 as well as the periodic disturbance $\bar{d}$. Finally, representing the system (B.11) in a compact form,

$$\begin{aligned} x_{k+1} &= A_e x_k + B_e u_k, \\ y_k &= C_e x_k + D_e u_k, \end{aligned} \tag{B.12}$$

where $x \in \mathbb{R}^{m_a}$, $u \in \mathbb{R}^{n_a}$. The feedback control policy is given by the following controller

$$u_k = -K x_k. \tag{B.13}$$

3 Control

As previously mentioned the main goal of a WDN is to deliver water to the end-users. However, the management of a WDN comprises multiple competing objectives that must be considered, such as economic, water quality and safety. This work proposes a reward function to regulate the trade-off between different management objectives. The first objective is to ensure the supply at the district by regulating the level in the tank, the second is to reduce the network operation cost by minimising the pump effort. Recall that pumping stations are locally controlled, the controller proposed in this section acts as a supervisory control that regulates the WDN management.

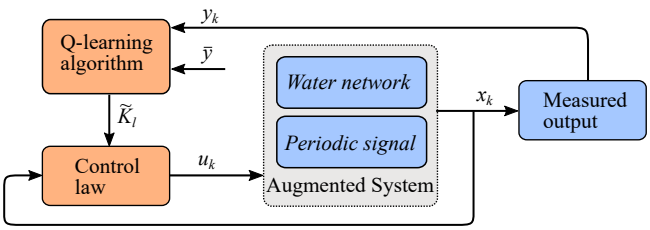


Figure B.4: Block diagram representation of the control algorithm.

3.1 Cost function - Bellman Equation

Inspired by the formulation of LQR problem with Bellman equation presented in [10], this paper proposes a formulation of the Q-value function

which includes energy optimisation for systems with an equilibrium point different from zero. In order to obtain an optimal control policy which leads the system to fulfil the desired control objectives, a cost (reward) function is defined as

$$V^{\pi}(x_k) = \sum_{i=k}^{\infty} \gamma^{i-k} \left[(y_i - \bar{y})^T Q_1 (y_i - \bar{y}) + u_i^T Q_2^T y_i + y_i^T Q_2 u_i + u_i^T R u_i \right], \quad \text{(B.14)}$$

where $\bar{y}$ is a vector with constant reference values, $Q_1 > 0$,$Q_2 > 0$ and $R > 0$ are weight matrices and $0 < \gamma < 1$ represents a discount factor that reduces the weight of the cost obtained further in the future. Subsequently, the instant reward is denoted as

$$\rho(x_k, u_k) = (y_i - \bar{y})^T Q_1 (y_i - \bar{y}) + u_k^T Q_2 y_k + y_k^T Q_2 u_k + u_k^T R u_k, \quad \text{(B.15)}$$

The first term represents the deviation of the tank level with respect to a given reference, it is used as a soft constraint to maintain the level within the range of operation. The middle terms represent the pump effort in terms of energy consumed (B.1) and the last term penalises high control actions. Note that, the terms corresponding to energy consumption and control action are not minimised to zero during the WDN operation, the discount factor γ bounds the cost function (B.14) from accumulating these non-zero rewards when time goes to infinity.

This control formulation assumes that the full-state feedback is available, since tank level h_k, pump pressure p_k are measured and the harmonics of the FS s_k can be computed by solving (B.10) for a given f_0. Note that the system must be at least detectable for solving ARE [6]. According to Bellman's optimality principle, the value can be determined using the HJB equation as follows

$$V^*(x_k) = \min_u (\rho(x_k, u_k) + \gamma V^*(x_{k+1})) \quad \text{(B.16)}$$

with the notation $(\cdot^*)$ representing the optimal value. Then, assuming that there exists a candidate solution to the value function (B.16), of the form

$$V(x_k) = x_k^T P x_k + G x_k + c, \quad \text{(B.17)}$$

the solution (B.17) can be introduced in (B.16) as follows,

$$V^*(x_k) = \min_u (\rho(x_k, u_k) + \gamma (x_{k+1}^T P x_{k+1} + G x_{k+1} + c) \quad \text{(B.18)}$$

The value function (B.16) is transformed to Q-value function where the control action is expressed explicitly.

$$Q^*(x_k, u_k) = \rho(x_k, u_k) + \gamma Q(x_{k+1}, u^*) \quad \text{(B.19)}$$

Additionally, the augmented system model (B.12) is introduced

$$\begin{aligned} Q(x_k, u_k) = &(Cx_k + Du_k - \bar{y})^T Q_1 (Cx_k + Du_k - \bar{y}) \\ &+ (Cx_k + Du_k)^T Q_2 u_k + u_k^T Q_2 (Cx_k + Du_k) + u_k^T R u_k \\ &+ \gamma[(Ax_k + Bu_k)^T P (Ax_k + Bu_k) + G(Ax_k + Bu_k) + c]. \end{aligned} \tag{B.20}$$

Then, the expression (B.20) can be reformulated in a matrix form

$$Q(x_k, u_k) = \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} M_{xx} & M_{xu} \\ M_{ux} & M_{uu} \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} N_x \\ N_u \end{bmatrix} + \begin{bmatrix} N_x \\ N_u \end{bmatrix}^T \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \tilde{c} \tag{B.21}$$

By expressing (B.21) in a compact form, quadratic, linear and constant components are identified. Let $z_k = [x_k, u_k]^T$, subsequently

$$Q(z_k) = z_k^T M z_k + 2N^T z_k + \tilde{c} \tag{B.22}$$

Then, the optimal control policy for (B.22) can be calculated as

$$u_k^* = \underset{u}{\operatorname{argmin}}\, Q(x_k, u_k) = M_{uu}^{-1}(M_{ux} x_k + N_u) \tag{B.23}$$

Note that, the resulting control law (B.23) in is affine, the offset in the control action represents the system regulation around an equilibrium point different from zero. Finally, by introducing the optimal control action (B.23) into the Q-value (B.20), the Q-value function results in an equation with the same structure as (B.17). Therefore, the proposed solution is proved to be a valid choice for this problem.

3.2 Q-value function approximation using linear architectures

The control approach presented thus far assumes the knowledge of the system dynamics. This paper proposes a model-free control strategy where the system dynamics are unknown. Thus, neither of the model matrices A,B,C and D are known.

For this model-free approach, this work uses a linear parametric approximation for the approximation of the Q-value function (B.22). The approximation function consists of a set Basis Functions (BFs) $\phi(x, u)$ and a coordinate vector θ,

$$\tilde{Q}(x_k, u_k) = \phi^T(x_k, u_k)\theta, \tag{B.24}$$

where $\phi \in \mathbb{R}^{n_b}$ is a column vector and $\theta \in \mathbb{R}^{n_b}$ with the number of bases $n_b = (m_a + n_a + 1)(m_a + n_a)/2$.

A polynomial architecture is selected over black-box methods such as neural networks for numerical convenience. Although the latter method could provide a better approximation of the non-linearities in a water network, the

level uncertainty increases considerably with these basis making the learning of the optimal parameters more difficult.
The Q-value function structure is available from (B.22) and can be used as a reference for building the polynomial approximation scheme [3].

$$\phi(x_k, u_k) = [x_{1,k}^2, x_{1,k}x_{2,k}, \ldots, x_{m_a,k}^2, x_{m_a,k}u_k, u_k^2]^T \tag{B.25}$$

The vector of BFs (B.25) consist of a finite set of 2^{nd} degree polynomials which are formed with system states and control action.

Remark 1. *The augmented state vector x includes a constant term, in this particular case x_2=c_0. Then, the product in* (B.24) *results in combination of quadratic, linear and constant terms*

$$\tilde{Q}(x_k, u_k) = \theta_1 x_{1,k}^2 + \theta_2 x_{1,k}x_{2,k}, \ldots, \theta_{n_b} u_k^2. \tag{B.26}$$

This provides an equivalent approximation scheme to the one developed in (B.22). *This constant term allows to compact a function with quadratic, linear and constant terms into a quadratic function. This simplification is used in the remainder of the paper for conciseness.*

Subsequently, the control law based on the approximated Q-value function (B.24) can be computed by

$$u_k = \underset{u}{\operatorname{argmin}}\, \tilde{Q}(x_k, u_k) = \underset{u}{\operatorname{argmin}}\, \phi^T(x_k, u_k)\theta \tag{B.27}$$

This yields to an optimal feedback control policy,

$$u_k = \tilde{K}(\theta)x_k, \tag{B.28}$$

which has an affine vector field implicit due to Remark 1, this provides an equivalent control law to (B.23). Alternatively, the approximated Q-function can be rearranged in a matrix form since (B.24) is quadratic with respect to x and u

$$\tilde{Q}(z_k) = z_k^T \tilde{H}(\theta) z_k, \tag{B.29}$$

Likewise, Q-value function (B.22) and the approximated (B.29) share the same structure. Since one of the BF is a constant, the approximated Q-value expression can be compacted in a quadratic form.

3.3 Parameter Update

The coordinate vector θ is a variable initially unknown, and therefore it must be learned using past experiences. For this purpose, the optimal value of the coordinate vector θ is determined in real time using Temporal Difference (TD) methods [12]. These methods aim at reducing the approximation error

which is defined with the Bellman equation and the Q-value approximator (B.24)

$$e_k = \rho(x_k, u_k) + \gamma\phi^T(x_{k+1}, u'_k)\theta - \phi^T(x_k, u_k)\theta \tag{B.30}$$

then, (B.16) can be formulated similarly for the temporal difference error.

$$0 = \min_{u_k}(\rho(x_k, u_k) + \gamma\phi^T(x_{k+1}, u'_k)\theta - \phi^T(x_k, u_k)\theta) \tag{B.31}$$

This problem is described as a contraction map of the projected Bellman equation in [3], as such it can be solved by successive approximations methods such as policy iteration algorithms. Then, a learning rate parameter α, which weights the new experiences versus past experiences, is introduced as follows

$$\phi^T(x_k, u_k)\theta_{k+1} = (1-\alpha)\phi^T(x_k, u_k)\theta_k + \alpha\left[\rho(x_k, u_k) + \gamma\phi^T(x_{k+1}, u'_k)\theta_k\right], \tag{B.32}$$

where $0 < \alpha < 1$ is a constant learning rate. Finally, (B.32) is solved by executing a Least Squares Temporal Difference (LSTD) algorithm [8]. This algorithm is an online learning method which recursively applies a policy and collects data, until a batch of n_s samples is completed. Then, by introducing the collected data into (B.32), the update law becomes,

$$\Phi_l^T\theta_{l+1} = (1-\alpha)\Phi_l^T\theta_l + \alpha\left[J_l + \gamma\Phi'^T_l\theta_l\right], \tag{B.33}$$

where l is the iteration number, $\Phi_l = [\phi_l, \ldots, \phi_{l+n_s}]$ and $J_l = [\rho_l, \ldots, \rho_{l+n_s}]^T$ are a matrix and a vector generated by evaluating the collected data into the polynomial BFs (B.25) and reward functions (B.15) respectively. The optimal solution for θ_{l+1} is calculated by applying Least Squares. This process is repeated until the convergence of the coordinate vector θ is considered satisfactory.

Algorithm 4 LSTD for Q-function.

Input: γ, α, n_s,
Initialisation: $l \leftarrow 0, x_0, \theta_0$ where $\tilde{\pi}(\theta_0)$ must be an admissible policy.
repeat at every iteration k = 0,1,2, ...
  apply $u_k = \tilde{K}(\theta)x_k + \epsilon_k$ and measure x_{k+1}
  $Y_{l_s} \leftarrow \rho(x_k, u_k) + \gamma\tilde{q}(x_{k+1}, \tilde{K}_l x_{k+1})$
  if $k = (l+1)n_s$ **then**
    $\theta_{l+1} \leftarrow (1-\alpha)\theta_l + \alpha(\Phi_l\Phi_l^T)^{-1}\Phi_l Y_l$
    $\tilde{\pi}(\theta_{l+1}, x) \leftarrow \text{argmin}_u\, \phi(x, u)^T\theta_{l+1}$
    $l \leftarrow l+1$
  end if
until $\|\tilde{K}_{l+1} - \tilde{K}_l\| < \delta$

Note that a persistent excitation ϵ_k is introduced in the control signal. This perturbation is not a system property but it is part of the experiment design. Therefore, a reasonable generation of persistent excitation signals must be considered to have a proper balance between exploration and exploitation. The experiment aims to generate data batches with sufficient identifiability without compromising the control objectives. In this study case, the system dynamics include both fast and slow dynamics. This stiff system requires both high and low frequency noise for the adequate perturbation. The states corresponding to the FS, which are not affected by the control action, are artificially excited with noise.

4 Numerical & experimental results

The proposed RL controller is validated first using a numerical simulation. The network used for the simulation and test-bed is representing a WDN with elevated reservoir, see Figure B.1.

4.1 Numerical results

The network model used in the simulation is a non-linear model of the WDN emulated in the laboratory test bed described in [14]. In this simulation the only flow measured is the input flow at the pumping station q_1, pressure sensors are placed at the supply p_1 and at the elevated reservoir node p_n. The water consumption profile is simulated by the output signal of a Fourier Series of 2^{nd} order (B.9).
The top graph in Figure B.5 shows that the tank level is regulated to the reference after a learning transient. During this transient the system is controlled with a bad initial policy and the level reaches a low value near the operational boundaries. The middle graph shows the input flow and the total consumed water. Once the algorithm learns an optimal policy, the control inflow follows the periodic trajectory described by the demand flow. Note that at time = 47 days, when the learning phase is completed, the persistent excitation is no longer applied, showing a smoother regulation. The top and middle graphs in Figure B.6 show the convergence of the control policy. The bottom graph shows the Q-values reaching its minimum during the operation once the approximation parameters converge to its optimal value.
Recall that the objective of the controller is to reduce the energy usage of the pumping station and to maintain the tank level within safe levels without a model. The validation of the controller shows promising results on a simulation framework, the proposed linear architecture provides a satisfactory trade-off between numerical accuracy and performance.

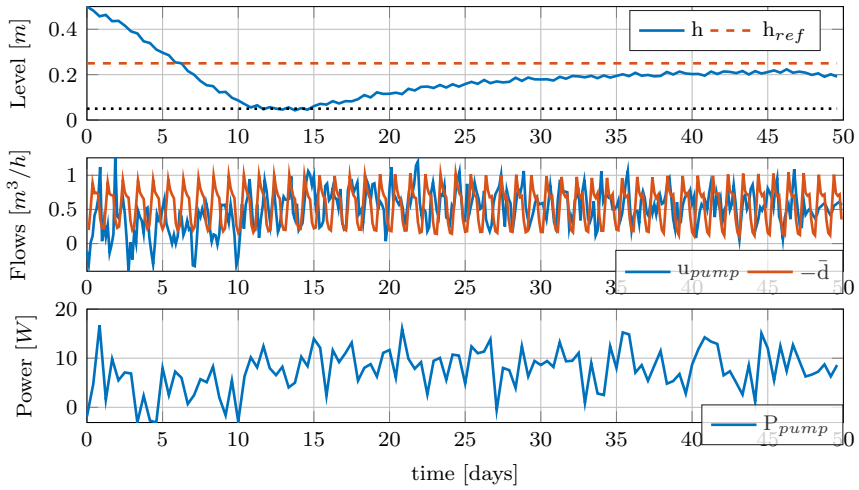


Figure B.5: Simulation results. Top: Tank level and level reference Middle: Network flows control input and disturbance Bottom: Pump power consumption

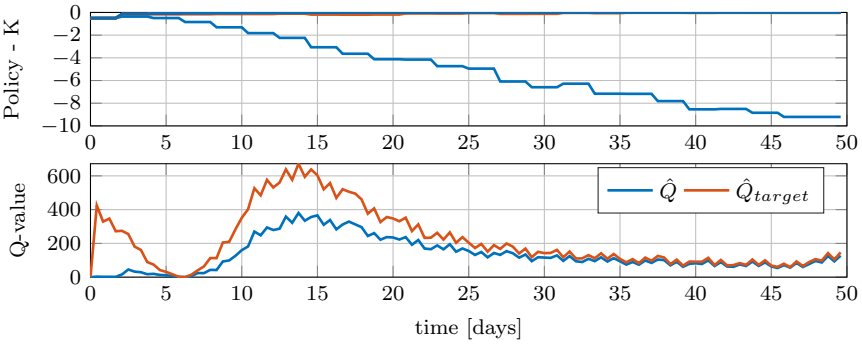


Figure B.6: Simulation results. Top: Control policy gain. Bottom: Q-values target and current

4.2 Experimental results

The controller is validated in the test bed described in [14], and it is implemented with the same control objectives as the simulation. However, in this case an admissible policy is obtained in a simulation framework. By having a preliminary policy with certain knowledge of the system dynamics, the controller reduces the exploration of state space areas where the system has physical limitations. In Figure B.7 (top), the level is kept around a constant value while the controlled input flow compensates the periodic water consumption from the two end users (bottom). Figure B.8 shows the convergence of the approximation parameters and control gains which stabilise to

near constant values.

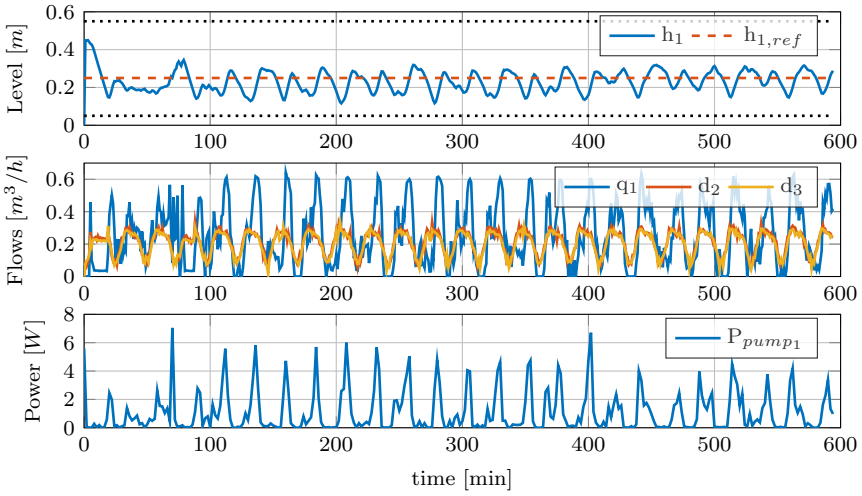


Figure B.7: Experimental Results Top: Tank level and level reference Middle: Network flows: control input q_1 and disturbances d_2, d_3. Bottom: Pump power consumption

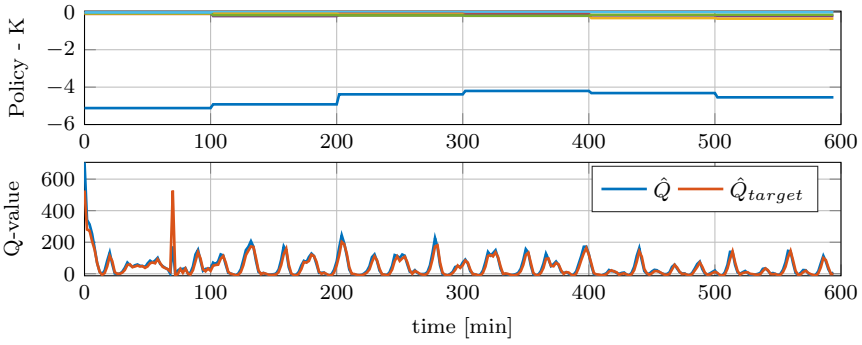


Figure B.8: Experimental Results. Top: Control policy gain. Bottom: Q-values target and current

5 Conclusion

This paper proposes a novel formulation of the RL control for linear systems with periodic disturbances. An augmented state space model is developed which includes both the dynamics of a WDN with storage unit and the periodic disturbances which are approximated with a FS. This paper uses a model-free control approach based on Q-learning. The augmented linear

model is used as a reference to build the basis functions that approximates the Q-value function.
The controller is tested in a simulation and laboratory setup, the performance of the WDN controller is satisfactory for a learning-based management. However, when tested against a real system, this method has more difficulties to find an optimal policy. The laboratory test bed used for the validation incorporates multiple features of a real WDN, such as sensor and actuator dynamics and saturation, communication delays or stochastic demands. Some of these factors are not modelled, and not considered in the approximation scheme proposed in (B.24). This mismatch between the function approximator and reality hinders the learning of an optimal policy. In order to capture these elements in the approximation, a non-linear approximation architecture defined by neural networks could be considered. However, a more complex approximation architecture may lead to longer learning periods. The proposed control method only uses the reference tracking as soft-constraint. In the future, the applicability, learning time and robustness of this methods can be improved by including input and state constraints which reduce the risk of operating in unsafe areas.

References

[1] K. Balla, T. Nørgaard Jensen, J. Bendtsen, and C. Kallesøe, "Model predictive control using linearized radial basis function neural models for water distribution networks," in *IEEE Conference on Control Technology and Applications (CCTA)*.

[2] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in neural information processing systems*, 2017.

[3] D. P. Bertsekas, *Approximate dynamic programming*, fourth edition ed., ser. Dynamic programming and optimal control. Athena Scientific, 2012, no. Volume 2.

[4] G. Cembrano, G. Wells, J. Quevedo, R. Pérez, and R. Argelaguet, "Optimal control of a water distribution network in a supervisory control system," *Control Engineering Practice*, vol. 8, no. 10, 2000.

[5] J. A. Farrell and M. M. Polycarpou, *Adaptive Approximation Based Control: Unifying Neural, Fuzzy and Traditional Adaptive Approximation Approaches*. John Wiley & Sons, Inc., 2006. [Online]. Available: http://doi.wiley.com/10.1002/0471781819

[6] H. Kano, "Existence condition of positive-definite solutions for algebraic matrix riccati equations," *Automatica*, vol. 23, no. 3, 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0005109887900136

[7] G. Konidaris, S. Osentoski, and P. Thomas, "Value function approximation in reinforcement learning using the fourier basis," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011.

[8] M. G. Lagoudakis and R. Parr, "Least-Squares Policy Iteration," *Journal of Machine Learning Research*, p. 43, 2003.

[9] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[10] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement Learning for Partially Observable Dynamic Processes: Adaptive Dynamic Programming Using Measured Output Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, 2011.

[11] T. Maschler and D. A. Savic, "Simplification of water supply network models through linearisation," University of Exeter, Tech. Rep., 1999.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, second edition ed., ser. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press, 2018.

[13] P. K. Swamee and A. K. Sharma, *Design of Water Supply Pipe Networks*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2008. [Online]. Available: http://doi.wiley.com/10.1002/9780470225059

[14] J. Val, R. Wisniewski, and C. Kallesøe, "Optimal control for water distribution networks with unknown dynamics," ser. IFAC World Congress, 2020.

[15] Y. Wang, C. Ocampo-Martinez, and V. Puig, "Robust model predictive control based on gaussian processes: Application to drinking water networks," in *2015 European Control Conference (ECC)*.

[16] Y. Wang, V. Puig, and G. Cembrano, "Non-linear economic model predictive control of water distribution networks," *Journal of Process Control*, 2017.

Paper C

Real-Time Reinforcement Learning Control in Poor Experimental Conditions

Jorge Val, Rafał Wisniewski and Carsten S. Kallesøe.

The paper has been published in the
2021 European Control Conference (ECC), pp. 126-131, 2021.

The layout has been revised.

Abstract

Reinforcement Learning (RL) is a widely used method for solving optimal problems without system knowledge. However, the use of RL for control of industrial applications is still reduced. One of the reasons for limited applicability of RL in this field is the difficulty of learning the system behaviour under poor experimental conditions. This paper proposes two methods to cope with scenarios where the data collected is not contributing to the learning in linear systems. The first method identifies the periods where the learning is not efficient and pauses the policy update, the second method applies a reduction of the approximation space to continue with the learning. The proposed methods are validated in a simulation environment of a water distribution network. Both methods show similar performance and provide a reliable operation during steady state or poor experimental conditions.

1 Introduction

Reinforcement Learning (RL) is a widely used method to solve optimisation problems when an environment is unknown. The use of RL methods is also extended to control in robotics and other industrial applications. RL can learn the control policy of complex systems where the development of a model is tedious or it is simply not possible [13], [18]. A great advantage of the use of RL in industrial applications is the capacity to adapt to changes without the need of calibrations. For instance, large scale systems such as water networks are time-variant systems which require of continuous learning to adapt to different operating conditions such as changes in consumption profile, inclusion of new distribution areas or ageing of the pipe network. Nevertheless, the deployment of RL methods in industrial applications is a challenging task. The reasons are manifold: the training of these algorithms in real applications is significantly more complex than the training in a virtual domain with a great number of episodes, high dimensional state-action spaces [8], [4], safety constraints [15]. The learning of an optimal policy using RL methods comprises of two phases: exploitation and exploration [21]. The exploitation phase utilises the knowledge of the system to achieve an optimal operation and exploration phase expands the knowledge of the system and the Q-value function. This function is an indicator of the system's performance based on operation costs or rewards. The data collected during the operation is used to identify the Q-value function and compute the best policy for a particular system. The identification of the Q-value function in a real-time operation is not a simple task since the controller must achieve certain control objectives while exploring out of the optimal regions. A good exploration strategy that balances the trade-off between exploration and exploitation must be designed in order to achieve the control objectives while

learning. The exploration strategy can be defined in different ways, some methods use a random switch between a greedy policy and a random action [21], other methods require persistent excitation similarly to adaptive controllers [17], [12]. The identification of the function approximation requires adequate experimental conditions. Adequate experimental conditions are a good sampling time and a persistent excitation signal that deviates the system from the nominal operation in order to explore new areas. Solutions to deal with the lack of system excitation are presented in [16], [1] with experience replay technique. Another issue with the identification process is that the approximation scheme used for the estimation is not adequate for the collected data. The Bierman-Thornton UD factorization provides a greater numerical stability to Kalman filters [5], other studies propose a reduction of the Q-value approximation scheme by selecting the most relevant features (basis) for the application [20], [2].
Driven by these identification challenges during closed-loop operation and learning, this paper presents a method for identifying the learning efficiency and cope with poor experimental conditions. The lack of information in the data is due to the system operating in steady state, hence there are only slight variations in its operating points or the system is not sufficiently perturbed, therefore the learning algorithm can not correctly identify the Q-value function. Although an adequate the sampling time has a great contribution for having good experimental conditions, this work considers a fixed sampling rate restricted by the application domain. The proposed method consist of detecting the learning periods where the information obtained for the estimation is poor. Additionally, this paper presents two solutions to deal with these scenarios: a conservative strategy where the learning is paused until the collected data batch contains sufficient information and a greedy strategy where the learning continues. For the second strategy, a reduction of the approximation vector space is formulated such that the identification process uses only the dominant data, avoiding oversampling.
The remainder of this paper is organised as follows. In Section 2 some concepts of estimation are recapitulated. In Section 3 the water network model is described. In Section 4 the control design and algorithms are presented. In Section 5 the simulation results are shown. In Section 6, the contributions of the work and ideas for the future work are discussed.

2 Preliminaries

In this section, some concepts of system identification and statistics are introduced in order to clarify the criteria that this works uses to evaluate the learning performance.

2.1 Fisher information matrix

Least Squares (LS) and Maximum Likelihood Estimation (MLE) are equivalent methods for estimating parameters for an i.i.d data with Gaussian distribution. The MLE is a method for estimating the parameters of a probability distribution by maximising a likelihood function [9]. This paper uses MLE over LS because of its efficiency analysis that allows assessing the consistency of the parameter estimation for an observed data batch. The Cramér-Rao bound indicates the lower bound on the covariance matrix of unbiased estimates, this can be used to measure the statistical efficiency of an unbiased estimation [19].

Lemma 1. *Let θ be an unknown coordinate vector which is estimated from m independent samples or measurements of Φ [19]. Let Φ be a stochastic vector-valued variable, the distribution of which depends on an unknown vector θ . Let $L(\Phi, \theta)$ denote the likelihood function, and let $\hat{\theta} = \hat{\theta}(\Phi)$ be an arbitrary unbiased estimate of θ determined from Φ. Then, the bound for the covariance is expressed as follows*

$$cov(\hat{\theta}) \geq I(\theta)^{-1} \tag{C.1}$$

where $I(\theta)$ is the Fisher information matrix, which is defined by

$$I(\theta) = -\left[E\frac{\delta^2 logL}{\delta\theta^2}\right]. \tag{C.2}$$

This matrix $I(\theta)$ is an indicator of the amount of information that the matrix Φ, built with the measurements of a random variable, contains about the unknown vector θ.

3 System model

This paper validates the proposed algorithm in the management of a Water Distribution Network (WDN) that has an elevated reservoir and the pipe network is defined by a ring topology.

3.1 Water network model

The water network model used in this work is similar to the model presented in [11]. This is a low dimension model of a network that simplifies the end-users by aggregating geographically close end-users into single nodes. The water inflow at the pumping station is assumed to be locally controlled and its dynamics are considerably faster than the elevated reservoir dynamics, thus the dominant dynamics of the system is given by the elevated reservoir. Figure C.1 shows a standard ring topology network where the demand from

the city district is represented by multiple end-user demands d_j connected to the main pipes, the inflow from the pumping station is denoted with q_1, the flow to the tank is denoted with d_{n+1} and outflow to the pressure zone 2 is denoted with d_{n+2}.

Due to the mass conservation in the water network, the relation of the flows

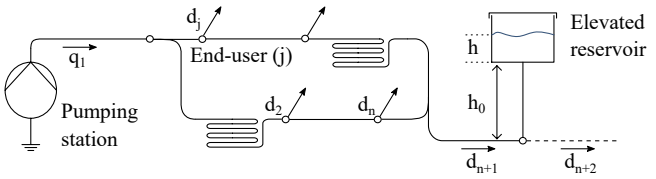


Figure C.1: Scheme of the main elements in a WDN with a pumping station, a pipe network, multiple end-users and an elevated reservoir.

in the network is represented as

$$q_1(t) + d_{n+1}(t) = -\sum_{j=1}^{n_d} d_j(t), \tag{C.3}$$

where n_d is the total number of end-users connected to the network and the total water consumption is denoted by $\bar{d} = -\sum_{j=1}^{n_d} d_j$. The elevated reservoir dynamics is given by the following expression

$$A_{er}\dot{h} = d_{n+1}(t) - d_{n+2}(t), \tag{C.4}$$

where A_{er} is the cross-sectional area of the elevated reservoir.

4 Control

The management of WDNs must guarantee the supply of water to the end-users with sufficient pressure. This paper presents a cost function whose main objective is to maintain the network pressure at a certain level. This objective is achieved by regulating the water level of the elevated reservoir. Additionally, by having a volume of water stored in the elevated reservoir, the supply system becomes more robust against future stochastic demands. This work considers a constant demand profile as disturbance. This profile corresponds to the mean water consumption of a pressure zone. This work uses the state space model structure and problem formulation presented in [22] and [14] to validate the learning efficiency of Q-learning algorithms.

4.1 State space augmentation

This model consists of an extension of the WDN model (C.4) and includes a trajectory reference state r and an integral error ξ. The reference trajectory indicates the nominal level of the tank and it is defined by a linear function

$$\dot{r} = L_c r, \tag{C.5}$$

where $r \in \mathbb{R}$, in this work the trajectory is a constant, $L_c = 0$. Then, defining the integral error

$$\dot{\xi} = h(t) - r(t). \tag{C.6}$$

The main purpose of this integral error is to have an integral action which compensates the constant disturbances corresponding to the mean water demand of a city district. By combining equations (C.4), (C.5) and (C.6) the augmented model is built.

$$\begin{bmatrix} \dot{h} \\ \dot{r} \\ \dot{\xi} \end{bmatrix} = \begin{bmatrix} A_c & 0 & 0 \\ 0 & L_c & 0 \\ C_c & -I & 0 \end{bmatrix} \begin{bmatrix} h \\ r \\ \xi \end{bmatrix} + \begin{bmatrix} B_c \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} u \end{bmatrix} + \begin{bmatrix} W_c \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} d \end{bmatrix} \tag{C.7}$$

where $h \in \mathbb{R}$ represents the tank level, $u \in \mathbb{R}$ the controlled inflow q_1 and $d \in \mathbb{R}$ represents all the system disturbances, the district demands ($d = \bar{d} + d_{n+2}$). Finally, expressing the state space representation (C.7) in a more compact form for discrete time

$$\begin{aligned} x_{k+1} &= A_e x_k + B_e u_k + W_e d_k \\ y_k &= C_e x_k, \end{aligned} \tag{C.8}$$

where $x_k \in \mathbb{R}^{m_a}$, $u_k \in \mathbb{R}^{n_a}$, in this case $x = \begin{bmatrix} h, & r, & \xi \end{bmatrix}^T$ is the augmented state vector, with A_e,B_e,W_e and C_e constant matrices with compatible dimensions. The control law for this system is given by the following linear controller

$$u_k = \pi(x_k) = -K x_k \tag{C.9}$$

This model formulation assumes that the full-state feedback is available, since tank level h_k is measured and the reference trajectory r is known.

4.2 Problem formulation

The application objective is to provide an adequate pressure management in the network by regulating the tank level. Therefore, the following cost function is defined

$$V(x_k) = \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} (x_i^T Q_e x_i + u_i^T R u_i) \tag{C.10}$$

where Q_e and R are weight matrices penalising tracking error and control action, and $0 < \gamma < 1$ represents a discount factor that reduces the weight of the cost obtained further in the future, this means that γ bounds the cost function (C.10) from accumulating non-zero rewards when time goes to infinity. Subsequently, the instant reward is denoted as

$$\rho(x_k, u_k) = \frac{1}{2} x_k^T Q_e x_k + u_k^T R u_k. \tag{C.11}$$

By using the previous problem formulation and Bellman's optimality principle, the value can be determined using the HJB equation as follows

$$V^*(x_k) = \min_u (\rho(x_k, u_k) + \gamma V^*(x_{k+1})) \tag{C.12}$$

with the notation $(\cdot^*)$ representing the optimal value. By using the Q-value formulation proposed in [14] and then by defining the vector $z_k = [x_k, u_k]^T$, the expression (C.12) is compacted in a quadratic form

$$Q(z_k) = z_k^T M z_k. \tag{C.13}$$

Then, the optimal control policy for (C.13) is determined as follows

$$u_k^* = \underset{u}{\operatorname{argmin}}\, Q(x_k, u_k) = K x_k \tag{C.14}$$

4.3 Function Approximation structure

The optimal control solution presented in the previous subsection is developed with the knowledge of the system dynamics, system matrices A_e, B_e and C_e. This paper studies a model-free control approach where the system dynamics are unknown, thus the Q-value function cannot be developed with the previous methodology. Nevertheless, this paper uses the same function structure to define the approximation scheme. This approximation scheme consists of a set of Basis Functions (BFs) ϕ and a coordinate vector θ,

$$\hat{Q}(x_k, u_k) = \phi^T(x_k, u_k)\theta, \tag{C.15}$$

where $\phi \in \mathbb{R}^{n_b}$ is a column vector and $\theta \in \mathbb{R}^{n_b}$ with the number of bases $n_b = (m_a + n_a + 1)(m_a + n_a)/2$.
A polynomial architecture is selected to describe the system performance based on the linear model developed in Section 3. Although this approximation scheme introduces certain error, since WDNs are non-linear, a complex approximation considerably increases the learning uncertainty. For this reason, a simple approximation is chosen for achieving a fast identification by learning only the elementary system dynamics.

The Q-value function structure is available from (C.13) and can be used as a reference for building the polynomial approximation scheme [3].

$$\phi(x_k, u_k) = [x_{1,k}^2, x_{1,k}x_{2,k}, \ldots, x_{m_a,k}^2, x_{m_a,k}u_k, u_k^2]^T \tag{C.16}$$

The vector of BFs (C.16) consists of a finite set of 2^{nd} degree polynomials built with combinations of x and u. Subsequently, the optimal control law for the approximation of the Q-value function is given by

$$u_k = \underset{u}{\operatorname{argmin}}\, \hat{Q}(x_k, u_k) = \underset{u}{\operatorname{argmin}}\, \phi^T(x_k, u_k)\theta \tag{C.17}$$

This yields to an optimal feedback control policy,

$$u_k = \hat{K}(\theta)x_k, \tag{C.18}$$

By rearranging (C.15) in a quadratic form with respect to z, the approximated Q-value function shares the same structure as the Q-value in (C.13)

$$\hat{Q}(z_k) = z_k^T \hat{H}(\theta) z_k, \tag{C.19}$$

4.4 Parameter Update

The coordinate vector θ is initially unknown and has to be learned iteratively using previously measured data. A Temporal Difference (TD) algorithm is used for solving in real-time the approximation of the Q-value function. The algorithm consists of minimising the approximation error between different iterations (C.20), then introducing the Q-value approximation (C.15) the TD update law is defined as follows

$$\phi^T(x_k, u_k)\theta_{k+1} = (1-\alpha)\phi^T(x_k, u_k)\theta_k + \alpha\left[\rho(x_k, u_k) + \gamma\phi^T(x_{k+1}, u_k')\theta_k\right] \tag{C.20}$$

where $0 < \alpha < 1$ is a constant learning rate. The Q-value function equation (C.20) is solved by executing the Least Squares Temporal Difference. Both algorithms 5 and 6 use the same principle for identifying the coordinate vector parameters. This method applies a control action, collects data measurements and the rewards until a batch with m samples is completed. By evaluating (C.20) with the batch of data, the batch update law becomes,

$$\Phi_l^T\theta_{l+1} = (1-\alpha)\Phi_l^T\theta_l + \alpha\left[J_l + \gamma\Phi_l'^T\theta_l\right] \tag{C.21}$$

where l is the iteration number, $\Phi_l = [\phi_l, \ldots, \phi_{l+m}]$ and $J_l = [\rho_l, \ldots, \rho_{l+m}]^T$ are a matrix an a vector generated by evaluating the collected data into the polynomial BFs (C.16) and reward functions (C.11) respectively.

4.5 Learning Efficiency Analysis

As presented in Section 2 the Fisher information matrix is used to analyse the identification efficiency [19]. This section presents the development of this analysis applied to the Q-value estimation, where e is the TD approximation error. This error is assumed to have Gaussian distribution with the variance $var(e) = var(\hat{Q}_{l+1} - \hat{Q}_l)$. Then, the Fisher information matrix (C.2) is computed. Remark that, this matrix is an indicator of the amount of information that a batch of data Φ carries about a set of parameters θ.
According to [6] there are several criteria to analyse the data and design an optimal experimental condition. One of these criteria is the E-optimalilty, which aims to maximise the eigenvalues of the Fisher matrix in order to extract the maximum information from the collected data. This paper uses the minimum eigenvalue of the matrix I as indicator of the worst case scenario. This means that this value is expected to be low when the collected data has low variations and cannot be used for a proper parameter estimation. In Algorithm 5, the threshold for low information I_{low} is set when the rank of the covariance matrix $\Phi_l \Phi_l^T$ is singular or close to singular.

$$\lambda_{low} = \min \lambda(I_{low}) \tag{C.22}$$

where λ_{low} is a threshold for indicating low estimation efficiency.

Algorithm 5 LSTD for Q-function using Fisher information.

Input: γ, α, m,
Initialisation: $l \leftarrow 0, x_0, \theta_0$ where $\hat{\pi}(\theta_0)$ must be an admissible policy.
repeat at every iteration k = 0,1,2, ...
  apply $u_k = \hat{K}(\theta)x_k + \epsilon_k$ and measure x_{k+1}
  $Y_{l_s} \leftarrow \rho(x_k, u_k) + \gamma \hat{Q}(x_{k+1}, \hat{K}_l x_{k+1})$
  if $k = (l+1)m$ **then**
    if $\min \lambda(I_l) > \lambda_{low}$ **then**
      $\theta_{l+1} \leftarrow (1-\alpha)\theta_l + \alpha(\Phi_l \Phi_l^T)^{-1} \Phi_l Y_l$
    else
      $\theta_{l+1} \leftarrow \theta_l$
    end if
    $\hat{\pi}(\theta_{l+1}, x) \leftarrow \text{argmin}_u \, \phi(x,u)^T \theta_{l+1}$
    $l \leftarrow l+1$
  end if
until

4.6 Singular Value Decomposition and BFs selection

The previous subsection shows how data batches with little or redundant information can affect the identification process. In linear algebra, Singular Value Decomposition (SVD) is a widely used technique for data processing. This work uses this method to segregate the collected data into two parts: one containing high and the other containing low amount of system information. The matrix partition is developed firstly by approximating the data batch Φ matrix using SVD in its compact form.

$$\Phi_l = U\Sigma V^T, \tag{C.23}$$

where $\Phi_l \in \mathbb{R}^{n_b \times m}$ and $n_b \leq m$ is a matrix with the collected data, with n_b the number of features (BFs) and m the number of collected samples, $U \in \mathbb{R}^{n_b \times n_b}$ and $V \in \mathbb{R}^{m \times n_b}$ are unitary left and right singular matrices and $\Sigma \in \mathbb{R}^{n_b \times n_b}$ is a diagonal matrix with weights ordered by importance. Then, the partition of the data batch matrix Φ in two parts is expressed as follows.

$$U = [\bar{U}, \underline{U}], \quad \Sigma = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \underline{\Sigma} \end{bmatrix}, \quad V^T = \begin{bmatrix} \bar{V}^T \\ \underline{V}^T \end{bmatrix} \tag{C.24}$$

where $\bar{U} \in \mathbb{R}^{m \times p}$, $\bar{\Sigma} \in \mathbb{R}^{p \times p}$ and $\bar{V} \in \mathbb{R}^{p \times n}$. The sub-index notations $\bar{(\cdot)}$ and $\underline{(\cdot)}$ represent high and low amount of system information respectively. The singular matrix Σ is hierarchically organised, thus the first p elements contain most of the information. The value of p can be determined in several ways [10], [7]. In Algorithm 6, the rank of the collected data $p = Rank(\Phi_l \Phi_l^T)$ is used as a reference to create the matrix partition. Then, by substituting the SVD approximation into (C.21), the following linear transformation is deducted.

$$V\Sigma U^T \theta_{l+1} = (1-\alpha) V\Sigma U^T \theta_l + \alpha \mathrm{Y}_l \tag{C.25}$$

where $\mathrm{Y}_l = J_l + \gamma \Phi'^T_l \theta_l$. By rearranging (C.25) with the (C.24), the partitioned matrix is expressed in the SVD sub-spaces,

$$\begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \underline{\Sigma} \end{bmatrix} \begin{bmatrix} \bar{\theta}_{l+1} \\ \underline{\theta}_{l+1} \end{bmatrix} = (1-\alpha) \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \underline{\Sigma} \end{bmatrix} \begin{bmatrix} \bar{\theta}_l \\ \underline{\theta}_l \end{bmatrix} + \alpha \begin{bmatrix} \bar{V}^T \\ \underline{V}^T \end{bmatrix} \mathrm{Y}_l \tag{C.26}$$

where $U^T\theta = \begin{bmatrix} \bar{\theta} & \underline{\theta} \end{bmatrix}^T$. In this way the parameter identification is computed separately, the upper partition, with $\bar{\theta}$, is updated with standard Temporal Difference (C.20) while the lower partition, with $\underline{\theta}$, is discarded and its parameters are not updated. The partitioned update law is shown in lines *10-11* on Algorithm 6. Note that this method introduces an additional approximation error, hence there is a lower limit where the approximation can no longer be reduced.

Algorithm 6 LS-TD for Q-function using SVD.

Input: γ, α, n_s,
Initialisation: $l \leftarrow 0, x_0, \theta_0$ where $\hat{\pi}(\theta_0)$ must be an admissible policy.
repeat at every iteration k = 0,1,2, ...
  apply $u_k = \hat{K}(\theta)x_k + \epsilon_k$ and measure x_{k+1}
  $Y_{l_s} \leftarrow \rho(x_k, u_k) + \gamma\hat{Q}(x_{k+1}, \hat{K}_l x_{k+1})$
  if $k = (l+1)n_s$ **then**
    if $inv(\Phi_l\Phi_l^T)$ == TRUE **then**
      $\theta_{l+1} \leftarrow (1-\alpha)\theta_l + \alpha(\Phi_l\Phi_l^T)^{-1}\Phi_l Y_l$
    else
      $\bar{\theta}_{l+1} \leftarrow (1-\alpha)\bar{\theta}_l + \alpha\bar{\Sigma}^{-1}\bar{V}^T Y_l$
      $\theta_{l+1} \leftarrow U\left[\bar{\theta}_{l+1}; \underline{\theta}_l\right]$
    end if
    $\hat{\pi}(\theta_{l+1}, x) \leftarrow \text{argmin}_u\, \phi(x,u)^T\theta_{l+1}$
    $l \leftarrow l+1$
  end if
until

5 Results

A simulation environment is developed in order to validate the proposed control algorithm against different scenarios. The study case of this simulation reproduces a WDN of a small water utility. In particular, this simulation is constructed with the network information provided by Bjerringbro's water utility, a small urban district in Denmark, the structure of this network is illustrated in Figure C.2. This distribution area is divided into pressure zone 1 and pressure zone 2, the total consumption for each district is represented with $\bar{d}$ and d_{n+2} respectively.

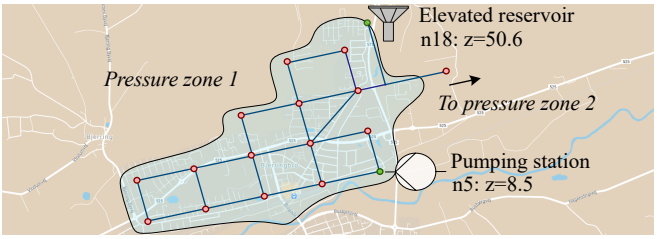


Figure C.2: Illustration of a simplified city district from Bjerringbro (Denmark). The pipe network is represented with blue lines, the end-users with red dots, a pumping station, an elevated reservoir with green dots.

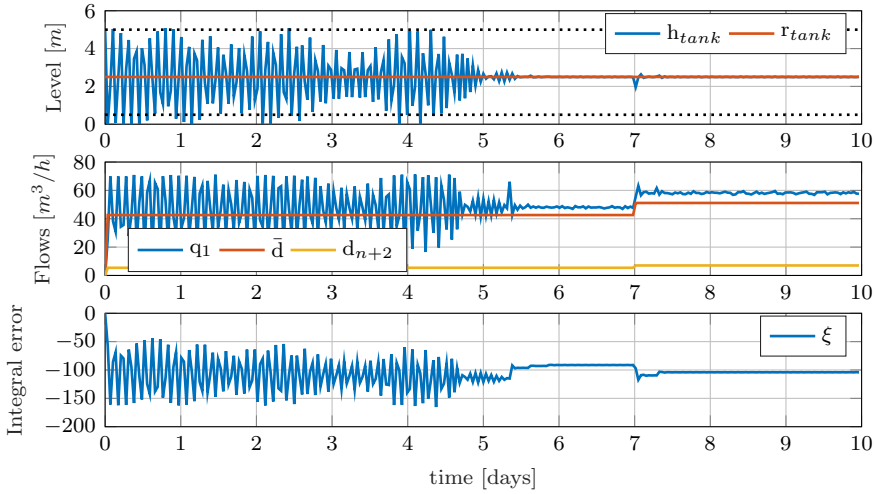


Figure C.3: Algorithm 5 simulation. Top: shows the tank level and the reference level. Middle: Controlled input flow and demand flows. Bottom: Integral error state.

5.1 Numerical results

In graph Figure C.3 the regulation of the system states is shown, a long oscillatory transient is observed during the first 5 days where the system is regulated with a non-optimal control policy. The middle graph in Figure C.3 shows a noisy control action, water inflow q_1, compared with the disturbances, water demands $\bar{d}$ and d_{n+2}. A bad policy on the integral error also increases the effect of the transient oscillations, similarly to a poorly calibrated PI controller. During the same simulation period, Figure C.4 shows the convergence of the approximation parameters θ, thus the learning of a better control policy $\hat{K}$, when the parameters converge to a nearly constant value, the tank level and control action remain in steady state. During nominal operation, some of the measurements that are used for the approximation, such as tank level or integral error, are nearly constant leading to collected data batches with poor condition. Red marks in Figure C.4 (top) show these periods where the policy is not updated (middle). Once the operating point, where the learning efficiency is poor, is reached, the parameter update is paused. The algorithm evaluates the efficiency of the identification from the collected data based on the Fisher Information matrix. The graph in Figure C.4 (bottom) shows the minimum eigenvalue of the Fisher information matrix (C.2) that is computed each batch update of the Q-value function (C.15). During the first part of the simulation, where the algorithm is learning a good policy, the eigenvalues are high, once the parameters converge and the system operates in steady state a decay of the values in time is observed.

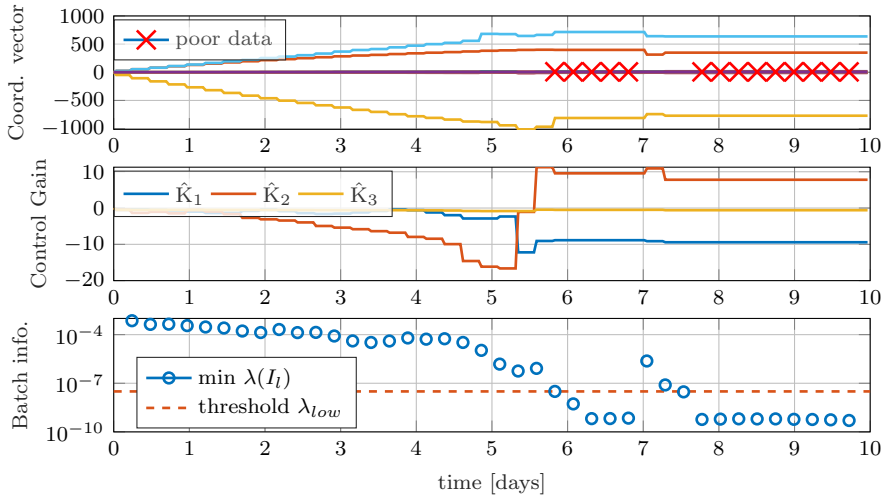


Figure C.4: Algorithm 5 simulation. Top: Coordinate vector of the Q-value function approximator. Middle: Control policy. Bottom: Batch information measured with the minimum eigenvalue of the Fisher matrix.

The eigenvalue remains below the threshold until a change in the operation is introduced ($t = 7days$), the data batches carry more new information and the learning is reactivated.

The simulation results with Algorithm 6 show similar results as the previous simulation. Although the control gain is slightly different, the learned policy does not have an impact on the system performance.

6 Conclusion

Large scale or complex systems require of *a priori* knowledge of the system to build an adequate approximation scheme for the application. By increasing the complexity of the approximation structures such as neural network or polynomial approximation, the algorithm can achieve a more accurate approximation of the environment. However, sometimes complex approximation structures penalise the learning efficiency leading to numerical issues or long training periods.

This paper proposes two solutions to extend the applicability of RL methods in real-time control systems. The two methods are validated in a simulation framework that reproduces a WDN and scenarios with poor data. This work copes with scenarios where the learning is limited by the lack of excitation of the collected signals. Additionally, a reduction of the approximation scheme is proposed that reduces the approximation scheme and selects only the relevant BFs or features. This method, with a low dimensional approximation

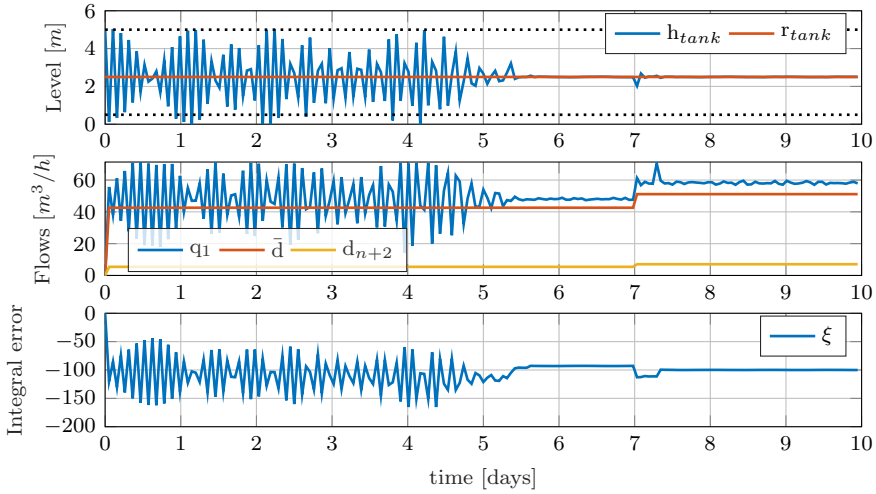


Figure C.5: Algorithm 6 simulation. Top: shows the tank level and the reference level. Middle: Controlled input flow and demand flows. Bottom: Integral error state.

model, increases the learning efficiency and numerical robustness by using only relevant data in steady state periods. However, the reduction of the approximation subspace narrows the flexibility of the controller to incorporate new changes in the system dynamics such as disturbance variation.
This method requires of further validation with experimental data. In the future, the Algorithm 6 can extended by using other methods like Principal Component Analysis, Independent Component Analysis or LASSO regularisation for variable selection and regression accuracy. Another factor to address in the future is the design of an adequate exploration signal. This signal can significantly improve the identification process. However, knowing the characteristics of this signal, gain and frequencies affecting the system is challenging when the system dynamics are unknown.

References

[1] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 2, pp. 201–212, 2011.

[2] B. Behzadian and M. Petrik, "Feature selection by singular value decomposition for reinforcement learning," in *Proceedings of the ICML Prediction and Generative Modeling Workshop*, 2018.

[3] D. P. Bertsekas, *Approximate dynamic programming*, fourth edition ed., ser. Dynamic programming and optimal control. Athena Scientific, 2012, no. Volume 2.

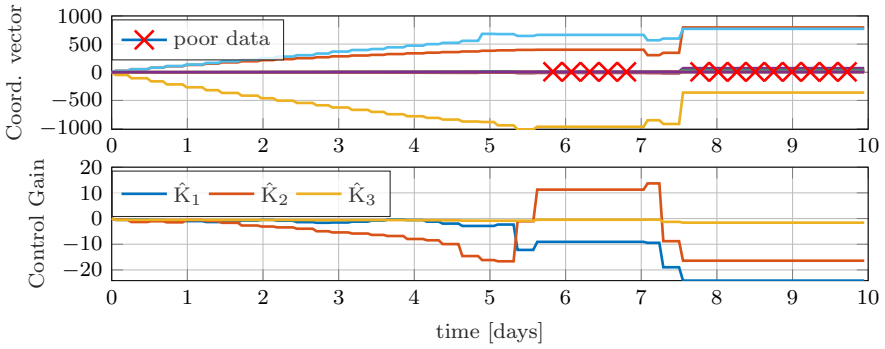


Figure C.6: Algorithm 6 simulation. Top: Coordinate vector of the Q-value function approximator and the $\times$ marks represent batches with poor condition data. Bottom: Control policy.

[4] L. Busoniu, D. Ernst, B. De Schutter, and R. Babuska, "Continuous-state reinforcement learning with fuzzy approximation," 2007.

[5] D. S. Chiu and K. P. O'Keffe, "Bierman-thornton ud filtering for double-differenced carrier phase estimation accounting for full mathematical correlation," in *Proceedings of the 2008 National Technical Meeting of The Institute of Navigation*, 2008, pp. 756–762.

[6] H. Dette and W. J. Studden, "Geometry of e-optimality," *The Annals of Statistics*, pp. 416–433, 1993.

[7] D. L. Donoho and M. Gavish, "The optimal hard threshold for singular values is $4/\sqrt{3}$," 2013.

[8] G. Dulac-Arnold, R. Evans, H. van Hasselt, P. Sunehag, T. Lillicrap, J. Hunt, T. Mann, T. Weber, T. Degris, and B. Coppin, "Deep reinforcement learning in large discrete action spaces," 2015.

[9] D. F. Hendry and B. Nielsen, *Econometric modeling: a likelihood approach*. Princeton University Press, 2007.

[10] P. D. Hoff, "Model averaging and dimension selection for the singular value decomposition," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 674–685, 2007.

[11] C. S. Kallesøe, T. N. Jensen, and J. D. Bendtsen, "Plug-and-Play Model Predictive Control for Water Supply Networks with Storage," *IFAC-PapersOnLine*, vol. 50, no. 1, 2017.

[12] T.-H. Lee and K. S. Narendra, "Robust adaptive control of discrete-time systems using persistent excitation," *Automatica*, vol. 24, no. 6, pp. 781–788, Nov. 1988. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0005109888900544

[13] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[14] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement Learning for Partially Observable Dynamic Processes: Adaptive Dynamic Programming Using Measured Output Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, 2011.

[15] Z. Li, U. Kalabić, and T. Chu, "Safe reinforcement learning: Learning with supervision using a constraint-admissible set," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 6390–6395.

[16] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[17] K. S. Narendra and A. M. Annaswamy, "Persistent excitation in adaptive systems," *International Journal of Control*, vol. 45, no. 1, pp. 127–160, 1987.

[18] J. Shin, T. A. Badgwell, K.-H. Liu, and J. H. Lee, "Reinforcement learning – overview of recent progress and implications for process control," *Computers & Chemical Engineering*, 2019.

[19] T. Söderström and P. Stoica, *System identification*. Prentice-Hall International, 1989.

[20] Z. Song, R. E. Parr, X. Liao, and L. Carin, "Linear feature encoding for reinforcement learning," *Advances in neural information processing systems*, vol. 29, pp. 4224–4232, 2016.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[22] J. Val, R. Wisniewski, and C. Kallesøe, "Optimal control for water distribution networks with unknown dynamics," ser. IFAC World Congress 2020, 2020.

Paper D

Safe Reinforcement Learning Control for Water Distribution Networks

Jorge Val, Rafał Wisniewski and Carsten S. Kallesøe.

The paper has been published in the
2021 IEEE Conference on Control Technology and Applications (CCTA) pp.
1148-1153, 2021.

The layout has been revised.

Abstract

Reinforcement Learning (RL) is an optimal control method for regulating the behaviour of a dynamical system when the system model is unknown. This feature is a strong advantage for controlling systems, such as Water Distribution Networks, where it is difficult to have a reliable model. When learning an optimal policy with RL, the exploration phase implies high degree of uncertainty in the system operation. Large scale infrastructures such as WDN require a robust operation since they cannot afford fails during the operation. This paper presents a model-free control method which provides safety in the operation while learning an optimal policy. This method introduces a policy supervisor block in the control loop which assesses the safety of the learned policy in real-time. The safety verification consists of evaluating the trajectory on a standard linear model. In this model only the fundamental linear dynamics are represented and the system's dimensions do not require to be expressed with high accuracy. If the predicted trajectory violates the boundaries, the supervisor provides a safe control action. Simulation and experimental results prove the applicability of the proposed method.

1 Introduction

Water Distribution Networks (WDNs) are large scale infrastructures that transport drinking water from the waterworks to the urban districts. The operation of these infrastructures is challenged by several factors such as uncertainty in the water demand, operation cost, quality of the water or smoothness in the management [12].

Some studies argue that water consumption uncertainty is one of the governing factors in the WDN management [15]. Modelling the uncertainty in the water demand is a major task, some have used the periodic pattern that the demand describes during the daily operation to model the demand dynamics [10], [5]. The management of WDNs is addressed in [17], [18], [4], these studies provide efficient solutions that regulate the operation of the network, many of them in a Model Predictive Control (MPC) framework. However, these approaches rely on a model to compute the control law. The network models are not always available or their continuous calibration is a laborious task. Robust MPC techniques deal with this issue by considering a system model with uncertainty, but the resulting control policy can be very conservative [16].

This work proposes a model-free control method for the optimal management of WDNs, therefore this controller must satisfy the operation objectives without knowledge of the particular network model and nominal conditions. The use of Reinforcement Learning (RL) in control brings a great advantage in comparison to other methods because of its capacity of providing an op-

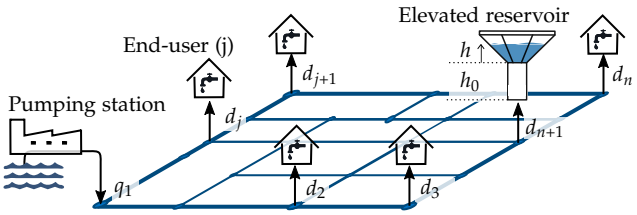


Figure D.1: Illustration of a simplified WDN with a pumping station, an elevated reservoir and multiple end-users.

timal policy without a system model. RL for continuous state-action space relies on the approximation of a value function, neural networks (NNs) are frequently used as approximation architecture to estimate complex systems. Alternative architectures or Basis Functions (BFs) can also be used to approximate the value function. Polynomial basis [3] or Fourier basis [7] are less accurate than NNs but they can be suitable approximations for certain applications. This paper uses an approximation scheme based on Fast Fourier Transform (FFT) to consider the effect of a periodic demand (disturbance) in the learning controller. RL controllers provide an optimal policy according to previous experiences, this means that the initial policy is typically bad. Operating with such policy can be challenging, since a non-optimal policy can easily drive the system to unsafe regions. Such uncertainty in the behaviour during the early operation limits the deployment of this technology in industrial applications like WDN which require consistent robustness in the operation. Ensuring safety during the RL operation without any knowledge of the system is a challenge that has been addressed in in different ways, by using safe optimisation based on an underlying Gaussian models [2], using barrier functions on the reward [11], [22], or including safety in the exploration by combining learning and model-based control [14], [13].
Inspired by the control loop from [9], this paper proposes a supervisor module that prevents the applied control action to drive the system into unsafe regions. This module assesses the safety based on a standard linear system. The control solution is first validated in a simulation framework and subsequently in the Smart Water Infrastructures Laboratory test-bed which emulates a real WDN.
The remainder of this paper is organised as follows. In Section 2 the WDN model is introduced. In Section 3 the control design and algorithms are developed. In Section 4 the simulation, and experimental results are shown. In Section 5, the contributions of the study and ideas for the future work are discussed.

2 System model

In this section a model of a WDN is presented. The WDN used as study case includes the following elements: a pumping station, an elevated reservoir and multiple end-users.

2.1 Water network model

The main network components are represented in Figure D.1, the illustration shows a conventional ring topology network to which the multiple end-users are connected to, the demands are denoted with d_j, the water inflow is denoted with q_1 and is located at the pumping station, the outflow to the elevated reservoir is denoted with d_{n+1}.
This work uses the WDN model presented in [6]. This is a simplified model where the end-users that geographically close are aggregated in a single consumption node. The input flow is ideally regulated at the pumping station. Assuming that mass conservation holds in the pipe network, the relation between the flows in the network is given by

$$q_1(t) + d_{n+1}(t) = -\sum_{j=1}^{n_d} d_j(t), \tag{D.1}$$

where n_d is the total number of end-users connected to the network. Therefore, the total amount of water consumption is $\bar{d} = -\sum_{j=1}^{n_d} d_j$. The elevated reservoir consists of a raised tank storing drinking water and its dynamics are given by

$$A_{er}\dot{h} = d_{n+1}(t), \tag{D.2}$$

where A_{er} is the cross-sectional area of the tank and h is the tank level. The pumping station dynamics are considered much faster than the dynamics of the elevated reservoir. Therefore, the reduced order model includes only the elevated reservoir dynamics. The following expression shows the simplified network model in its discrete-time state space form.

$$h_{k+1} = Ah_k + Bq_k + E\bar{d}_k, \tag{D.3}$$

where $h_k \in \mathbb{R}$ is the system state, $q_k \in \mathbb{R}$ is the controlled input flow and $d_k \in \mathbb{R}$ represents the system disturbances, and A,B and E are constant matrices with compatible dimensions.

2.2 Disturbance model

The disturbance of a WDN is the water consumption or demand of the multiple end-users. The demand of the individual end-users is unknown in advance. However, the signal described by total demand typically follows a

pattern from day to day. This study assumes that the signal described by this stochastic process can be approximated by a Fourier Series (FS) of order N. The signal approximation is performed as follows, let (D.4) be a FS continuous signal

$$\bar{d}(t) = a_0 + \sum_{n=1}^{N}(a_n \cos(\omega_n t) + b_n \sin(\omega_n t)) + w, \tag{D.4}$$

where a_0, a_n and $b_n \in \mathbb{R}$ are the Fourier coefficients, $\omega_n = 2\pi n f_0$ and f_0 represents the fundamental frequency and w is normally distributed and independent noise. In this case study, the frequency f_0 is calculated for a period of a day. By computing the mean of (D.4), and then representing the signal on a discrete-time state form

$$\begin{aligned} s_{k+1} &= A_d s_k, \\ d_k &= C_d s_k, \end{aligned} \tag{D.5}$$

where the system matrix $A_d = \text{diag}(1, F_1, \ldots, F_N)$, with $F_n = \begin{bmatrix} \cos(\omega_n \Delta t) & -\sin(\omega_n \Delta t) \\ \sin(\omega_n \Delta t) & \cos(\omega_n \Delta t) \end{bmatrix}$ where Δt is the sampling time, and the output matrix C_d includes the Fourier coefficients. The state vector $s_k \in \mathbb{R}^{n_d}$, with $n_d = 2N + 1$, is subject to the following initial condition

$$s_{i,t_0} = \begin{cases} c_0 & \text{if } i = 0 \\ \cos(\omega_n t_0) & \text{if } i > 0, \quad i \text{ odd} \\ \sin(\omega_n t_0) & \text{if } i > 0, \quad i \text{ even} \end{cases} \tag{D.6}$$

where c_0 is a constant, t_0 is the initial time value and the index vector $i \in \mathbb{Z}, [0, n_d]$.

3 Control

The management of a WDN includes several operational objectives. In this work, the control objectives are formulated such that the main priority of the management is to ensure a robust water supply to the end-users. By regulating the tank level to a certain reference level, the pressure in the network is maintained to an appropriate pressure that guarantees the supply at the end-users. Moreover, by having certain volume of water stored at the elevated reservoir the management overcomes unexpected peaks in the demand. The second objective is to reduce high peaks of pressure in the network that increases the probability of pipe burst and therefore water leakages.
In addition to the aforementioned operational objectives, there are some physical boundaries that the network operation cannot surpass such as tank

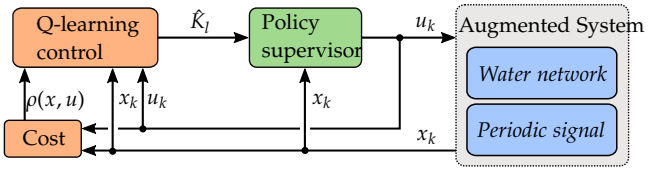


Figure D.2: Block diagram of the control scheme.

capacity or saturation of the actuators.
The control strategy that this paper proposes aims to learn an optimal controller which satisfies the given objectives while respecting certain safety boundaries. The control scheme consists of a Q-learning algorithm and a supervisor block that assesses the safety of the policy based on a standard linear model of the network system.

3.1 Augmented State Space

In this subsection, the two models presented in Section 2, WDN model (D.3) and disturbance model (D.5) are combined such that the disturbance signal is included in the model as part of the state vector. This rearrangement aims to provide an approximation scheme to the Q-learning algorithm that describes the behaviour of the linear system with periodic disturbances.

$$\begin{bmatrix} h_{k+1} \\ s_{k+1} \end{bmatrix} = \begin{bmatrix} A & EC_d \\ \mathbf{0} & A_d \end{bmatrix} \begin{bmatrix} h_k \\ s_k \end{bmatrix} + \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix} u_k, \tag{D.7}$$

where u_k is the controlled input. This work assumes that all the states included in h and s are measurable. Finally, by defining the state vector $x_k = \begin{bmatrix} h_k & s_k^T \end{bmatrix}^T$, the system (D.7) is represented in a compact form,

$$x_{k+1} = A_e x_k + B_e u_k, \tag{D.8}$$

where $x \in \mathbb{R}^{m_a}$, $u \in \mathbb{R}^{n_a}$. The feedback control policy is given by the following controller

$$u_k = \pi(x_k) = -Kx_k. \tag{D.9}$$

3.2 Problem formulation

A cost function that contains the control objectives is formulated as follows.

$$\begin{aligned} V(x_k) &= \sum_{i=k}^{\infty} \gamma^{i-k} ((x_i - r)^T Q_e (x_i - r) + u_i^T R u_i) \\ &= \sum_{i=k}^{\infty} \gamma^{i-k} \rho(x_i, u_i), \end{aligned} \tag{D.10}$$

where the weight Q_e in the first term in (D.10) penalises the deviation of the level h from a reference r_{tank}. The reference vector $r = \begin{bmatrix} r_{tank} & \mathbf{0} \end{bmatrix}^T$ contains constant references. The weight in the second term R penalises high values in the pump actuation, and γ is a constant discount factor $0 < \gamma < 1$ which reduces the cost obtained further in the future. Therefore, this parameter bounds the accumulated cost obtained when time goes to infinity. Subsequently, the instant reward ρ is defined as

$$\rho(x_k, u_k) = (x_k - r)^T Q_e (x_k - r) + u_k^T R u_k. \tag{D.11}$$

By combining the previous problem formulation and Bellman's optimality principle, the optimal value $V^*(x_k)$ can be calculated using the HJB as presented in [8]

$$V^*(x_k) = \min_u (\rho(x_k, u_k) + \gamma V^*(x_{k+1})), \tag{D.12}$$

with the notation $(\cdot^*)$ representing the optimal value. A candidate solution to the value function (D.12) is proposed. By assuming that there exists a candidate solution to the value function (D.12) of the form [20]

$$V(x_k) = x_k^T P x_k + G x_k + c, \tag{D.13}$$

the candidate solution (D.13) is combined with the Bellman equation (D.12) as follows,

$$V^*(x_k) = \min_u (\rho(x_k, u_k) + \gamma (x_{k+1}^T P x_{k+1} + G x_{k+1} + c). \tag{D.14}$$

Additionally, as presented in [8], the system dynamics of the model (D.8) is introduced in (D.14) which leads to

$$\begin{aligned} Q(x_k, u_k) = (x_k - r)^T Q_e (x_k - r) + u_k^T R u_k \\ + \gamma [(A x_k + B u_k)^T P (A x_k + B u_k) + G(A x_k + B u_k) + c]. \end{aligned} \tag{D.15}$$

Note that the value function is now expressed in terms of x and u, and is referred as Q-value function. The later expression (D.15) is rearranged in a matrix form,

$$Q(x_k, u_k) = \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} M_{xx} & M_{xu} \\ M_{ux} & M_{uu} \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} N_x \\ N_u \end{bmatrix} + \begin{bmatrix} N_x \\ N_u \end{bmatrix}^T \begin{bmatrix} x_k \\ u_k \end{bmatrix} + d, \tag{D.16}$$

then, by defining the vector $z_k = [x_k, u_k]^T$, the expression (D.16) is compacted in a quadratic form,

$$Q(z_k) = z_k^T M z_k + 2 N^T z_k + d. \tag{D.17}$$

Then, the optimal control policy for (D.17) is calculated as

$$u_k^* \in \operatorname*{argmin}_u Q(x_k, u_k) = M_{uu}^{-1} (M_{ux} x_k + N_u) \tag{D.18}$$

3.3 Function approximator

In contrast with the previous development of (D.17) where the system dynamics, matrices A_e, B_e, are known, the following method proposes a control approach based on the approximation of a Q-value function when the system dynamics are unknown. A detailed description of the safe operational domain is given in Section 3.4.
This model-free method builds a parametric approximation which estimates the Q-value function. The approximated Q-value function is developed with the same structure as (D.17) [3]. The function approximator consists of a linear parametric approximation,

$$\hat{Q}(x_k, u_k) = \phi^T(x_k, u_k)\theta, \tag{D.19}$$

where $\phi \in \mathbb{R}^{n_b}$ is a column vector with the BFs and $\theta \in \mathbb{R}^{n_b}$ is the coordinate vector with the number of bases $n_b = (m_a + n_a + 1)(m_a + n_a)/2$,

$$\phi(x_k, u_k) = [x_{1,k}^2, x_{1,k}x_{2,k}, \ldots, x_{m_a,k}^2, x_{m_a,k}u_k, u_k^2]^T. \tag{D.20}$$

The BFs vector consists of a finite set of 2^{nd} degree polynomials which are built considering the linear system (D.8). Remark that the dynamics of a real WDN are non-linear $x_{k+1} = f(x_k, u_k, d_k)$, therefore an error in the approximation is introduced. On the other hand, a larger approximation scheme based on a non-linear model considerably increases the learning uncertainty and time. The control strategy proposed in this paper prioritises a fast adaptation over optimality in the long term. This is done by learning only the system dynamics at the operation domain/nominal operation. Subsequently, the optimal control law for the approximated Q-value function is determined by

$$u_k \in \underset{u}{\operatorname{argmin}}\, \hat{Q}(x_k, u_k) = \underset{u}{\operatorname{argmin}}\, \phi^T(x_k, u_k)\theta \tag{D.21}$$

Then, the optimal policy is given by

$$u_k = \hat{\pi}(\theta, x_k) = \hat{K}(\theta)x_k. \tag{D.22}$$

Note that, by reformulating (D.19) in a quadratic form with respect to z. The approximated Q-value function has the same form as Q-value in (D.17),

$$\hat{Q}(z_k) = z_k^T \hat{H}(\theta) z_k. \tag{D.23}$$

3.4 Safety operation

The controller developed in (D.22) is an optimal control for a continuous state-action space where no boundaries of the domain are defined. However, real systems have physical limitations that the operation cannot cross. In this

work, a state x is considered safe if it belongs to the compact set $\mathcal{X}$ and an action u is a feasible control if it belongs to the compact set $\mathcal{U}$.

$$x_k = \mathcal{X} \triangleq \{x_k \in \mathbb{R}^{m_a} | \underline{x} \leq \hat{x}_k \leq \bar{x}\}, \quad \forall k \tag{D.24a}$$

$$u_k = \mathcal{U} \triangleq \{u_k \in \mathbb{R}^{n_a} | \underline{u} \leq u_k \leq \bar{u}\}, \quad \forall k, \tag{D.24b}$$

where the notation $(\underline{\cdot})$ and $(\bar{\cdot})$ define lower and upper bounds respectively. Until the Q-value function is properly mapped with the collected cost, the control algorithm might generate policies which drive the system out of the safe region. Therefore, this paper proposes a policy supervisor that verifies the risk of the control action applied and corrects the state trajectory if necessary, thus allowing only exploration of the safe set.
Assuming that the core system dynamics are known at the boundaries, the supervisor can assess the safety based on the prediction of a linear model. The standard linear model is defined as

$$\hat{x}_{k+1} = \hat{A}x_k + \hat{B}(u_k - d_{avg}) \tag{D.25}$$

where d_{avg} is the average demand. Since the specific system dynamics are unknown, the dimensions of the system matrices $\hat{A}$ and $\hat{B}$ are selected such that the predicted trajectory represents a worst case scenario. In this study only the direction of the control action and a broad estimation of the d_{avg} are sufficient to repel the system from unsafe areas. If the error between the real system and (D.25) is large, the supervisor policy nearby the boundaries can be either very conservative or very slow.
The Figure D.2 represents the control structure where a Q-learning block computes a policy based on the approximated Q-value function, a supervisor that predicts the next state and decides the control action to be applied. The `lines 16-20` in Algorithm 7 show the decision criteria.
This method aims to correct the bad behaviour of the initial policies until the collected punishments around the area improve the Q-value function approximation and subsequently the controller actions. The value of u_{safe} is the solution of the constrained optimisation problem (D.26), which is is computed with [1],

$$u_{safe} \in \underset{u}{\text{argmin}} \quad z_k^T \hat{H}(\theta) z_k \tag{D.26a}$$

$$\text{s.t.} \quad \hat{x}_{k+1} = \hat{A}x_k + \hat{B}(u_k - d_{avg}) \tag{D.26b}$$

$$\hat{x}_{k+1} \in \mathcal{X} \tag{D.26c}$$

$$u_k \in \mathcal{U} \tag{D.26d}$$

An example where the system is near the safety boundary is illustrated in Figure D.3. In this example the supervisor predicts that the system, at state x_k and following a policy $\hat{\pi}(\theta, x_k)$, will cross to the unsafe area in the next

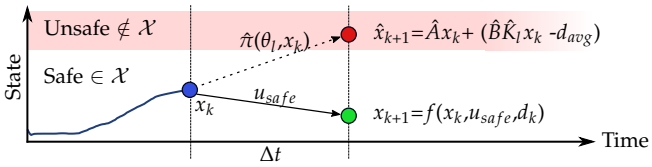


Figure D.3: Example of a policy supervision where safety of the system trajectory is assessed: Blue-dot represents the state at time k, red-dot represents the prediction based on a linear model and green-dot represents the state at time $k+1$ using a safe control action

time step. Thus, a safety event is triggered and a safe control action is applied which drives the system away of the unsafe area.

3.5 Parameter Update

The coordinate vector θ is initially unknown and it is iteratively approximated by using previously collected data. A Temporal Difference (TD) algorithm is used for approximating the the Q-value function in real-time. This algorithm minimises the approximation error between different iterations (D.27). Then, introducing the Q-value approximation (D.19) the TD update law is defined as follows

$$\phi^T(x_k, u_k)\theta_{k+1} = (1-\alpha)\phi^T(x_k, u_k)\theta_k + \alpha\left[\rho(x_k, u_k) + \gamma\phi^T(x_{k+1}, u'_k)\theta_k\right] \quad \text{(D.27)}$$

where $0 < \alpha < 1$ is a constant learning rate. The Q-value function equation (D.27) is solved by applying Least Squares Temporal Difference (LS-TD). This method applies a control action, collects data measurements and the costs until a batch with m samples is completed. By evaluating (D.27) with the batch of data, the batch update law becomes,

$$\Phi_l^T\theta_{l+1} = (1-\alpha)\Phi_l^T\theta_l + \alpha\left[J_l + \gamma\Phi'^T_l\theta_l\right] \quad \text{(D.28)}$$

where l is the iteration number, $\Phi_l = [\phi_l, \ldots, \phi_{l+m}]$ and $J_l = [\rho_l, \ldots, \rho_{l+m}]^T$ are a matrix and a vector generated by evaluating the collected data into the polynomial BFs (D.20) and reward functions (D.11) respectively.
Note that the applied control action u_k includes a persistent excitation term ϵ_k. This term ensures that the data collected contain sufficient information about the system. In some cases, when the experimental conditions are not adequate for the system identification, the parameter update might lead to inconsistent approximations. This paper uses a feature selection method presented in [19] to provide additional numerical robustness during the identification.

Algorithm 7 LS-TD for Q-function with safety supervision.

Input: γ, α, n_s,
Initialisation: $l \leftarrow 0, x_0, \theta_0$ where $\hat{\pi}(\theta_0)$ must be an admissible policy.
repeat at every iteration k = 0,1,2, ...
  apply u_k and measure x_{k+1}
  $Y_{l_s} \leftarrow \rho(x_k, u_k) + \gamma \hat{Q}(x_{k+1}, \hat{K}_l x_{k+1})$
  if $k = (l+1)n_s$ **then** ▷ Policy update
    if $inv(\Phi_l \Phi_l^T)$ == TRUE **then**
      $\theta_{l+1} \leftarrow (1-\alpha)\theta_l + \alpha(\Phi_l \Phi_l^T)^{-1} \Phi_l Y_l$
    else
      $\bar{\theta}_{l+1} \leftarrow (1-\alpha)\bar{\theta}_l + \alpha \bar{\Sigma}^{-1} \bar{V}^T Y_l$
      $\theta_{l+1} \leftarrow U \left[\bar{\theta}_{l+1}; \underline{\theta}_l\right]$
    end if
    $\hat{\pi}(\theta_{l+1}, x) \leftarrow \text{argmin}_u \, \phi(x,u)^T \theta_{l+1}$
    $l \leftarrow l+1$
  end if
  if $\hat{x}_{k+1} \in X$ **then** ▷ Policy supervisor
    $u_k = \hat{K}(\theta_l) x_k + \epsilon_k$
  else
    $u_k = u_{safe} + \epsilon_k$
  end if
until

4 Results

The proposed control algorithm is validated on both a computer simulation and a test-bed at the Smart Water Infrastructures Laboratory (SWIL). Both frameworks emulate a water distribution network with similar characteristics to the one shown in Figure D.1. A detailed description of the laboratory setup is provided in [21].

4.1 Numerical results

The network model used in the simulation is a non-linear model of the laboratory setup. The water consumption profile is generated with (D.5), a Fourier Series of 2^{nd} order. The simulation is initialised with an arbitrary policy in the safe domain $\mathcal{X}$. The left graph of Figure D.4 shows how the controller does not compensate the demand and the tank level tends to empty. After day 7, multiple safety events are triggered and the supervisor corrects the system's trajectory with a safe control action. In the period between day 7 and 16, the system is chattering near the safety boundary. After that, when the policy is improved, the tank level is regulated to the reference with a smooth control action, thus achieving the given objectives. The right graph of Figure D.4 shows that the TD error of the approximation is minimised while learning the optimal policy. Note that, in this case, the policy supervisor is fairly conservative around the boundaries, thus reducing the operation domain.

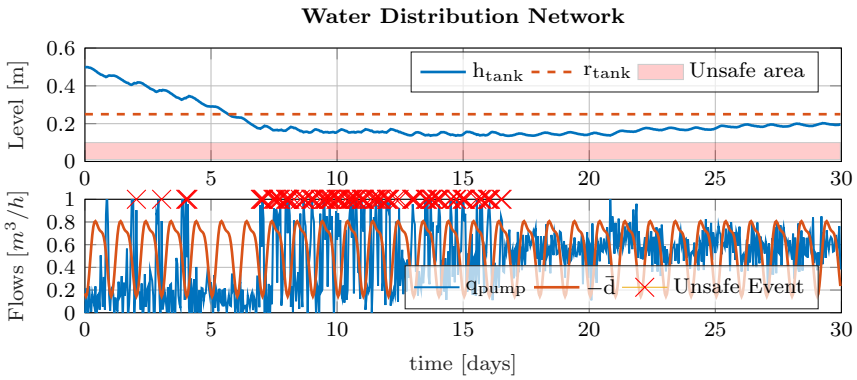


Figure D.4: Simulation results. water distribution network information.

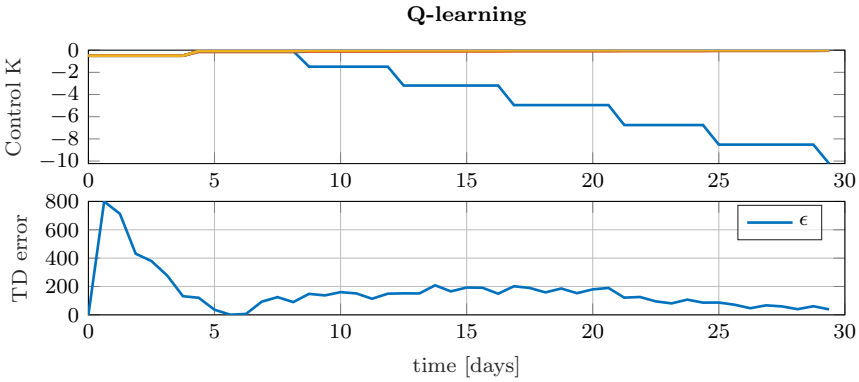


Figure D.5: Simulation results. Policy Learning.

4.2 Experimental results

A laboratory experiment is performed with a similar control criteria as the computer simulation, maintaining the initial policy and application objectives.
In Figure D.6 (left), at the start of the test, the water level is around the safety limit. This triggers frequent safety events that maintain the operation in the safe operation domain. After this, the learned policy is compensating the disturbances and tracking the reference. At this point, the TD-error shown in the right graph Figure D.6 is minimised. Figure D.7 zooms in on the learning stage, where the estimated $\hat{x}$ violates the safety boundary and several safety events are observed. The learning of an optimal policy is uninterrupted despite the safety events and the algorithm converges to an optimal policy despite the chattering nearby the boundary. During this time, the policy is gradually improved and the frequency of the safety events is reduced. Note that, in this case the safe actuation is less conservative, however the state eventually violates the safety limits. This shows that the standard linear model used in the safety constraints does not represent accurately the laboratory system.

5 Conclusion

This paper proposes an optimal control strategy to apply when the system dynamics of a WDN are unknown. The function approximators created with the reduced order linear model provide a fair estimation architecture when tested in a real system. Although it introduces an error, since the real system is non-linear, this error is negligible when the system operates around the operating point. Safety issues during the learning are addressed with a pol-

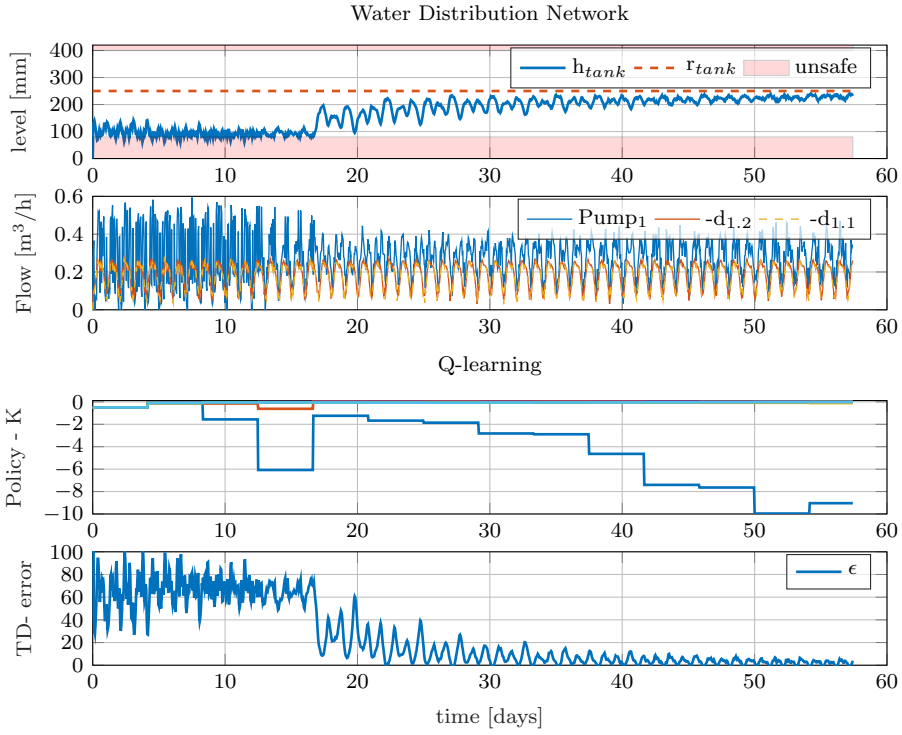


Figure D.6: Experimental results: (Top) water distribution network information. (Bottom) Policy Learning.

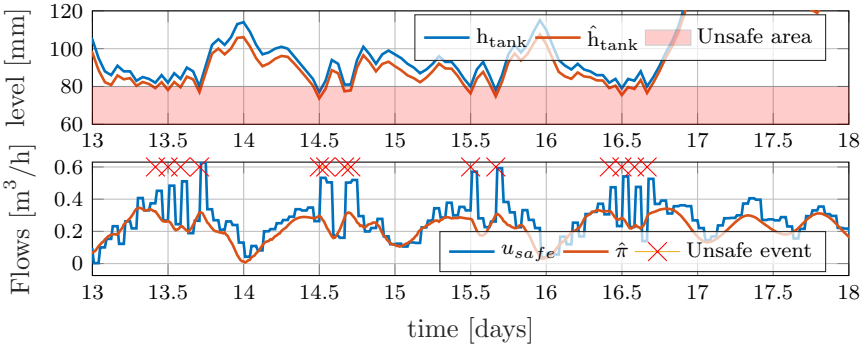


Figure D.7: Experimental results: Fragment of the learning transient with several safety events.

icy supervisor block in the control loop. It successfully filters most policies that violate the limits of the operation, and applies a safe action that drives the systems to a safe area. The safe policy is subject to a standard model. This paper assumes that a linear model is sufficient to describe the basic behaviour of the system around the safety boundaries. The model dimensions determine the performance of the policy supervisor. When large differences between real system and standard model are encountered, the supervisor might fail in the detection of the safety event, or on the contrary provide a conservative management. The validation of this method in a simulation and real framework shows the need of safety constraints that facilitate the learning, exploring only in the safe area.
In the future, the safety event detection can be improved by reducing the model uncertainly at the boundaries. The approximation method can also be extended to include a more complex geometry of the tank or include other management objectives such as water quality or operational costs.

References

[1] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl, "CasADi – A software framework for nonlinear optimization and optimal control," *Mathematical Programming Computation*, 2019.

[2] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017, pp. 908–918. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf

[3] D. P. Bertsekas, *Approximate dynamic programming*, fourth edition ed., ser. Dynamic programming and optimal control. Athena Scientific, 2012, no. Volume 2.

[4] G. Cembrano, G. Wells, J. Quevedo, R. Pérez, and R. Argelaguet, "Optimal control of a water distribution network in a supervisory control system," *Control engineering practice*, 2000.

[5] C. Hutton, L. Vamvakeridou-Lyroudia, Z. Kapelan, and D. Savic, "Uncertainty quantification and reduction in urban water systems (uws) modelling: Evaluation," *environment*, vol. 33, no. 2, pp. 1–15, 2004.

[6] C. S. Kallesøe, T. N. Jensen, and J. D. Bendtsen, "Plug-and-Play Model Predictive Control for Water Supply Networks with Storage," *IFAC-PapersOnLine*, vol. 50, no. 1, 2017.

[7] G. Konidaris, S. Osentoski, and P. Thomas, "Value function approximation in reinforcement learning using the fourier basis," in *Proceedings of the Twenty-Fifth*

AAAI Conference on Artificial Intelligence, ser. AAAI'11. AAAI Press, 2011, p. 380–385.

[8] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement Learning for Partially Observable Dynamic Processes: Adaptive Dynamic Programming Using Measured Output Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, 2011.

[9] Z. Li, U. Kalabić, and T. Chu, "Safe reinforcement learning: Learning with supervision using a constraint-admissible set," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 6390–6395.

[10] R. Lopez Farias, V. Puig, H. Rodriguez Rangel, and J. J. Flores, "Multi-model prediction for demand forecast in water distribution networks," *Energies*, vol. 11, no. 3, p. 660, 2018.

[11] Z. Marvi and B. Kiumarsi, "Safe off-policy reinforcement learning using barrier functions," in *2020 American Control Conference (ACC)*, 2020, pp. 2176–2181.

[12] C. Ocampo-Martinez, V. Puig, G. Cembrano, and J. Quevedo, "Application of Predictive Control Strategies to the Management of Complex Networks in the Urban Water Cycle," *Control Systems, IEEE*, 2013.

[13] Y. Okawa, T. Sasaki, and H. Iwane, "Control approach combining reinforcement learning and model-based control," *2019 12th Asian Control Conference (ASCC)*, pp. 1419–1424, 2019.

[14] ——, "Control approach combining reinforcement learning and model-based control," in *2019 12th Asian Control Conference (ASCC)*. IEEE, 2019, pp. 1419–1424.

[15] M. Pasha and K. Lansey, "Analysis of uncertainty on water distribution hydraulics and water quality," in *Impacts of Global Climate Change*, 2005, pp. 1–12.

[16] M. Pereira, D. M. de la Peña, D. Limon, I. Alvarado, and T. Alamo, "Application to a drinking water network of robust periodic mpc," *Control Engineering Practice*, vol. 57, pp. 50 – 60, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967066116301861

[17] F. K. Pour, V. Puig, and G. Cembrano, "Economic mpc-lpv control for the operational management of water distribution networks," *IFAC-PapersOnLine*, vol. 52, no. 23, pp. 88–93, 2019.

[18] C. C. Sun, V. Puig, and G. Cembrano, "Combining csp and mpc for the operational control of water networks," *Engineering Applications of Artificial Intelligence*, vol. 49, pp. 126–140, 2016.

[19] J. Val, R. Wisniewski, and C. Kallesøe, "Real-time reinforcement learning control in poor experimental conditions," in *2021 European Control Conference (ECC)*. IEEE, 2021.

[20] ——, "Reinforcement learning control for water distribution networks with periodic disturbances," in *2021 American Control Conference (ACC)*. IEEE, 2021.

[21] J. Val Ledesma, R. Wisniewski, and C. Kallesøe, "Optimal control for water distribution networks with unknown dynamics," vol. 53. Elsevier, Apr. 2021, pp. 6577–6582, 21th IFAC World Congress.

[22] Y. Yang, K. G. Vamvoudakis, H. Modares, W. He, Y. Yin, and D. C. Wunsch, "Safe intermittent reinforcement learning for nonlinear systems," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 690–697.

Paper E

Smart Water Infrastructures Laboratory: Reconfigurable Test-Beds for Research in Water Infrastructures Management

Jorge Val, Rafał Wisniewski and Carsten S. Kallesøe.

The paper has been published in the
Journal of Water, Special Issue Advances in the Real-Time Monitoring and Control of Urban Water Networks Vol. 13(13), 2021.

The layout has been revised.

Abstract

The smart water infrastructures laboratory is a research facility at Aalborg University, Denmark. The laboratory enables experimental research in control and management of water infrastructures in a realistic environment. The laboratory is designed as a modular system that can be configured to adapt the test-bed to the desired network. The water infrastructures recreated in this laboratory are district heating, drinking water supply, and waste water collection systems. This paper focuses on the first two types of infrastructure. In the scaled-down network the researchers can reproduce different scenarios that affect its management and validate new control strategies. This paper presents four study-cases where the laboratory is configured to represent specific water distribution and waste collection networks allowing the researcher to validate new management solutions in a safe environment. Thus, without the risk of affecting the consumers in a real network. The outcome of this research facilitates the sustainable deployment of new technology in real infrastructures.

1 Introduction

1.1 Motivation

A steadily growing population that is continuously demanding increasing living standards puts great pressure on availability of resources including energy and water [1]. The increasing demand and the need to provide it in a sustainable way challenge the urban infrastructures for transporting water, waste-water, and energy, leading to a need for their continuous development [2]. Energy savings together with renewable energy production are environmentally friendly strategies to meet these growing demands [3].

Many water resources are wasted due to leakages in the distribution network. It is estimated that around 35% in average, and in worst case up to 70%, of produced drinking water is wasted in the water infrastructures, summing up to 26.7 cubic kilometres per year in developing countries [4].

Uncontrolled sewage overflows have a severe impact on the ecosystem. The minimisation of waste water overflows is an important goal in the utility management [5]. In combined sewer systems, the rain-events are not always predicted and their uncertainty complicates the real-time control task [6].

In addition to the previous challenges, water utilities must ensure an adequate water quality during its distribution. The continuous use of chemicals in extensive agriculture and industry increases the environmental pollution and threatens the sources of potable water [7]. There are other elements that can cause loss of water quality, such as bio-film growth, corrosion, water age, or stagnation. Water age is one of the major causes of deterioration of water quality [8], and the utilities must maintain an adequate residual con-

centration of disinfectant, typically chlorine, to avoid microbial and bio-film growth [9]. However, the concentration of chlorine decays in time and so does the quality of water. Other water distribution networks, such as WDN with clean groundwater sources, that do not rely on chlorine for disinfection, are also affected by the degradation of the water quality in time. In order to avoid the water ageing, the utility management prevents the storage of volumes of water for a long period of time.

Understanding these challenges can help preventing faults in the operation of the infrastructures [10], for instance by helping the development of new solutions for monitoring and management of urban water networks. The use of decision-support tools can save time and resources to the utility management [7, 11]. These control solutions can also improve the infrastructures' resilience to changes that threaten the operation. In this way, interruption of service, water leakages, waste water overflow, inefficient operation, contamination, or cyber-attacks can be reduced or avoided.

Some researchers provide innovative control strategies for leakage detection [12–15], energy saving [16, 17], and water quality [18] for water distribution networks. In waste water collection, several researchers propose the use of advanced control strategies in the management of these infrastructures to improve their performance [19, 20] and [21]. Furthermore, the transformation of urban areas to *smart cities* [22], concept often linked with digitalisation and data collection, gives access to additional network information and enables the possibility of adopting *smart management* solutions [23], for example by using artificial intelligence (AI) techniques. These new solutions must be flexible such that the infrastructure management adapts to the dynamical needs of the city. For instance, by using the enhanced monitoring capabilities the management can provide a response to specific weather forecast or end-user demand [24].

Although the digitalisation of these critical infrastructures with AI, wireless networks and IoT sensors considerably improve the monitoring and management of the infrastructure, it can make them more vulnerable to malicious attacks including, among others, cyber-attacks [25]. Some studies have addressed the security in water systems for improving the next generation of cyber-physical systems [26].

1.2 Project Objectives

The aforementioned research studies are great contributions to the modernisation of water infrastructures. The future development of these techniques and the deployment of new technology in real infrastructures require of extensive validation. However, water utilities are cautious when testing new solutions that might put the robustness of the daily operation at risk. There are certain scenarios that practically cannot be studied due to their non affordable

consequences such as leakages, waste-water overflow, contamination of water, or interruption of the infrastructure service. The proposed methods can benefit from customised experimental tests that support the understanding of the problem and the proposed solution. The need of realistic test environments that allow the validation of the control methods on different networks motivates the smart water infrastructures laboratory (SWIL) project.

The SWIL at Aalborg University (AAU) is a facility that can replicate three types of water infrastructures: district heating, water supply and waste-water collection. Due to the domain of this journal, only water distribution networks and waste water collection are described in this paper, Figure E.1 shows the control room of the SWIL with two test-beds. This laboratory is built around three points:

1. Build a test facility which emulates the operation of three water infrastructures;
2. Flexibility to configure test-beds according to specific water networks;
3. Recreate real management problems.

Firstly, the laboratory emulates the operation of several water infrastructures. This means that the physical behaviour of the systems is qualitatively emulated and the real-time monitoring and control systems are replicated. Secondly, the SWIL is required to be versatile and replicate a wide variety of water networks. For this, this project proposes a modular laboratory which opens the possibility to replicate different topologies and network features. Modular architectures are also used in other disciplines in product development to increase the versatility and flexibility of the systems [27, 28]. Finally, the SWIL is required to have increased realism in the experiments. By using data from utilities, the test-bed can be tailored to the study case or water utility needs. This means that the laboratory test-beds must be able to emulate a particular network structure and then recreate a specific management problem in it. For instance, the real demand profiles for heat and water consumption, or rain-events can be included in the tests in a smaller-scale. Currently, the laboratory has access to datasets from several water utilities in Denmark, such as Randers, Aalborg, Fredericia, Bjerrinbro. Other institutions like EURAC in Italy [29] or iTrust located at the Singapore University of Technology [30] have advanced laboratories which are equipped with test-beds for the study of problems in water infrastructures. iTrust conducts multidisciplinary research and innovation in cyber-physical systems, monitoring, control, management, and security of critical infrastructures. However, up to our knowledge, none of the two aforementioned facilities can reconfigure the system in the same way as the SWIL.

Figure E.1: Picture of the SWIL with two test-beds and the SCADA-PC: (**Left**) Waste water collection. (**Center**) Water distribution network. (**Right**) SCADA-PC.

1.3 Research Objectives

The main objective of the laboratory is to facilitate the discovery and demonstration of optimal and resilient solutions for the development of the water infrastructures with special focus on management via automated control, computer science, and digitalisation. In this way, the laboratory allows for fast prototyping of new control solutions, such that newly developed technology can have a realistic proof of concept and verification without compromising the operation of a real network. Thus, facilitating the later scalability of the control solution to a real scale network.

This project sets a success criteria also on the scientific side, the laboratory aims to accommodate experiments from multidisciplinary research areas. Although the main focus of the laboratory is the discovery of monitoring and control solutions, other research fields like planning, civil and environmental engineering can benefit from the flexibility and data collected from the laboratory experiments. This paper highlights three theoretical research fields where the laboratory can substantially contribute:

- Optimal management;
- Fault detection and fault tolerant control;
- Security.

Some control problems related with water infrastructures that can be studied in the SWIL are: optimal pressure management, water quality, distributed control, leakage detection, contamination propagation, energy optimal operation (smart grid connection), optimal use of retention basis, overflow minimisation, or control with delays and backwater effect.

The remainder of the paper is structured as follows: Section 2 presents the design criteria and methods followed to develop the modular laboratory. Both hydraulic network and the instrumentation and the data acquisition system in this laboratory are designed to replicate the listed management problems. In Section 3 several case studies are presented and the corresponding validation of the methods with laboratory experiments is described for each case. These study cases are part of the aforementioned research problem list, this paper only gives evidence of the usefulness of the SWIL for optimal management and fault-tolerant control domains. In Section 4, the results are interpreted, the laboratory contributions are highlighted and ideas for future projects in the laboratory are also discussed. In Section 5 the conclusions of this work are summarised.

Remark that, this document does not present a collection of control solutions for water infrastructures. The scope of this paper is to inform about the development of a test facility and validation methods of control solutions

via laboratory tests, it demonstrates the functionality and applicability of the modular SWIL test-beds.

2 Materials and Methods

This section presents the methods used to develop a laboratory that meets the requirements presented in Section 1.2 and accommodates test of the research problems presented in Section 1.3. Firstly, the *module design* is presented and the main components of the water infrastructures are identified and described by means of a mathematical model. Then, the abstraction process that encapsulates the physical effects of the network components into four modules test-beds is described.

Secondly, the *network design* presents the factors that are considered important when scaling-down a real network. The mathematical models are used to build a simulation framework that supports the design of the test-beds. This includes the adequate sizing of the pipes and the maximum capacity of the test-beds. Finally, in *hardware design* the data acquisition system (DAQ), the instrumentation installed and communication architecture of the laboratory are described.

2.1 Module Design

The laboratory focuses on emulating the qualitative physical effects of the water infrastructures. For this reason, network components such as pipes, valves, and pumps are scaled to mimic the properties of any water network. In the design of the laboratory the main features of the real large scale systems are considered, the two water infrastructures studied in this paper and some of its main components are illustrated in Figure E.2.

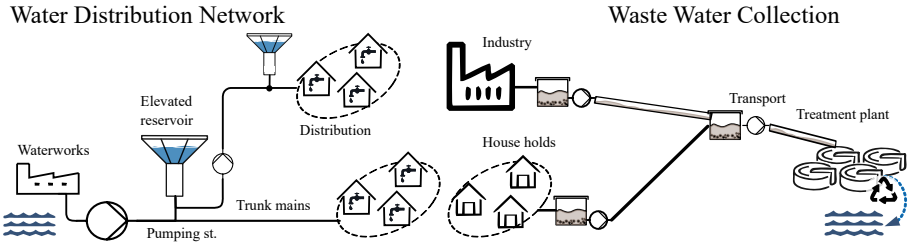


Figure E.2: Sketches of two water infrastructures: (**Left**) A water distribution network. (**Right**) A wastewater transport system.

Although these two networks differ in size, structure, and purpose, all of them are constructed of only a limited set of components. The networks

can be divided into a few basic components and this division is used in the SWIL to create test-beds. Despite of their differences, these critical infrastructures have structural similarities. They are composed of a transmission part (trunk main, transport, sewer), several supply units (water towers, pumping stations), a distribution or collection zone and storage units.

Moreover, the emulated network can be easily extended by integrating additional elements, such as tanks, consumers, suppliers, therefore, covering a wider range of test scenarios with multiple producers or interconnected networks.

The features of each infrastructure are encapsulated into five types of modules (units): A brief system description is given for each unit with mathematical models representing each network's main component.

Supply—Pumping Station/Storage

This unit has different functionalities: supply and storage. It consists of a set of pumps which boost the pressure in the pipe network. A model describing the pressure drop in a centrifugal pump is derived in [31]. Hence, the pump model is denoted by the polynomial

$$\Delta p_{pu,k} = -a_2 q_k^2 + a_1 q_k \omega - a_0 \omega^2, \tag{E.1}$$

where q_k is the flow through the pump k, $a_2 > 0$, a_1 and $a_0 > 0$ are constants describing the pump and ω is the rotational speed of the pump. Furthermore, the unit is equipped with a tank that can be used for water storage. The tank dynamics is given by the following differential equation

$$A_{er} \frac{d}{dt} h(t) = q(t), \text{with} \quad h(t_0) = h_0, \tag{E.2}$$

where A_{er} is the cross sectional area of the elevated reservoir, h is the tank level and q is the inflow to the tank. When working as an elevated reservoir, the pressure at the elevated reservoir node p_{er} is given by the algebraic relation,

$$p_{er}(t) = \mu \left(h(t) + z \right) + p_{air}, \tag{E.3}$$

where μ is a constant scaling the water level and pressure unit and z is the elevation of the tank inlet. The air pressure p_{air} inside the tank is locally regulated such that it emulates a real geodesic level or tower elevation z.

Moreover, this tank is equipped with an inner tank. The design of the tank is presented in Figure E.3. This feature allows using the tank as a retention tank/pond with a limited capacity and capture the overflow volume. The piping and instrumentation diagram of this unit is shown in the Appendix A— Figure E.24.

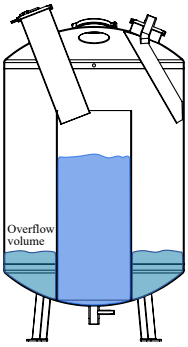


Figure E.3: Mechanical drawing of the pumping station tank.

Transmission—Pressurised Pipe

In this laboratory, there are two types of units dedicated to the water transport, pressurised pipe unit and gravity sewer unit. This division is due to the different system dynamics that characterise the transport of water. In Figure E.4 an illustration of the two types of pipes is shown.

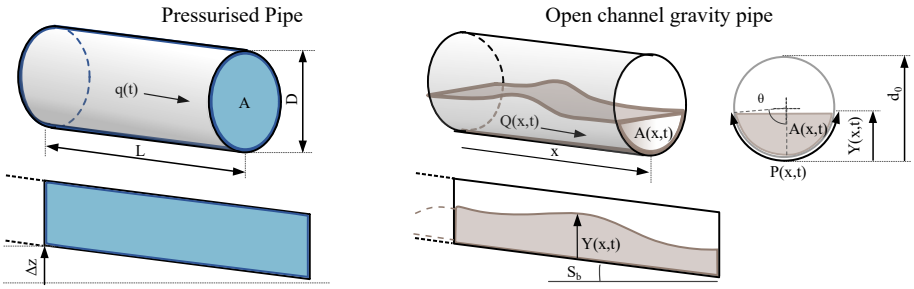


Figure E.4: (**Left**) Illustration of a pressurised pipe, Δz represents differential the elevation of the pipeline. (**Right**) Illustration of an open channel flow along a longitudinal axis x, S_b represents the bed slope.

The pipe unit emulates pressurised pipe lines in a network and it consists of a set of pipes of different diameter where several segments of pipes can be interconnected or bypassed in order to emulate different pipe length and configure the network topology.

In a water distribution network, surface roughness is not the only factor that produces resistance in the pipe during the operation; pipe bending, elbow, and fitting are also affecting the resistance. Form loss have the same structure as surface resistance, and when analysing long pipe lines as the ones modelled in this distribution network, form losses are considered negligible [32]. However, in the laboratory hydraulic circuit, the test-beds have

a number of bends and elbows that is worth considering. In this model, the flow regime is assumed turbulent and the head pressure drop through the pipe element k caused is given by Darcy-Weisbach equation [32],

$$h_{p,k}(q) = \underbrace{\frac{8 f_s L |q_k| q_k}{\pi^2} g D^5}_{\text{surface resistance}} + \underbrace{\frac{8 k_f |q_k| q_k}{\pi^2} g D^4}_{\text{form loss}} \tag{E.4}$$

where h_p is the head loss due to friction across the pipe element k, f_s is the coefficient of surface resistance, k_f is the coefficient of various form loss, L is the pipe length, D is the pipe diameter, q is the volumetric flow, g is the local acceleration due to the gravity. Then, the variation of the pipe resistance with respect to the quadratic flow through the pipe is given by [32].

$$h_{p,k}(q) = R_{p,head} |q_k| q_k \tag{E.5}$$

The piping and instrumentation diagram of this unit is shown in the Appendix A—Figure E.27.

Transmission—Gravity Sewer

This unit emulates the gravity sewers of a waste water collection. It consists of a set of pipelines that are dimensioned for open-channel flow and the slope of the pipes can be configured to cover several scenarios. The Saint-Venant equations are one of the most popular models to represent volumetric flow dynamics in open channel flow [33] where (E.6a) and represent the mass balance and (E.6b) is the momentum conservation, respectively:

$$\frac{\partial A(x,t)}{\partial t} + \frac{\partial Q(x,t)}{\partial x} = 0, \tag{E.6a}$$

$$\frac{1}{g A(x,t)} \left(\frac{\partial Q(x,t)}{\partial t} + \frac{\partial}{\partial x} \left[\frac{Q^2(x,t)}{A(x,t)} \right] \right) + \underbrace{\frac{\partial Y(x,t)}{\partial x} + \overbrace{S_f(x,t) - S_b(x)}^{\text{Kinematic wave}}}_{\text{Diffusion wave}} = 0, \tag{E.6b}$$

where $Q(x,t)$ is the volumetric flow, $Y(x,t)$ is the water depth, $A(x,t)$ is the cross section of the wetted area, $P(x,t)$ is the wetted perimeter, $S_f(x,t)$ the friction slope, $S_b(x,t)$ is the bed slope and g is the gravitational acceleration—these variables are represented in Figure E.4. This form of (E.6) is presented under the following hypotheses [34]:

1. The flow is one-dimensional. The velocity is uniform over the cross-section and the water across the section is horizontal;

2. The streamline curvature is small and vertical accelerations are negligible, hence the pressure is hydro-static;

3. The average channel bed slope is small, therefore the cosine of the angle can be approximated to 1;
4. The variation of channel width along x is small.

The algorithms that use these simplifications can be verified in the laboratory test-beds. For the design of the laboratory sewer, a steady flow solution of Saint-Venant equations is considered where $\frac{\partial}{\partial t}$ is replaced by 0 and constant water depth along the channel [34]. Then, the volumetric flow Q and wetted area A_w, the equations (E.6) are simplified to

$$S_f(x,t) - S_b = 0. \tag{E.7}$$

For these operating conditions, the uniform flow in open channels can be described by the Manning's equation [35].

$$q = A_w(K_n/n)R_h^{2/3}S_b^{1/2}, \tag{E.8}$$

where, for this work, the cross sectional area $A_w = \frac{1}{8}(2\theta - sin(2\theta))d_0^2$, the wetted perimeter $P_w = \theta d_0$, the hydraulic radius $R_h = \mathrm{A}/P_w$, K_n is constant coefficient corresponding to SI units, n is the Manning's roughness coefficient and the bed slope is defined as $S_b = \arctan\alpha$.

The piping and instrumentation diagram of this unit is shown in the Appendix A—Figure E.26.

City District—Consumer

This unit represents the end-users in a city district. The drinking water consumer consists of a valve that regulates the consumed water and a tank that collects it, the collected water is used as in-feed in the waste water system. Additionally, the geodesic level of the consumers can be emulated by introducing the equivalent air pressure in the tank.

In this project, controllable valves are used to represent the pressure drop generated at the end-users. Each valve varies its opening degree (OD) and it allows to control the pressure drop across it. The pressure drop due to the resistance factor is proportional to the quadratic term of the flow.

$$p_{cv,k} = \frac{1}{{K_{cv,k}}^2}\,|q_k|\,q_k, \tag{E.9}$$

where $K_{cv,k}$ is the conductivity of the valve k, q is the flow through the valve and Δp is the pressure drop over the component. Valve manufacturers provide an accurate parameter for the controllable valve conductivity K_{kv} which depends on the opening degree of the valve and relates the flow and pressure drop as shown in [36].

The piping and instrumentation diagram of this unit is shown in the Appendix A—Figure E.25.

2.2 Network Design

The operation of the main components, or laboratory modules, in the water infrastructures must be also analysed as part of a network. When working with a small scale network, this analysis must consider the scaling effect, a list of the main factors considered for the scaling process is given in *network scaling*.

Then, simulations of the scaled down networks are developed based on the mathematical models presented in the previous subsections [37]. The objective of these simulations is to design the correct size of the module components and evaluate the capacity of the modules when they are interconnected through a pipe network. Furthermore, having a simulation environment of the test-bed can support the preparation of the laboratory modules for a given study case. For example, by choosing certain topology, pipe length or magnitude of the signals input and disturbance signals.

Network Scaling

In order to transform a large-scale water infrastructure into a laboratory test-bed, this project has performed some simplifications to reduce the network size. This size reduction is based on four factors:

- Number of nodes: The end-users that are geographically close are aggregated and they are considered as a single consumer [38]. This node reduction does not affect the overall network structure;
- Number of pipe types: A real pipe network contains a large amount of pipe types which differ in size and material. In order to adapt the pipelines to a laboratory module, the piping is designed with a limited number of pipe diameters and lengths. The pipe networks at the laboratory are built with two pipe diameters for pressurised pipes (mains and branches), and one pipe diameter for gravity pipes (sewer pipes);
- Dimensionality: The magnitude of the network pressures and flows are reduced to meet the test-bed component requirements (sensors and actuators range). For instance, to get an idea of reduction in the magnitude, in the case of Bjerrinbro (a small water utility) the maximum supply pressure is approximately reduced from 5 bar to 4 m and the maximum supply flow from 80 m^3/h to 0.4 m^3/h;
- Time scale: The scaled-down test-beds allow accelerated tests. A test that would last several days in real-life can be replicated at the laboratory in hours. The tank modules have fixed dimensions, but its dynamics can be adjusted by varying the time scale of the tests.

All real networks differ in size and characteristics, and, therefore, the abstraction that transforms any full-scale water infrastructure into a test-bed introduces an error. In order to meet the laboratory physical limitations, an approximation of the network characteristics is required.

When choosing a smaller size for the modules, the focus of the design is to emulate most of the qualitative properties of a real network, such as pipe network topology, geodesic levels, flow regimes (turbulent and open-channel flow), system delays, actuation, and disturbance dynamics. For this reason, this study assumes that some errors introduced by the scaling, such as exact scale friction loss in the pipe network or pressure and flow ratio, have a minor impact on the tests of control solutions, since the verification method that this paper proposes relies on a proof of concept validation. The scalability of the solution is not addressed here.

Water Distribution Network Design

There are multiple elements that characterise a pipe network structure, such as ground levels, pipe size, or topology [32]. Two of the most representative topologies branched and looped geometry (ring) are illustrated in Figure E.5. Next to each topology, examples of equivalent networks constructed with laboratory modules are shown.

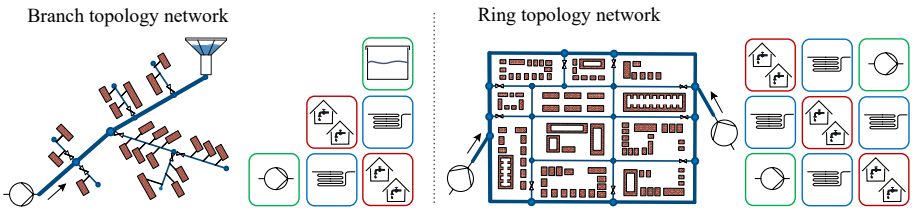


Figure E.5: (**Left**): Scheme of a standard branched topology and its laboratory equivalent. (**Right**): Scheme of a standard ring topology and its laboratory equivalent.

It can be observed that by changing the position of few modules, the structure and management of the network are significantly changed. The test-bed in Figure E.5 is transformed from being a branched structure with a single pumping station and elevated reservoir to a ring topology with two pumping stations and an increased number of end-users.

The ground levels of each urban district can be emulated by using the pressurised tank systems at the consumer and pumping modules.

Simulation A simulation of a WDN is built with the purpose of determining a reduced pipe size and the maximum capacity of the test-beds. The

network structure used for this simulation is a standard WDN with an arbitrary ring topology network that connects a pumping station (green block) and six end-users (red blocks), see Figure E.6. This network structure is inspired by the network examples studied in [32]. The capacity of this network is restricted by the pumping station, the number of consumers and pipe size. The pumping station and consumer operation is fixed and the length and diameter of the pipe units, mains and branches, are adjusted accordingly to fulfil the following conditions:

- The flow regime is turbulent in all the network pipes. Then, the friction losses are calculated with the model developed in (E.5);
- The total head is supplied by a set of *Grundfos-UPM3* pumps, its nominal operation is around q = 2 m^3/h and Δp_{pu} = 0.4 bar with speed ω = 80% for each pump, see curve in Figure E.7;
- A fraction of the total head loss (1/3) corresponds to friction loss (pipe), and the other fraction (2/3) corresponds to the pressure drop at the end-users (valves).

As mentioned on Section 2.2, a generalisation of the structure and sizing of a pipe network implies the introduction of some error since the ratio between these fractions varies from network to network. The simulations of the reference network models are developed in *modelica—Dymola*. Remark that the simulation package is built with a modular structure such that the network topology can be easily modified.

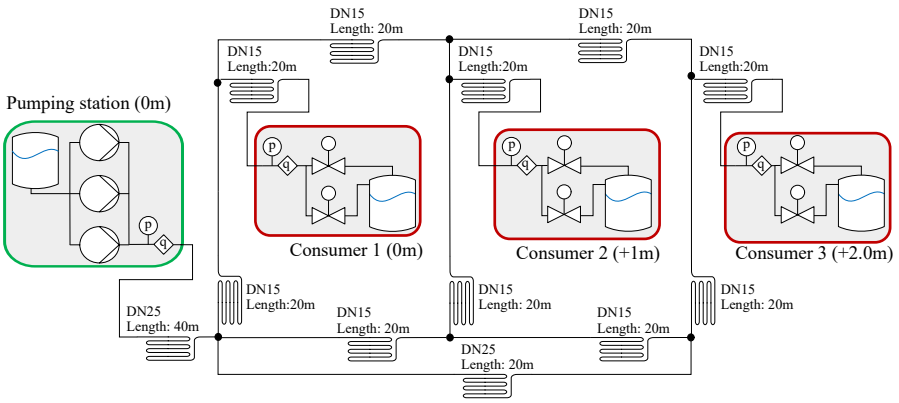


Figure E.6: Diagram of the hydraulic network used as reference for the design of the modules with a single pumping station and three consumer units representing aggregated end-users in a city district.

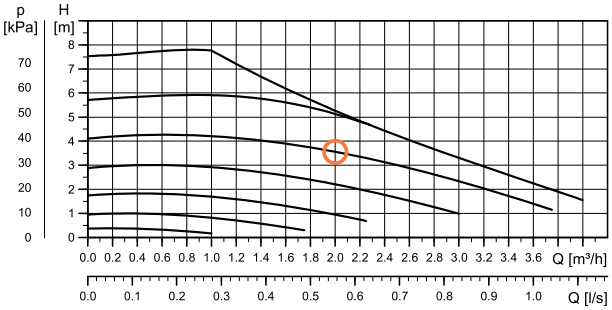


Figure E.7: Graph of a pump curve extracted from the data-sheet of *Grundfos-UPM3*. The performance of PWM controlled pumps is measured with A profile (heating) at eight PWM values: 5% (max.), 20%, 31%, 41%, 52%, 62%, 73%, 88% (min.), PWM regulates the speed of the pumps ω.

Waste Water Collection Design

The sanitary sewers are large networks of underground pipes that collect domestic sewage, industrial waste water, and rain-fall. This study focuses on a combined scheme to represent the characteristics of a typical sewer network in the laboratory, Figure E.8 illustrates the main elements in a combined sewer. In this model, the waste water conveys from different sources to the same pipe in order to be transported to retention tanks and a centralised treatment plant. The transport typically requires of a combination of both gravity sewers and pressurised pipes to overcome the elevations of the terrain. Additionally, several control elements such as retention tanks are introduced in critical locations along the network to regulate the discharge. Finally, a treatment plant receives all the water and rejects water when exceeds its capacity (overflow).

A WWC constructed with laboratory modules must comprise of the equivalent elements: water sources (green blocks) and storage elements representing retention tanks and treatment plant (red blocks). The laboratory blocks are interconnected with pressurised pipes (rising mains) or gravity sewers according to the application requirements.

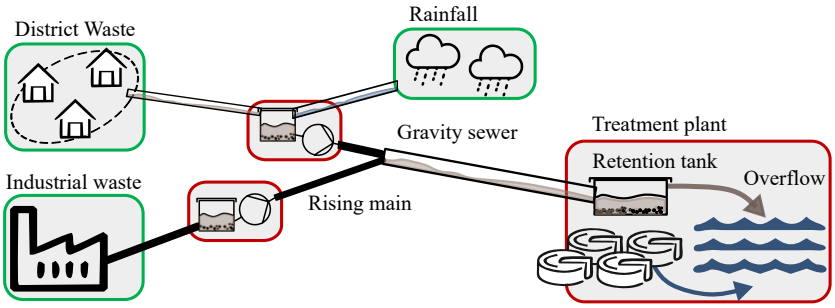


Figure E.8: Sketch of a reference waste water collection with three discharge sources, district, rain, and industry, several retention tanks and a treatment plant.

Simulation In the design of the sewer pipe module, some of the standards for sanitary sewers presented in [39] are considered. The dimensions of the sewer unit are represented in the mechanical drawing shown in the Figure E.9 (right), and the values of these parameters are calculated by solving the Manning's formula (E.8) with the following conditions:

- The maximum flow through the pipe given from the the nominal flow of a pumping station (3 pumps with 2 m^3/h each);
- The water covers half of the pipe ($\theta = \frac{\pi}{2}$) for a nominal volumetric flow;
- In this unit the bed slope S_b is constrained to the physical limitations of the laboratory unit. Due to the coiled shape of the conduct, the minimum height difference h_s for each loop is the diameter of the pipe. $S_{b,min} = \frac{d}{D\pi}$

The simulation results are shown in Figure E.9 (left). The pipe diameter d is of 8 cm (DN80) and a coil diameter D of 1 m are selected according to the given requirements. For a total pipe length L of 19.6 m, the estimated nominal delay t_{oc} = 29 s.

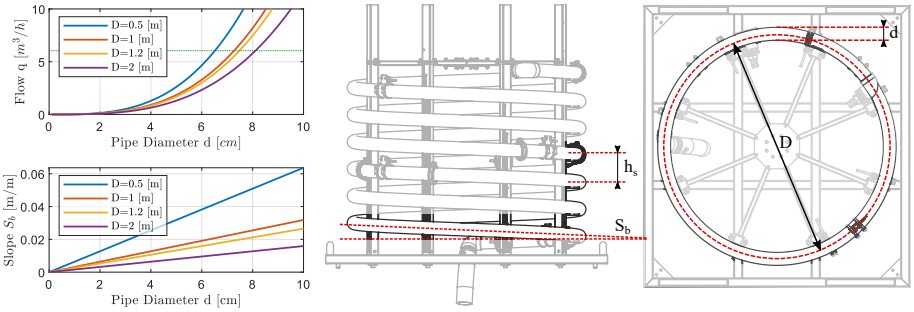


Figure E.9: (**Left**): Simulation results of the Manning's formula with different pipe sizing. (**Center**): Front view of the mechanical drawing of the gravity sewer. (**Right**): Top view of the mechanical drawing of the gravity sewer.

2.3 Hardware Design

This section describes the structure of the laboratory for control and real-time monitoring. First, a description of the data acquisition (DAQ) system and control structure is presented. Second, the communication architecture implemented is presented. Both systems are designed to emulate the control and monitoring hardware of a real water infrastructure.

Data Acquisition

The scheme in Figure E.10 shows the laboratory DAQ and control architecture that is divided into three levels:

At the management level, the central control unit (CCU)—SCADA gathers, processes, and monitors real-time data from the local units (LU).

At the local control level, the soft-PLCs perform three functions: data acquisition from the *Beckhoff* I/O Modules via Ethercat, communication with the CCU or other LUs and the control and safety of the LU. The soft-PLC consists of a Codesys runtime control installed on a *Raspberry Pi* (RPI) [40]. Moreover, the RPI is equipped with an HMI which provides a graphical interface for local monitoring, configuration, or manual control.

At the field level the I/O modules are connected to the sensors and actuators with different signals. The laboratory modules are provided with sensors to measure pressure, flow, temperature, conductivity, level. The complete list of the instrumentation equipment for each unit is shown in the Appendix A.

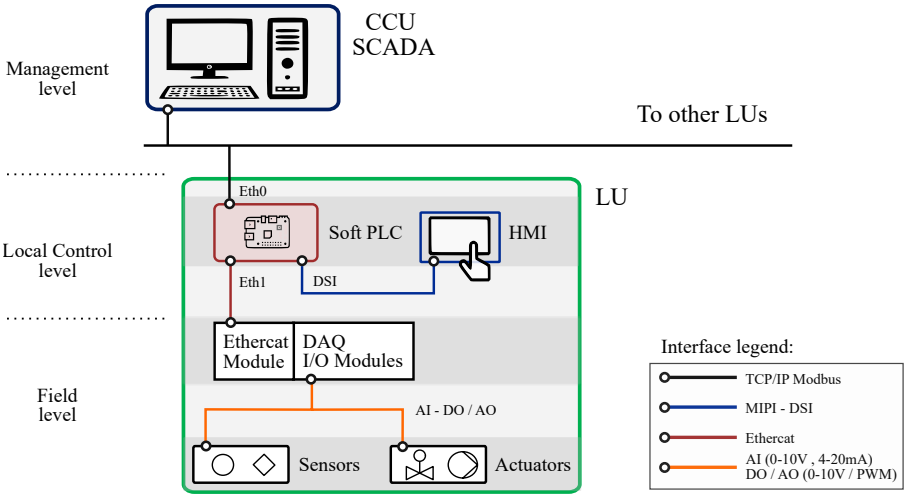


Figure E.10: Scheme of the laboratory control and DAQ architecture.

Communication Network

The communication architecture of the laboratory elements CCU and LUs is designed such that it can adapt to two control architectures: centralised or distributed control. Therefore, the communication interface is a modular system developed with MODBUS TCP-IP. Nevertheless, the hardware and software of the laboratory can also implement other communication protocols such OPC-UA.

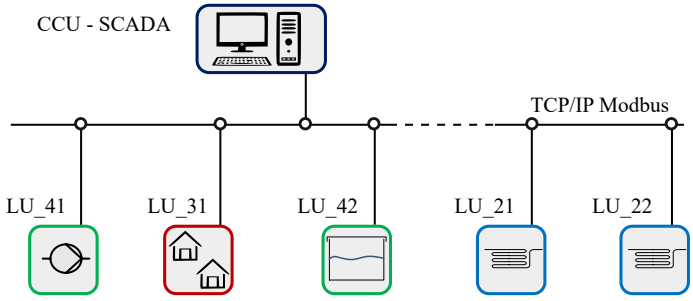


Figure E.11: Communication network architecture of the SWIL that connects the CCU and LUs via TCP/IP Modbus.

The Figure E.11 illustrates the modular communication architecture where the LUs are interconnected to a LAN together with a CCU that can be used for centralised management of the modules. The CCU is interfaced with *Simulink* for monitoring and fast implementation of the supervisory control

algorithms. An alternative interface method is developed in *Python* when the studied management requires distributed communication between LUs.

Safety and Local Control

During the design of the LUs, a risk assessment is performed with the purpose of avoiding failures and providing a secure operation of the laboratory. In order to guarantee a fail-safe operation, two protection layers are implemented on the laboratory equipment, hardware, and software safety controllers:

The first layer consists of safety valves that are installed at critical points of the laboratory such as boilers or pressurised tanks. These elements limit the maximum operating temperature and pressure, respectively. Additionally, hardware safety switches turn-off the power supply when the operation is not safe laboratory.

The second layer consists of software safety controllers that are implemented at the soft-PLC. They regulate the minimum and maximum tank level, avoiding air in the hydraulic circuit or water overflow.

3 Results

In this section, four case studies with laboratory experiments are presented. These experiments aim to reproduce different scenarios affecting the management of a real utility and illustrate the versatility of the SWIL.

First, the steps to transform real utility problems into laboratory experiments are explained. Then, a description of the laboratory experiments for each study case is given. This includes an introduction to the research contribution and a description of the laboratory customisation for each study case.

3.1 Test-Bed Configuration

First, information from the network structure is used to identify the main features of the studied infrastructure. The main components of the network, such as pumping stations, demand nodes, and rain collection or storage units, are replaced by their equivalent laboratory module and interconnected with pipe modules, emulating the real network topology. The pipe length, ground elevation and sewer's slope can be adjusted to meet the study requirements while considering the laboratory restrictions, recall scaling considerations in Section 2.2.

The laboratory test-beds are equipped with multiple sensors and actuators, this means that in most studies there is redundancy of the measure-

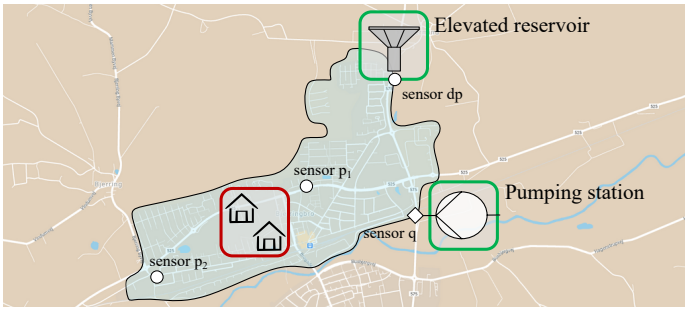


Figure E.12: Map of Bjerringbro and its water distribution network scheme. Green and red blocks represent the main elements of the network which can be emulated with laboratory modules.

ments. With a vast number of sensors it is possible to choose a subset that matches the configuration of a real system. These experiments restrict the use of the sensors according to the characteristics (number and relative position) of the studied network.

Then, real datasets of the network operation, such as water consumption, rain-events, or industrial discharge are used to adapt the test to the given problem. Note that, the laboratory experiments are performed in a smaller scale, this means that the magnitude of the real signals and time scale are adapted to meet the operation range of the test-bed.

3.2 Study Cases for Water Distribution Networks

This study case is inspired by the WDN at Bjerringbro, a small city district in Denmark, see Figure E.12. This water utility has one pumping station, one elevated reservoir and multiple end-users distributed along a pipe network. The main elements of Bjerringbro's network are identified (ring topology, number of pumping stations, elevated reservoirs, consumers, and sensors), and an equivalent scaled-down network is emulated with the laboratory modules as Figure E.13 shows.

A graph with real data from this district is presented in Figure E.14 as a reference of the water distribution network operation. This utility operates with an ON/OFF controller that regulates the elevated reservoir level.

Note that, the laboratory test-bed is equipped with an additional pumping station which is not existing in the real network, this component is added with the objective of evaluating the network management with two supply nodes.

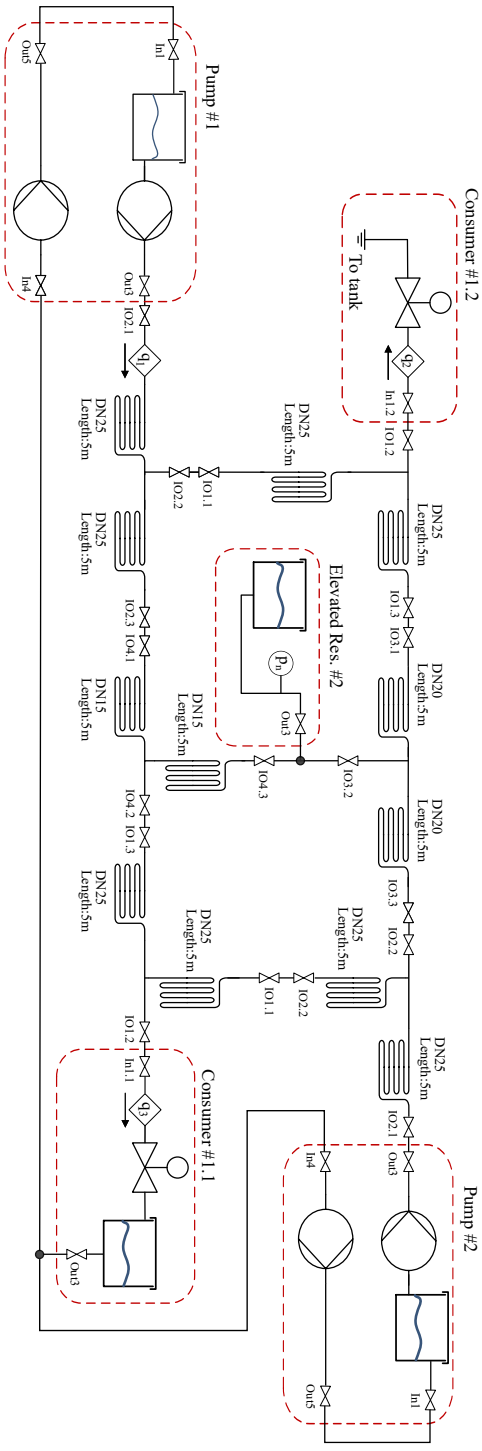


Figure E.13: Diagram of the a ring topology water distribution network that is built with laboratory modules. This network is inspired by the structure from Figure E.12.

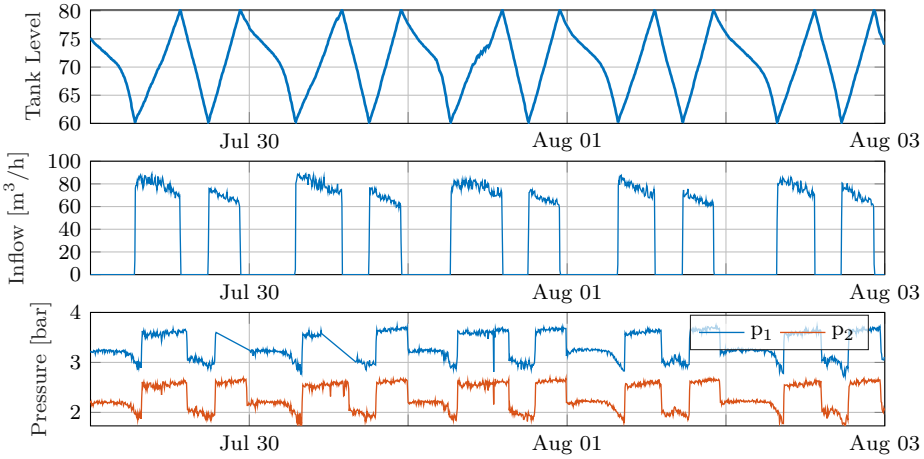


Figure E.14: Graphs of real data collected by Bjerringbro´s water utility. The measurement nodes are represented in Figure E.12 and the tank filling is regulated with a standard ON/OFF control.

Optimal Management with Multiple Supplies in Bjerringbro

This study case summarises the work presented in [41]. This study uses a network structure with two pumping stations and an elevated reservoir. This project proposes a distributed network management where a non-linear model predictive control (MPC) acts as supervisory control and PI controllers locally regulate the flow at the pumping stations. The supervisory control controls the tank level (h_{tank}) by regulating the inflows ($Pump_1$ and $Pump_2$), this controller is designed to minimise the operation cost and pressure variations at the end-users. The operational cost is evaluated using the power consumption of the pumping stations and the energy price. The district demands ($d_{1.1}$ and $d_{1.2}$) are emulated with real profile, and in the control, they are estimated using a Kalman filter. The control strategy is validated at the SWIL with the test-bed represented in Figure E.13. The experimental results in Figure E.15 show that the supervisory control schedules the pump actuation for the time-periods where the energy price is low, the study shows a reduction in the operational cost and pressure variations with respect to a standard ON/OFF tank filling that utilities typically operate with.

Optimal Management with Unknown Network Model in Bjerringbro

This study case summarises the work presented in [42], the objective of this experiment is to design an optimal controller without the knowledge of the system dynamics, in this case the network structure consists of a single pumping station. The network management is based on a reinforcement

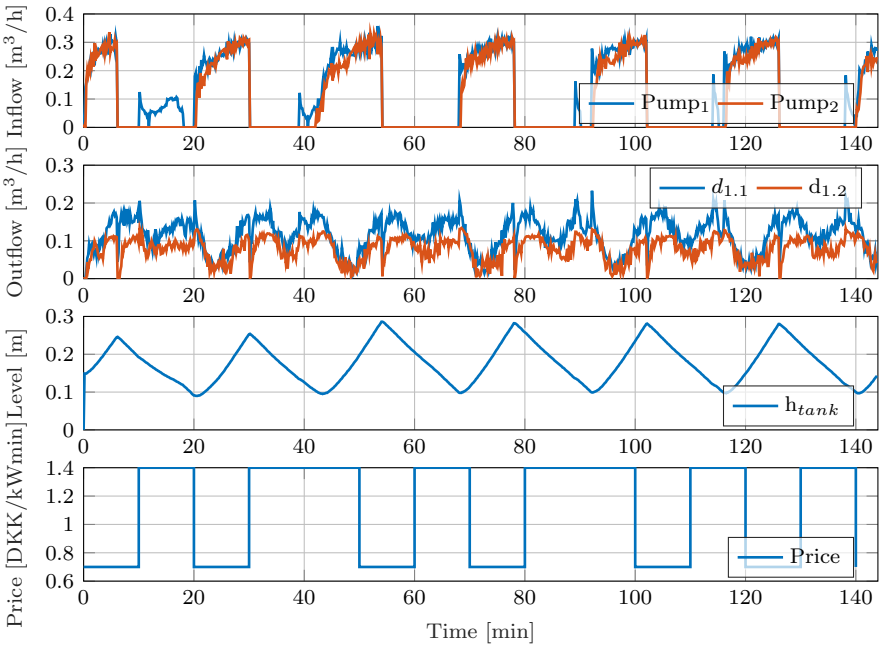


Figure E.15: Graphs of the experimental results with non-linear MPC control applied to Bjerringbro's study case [41].

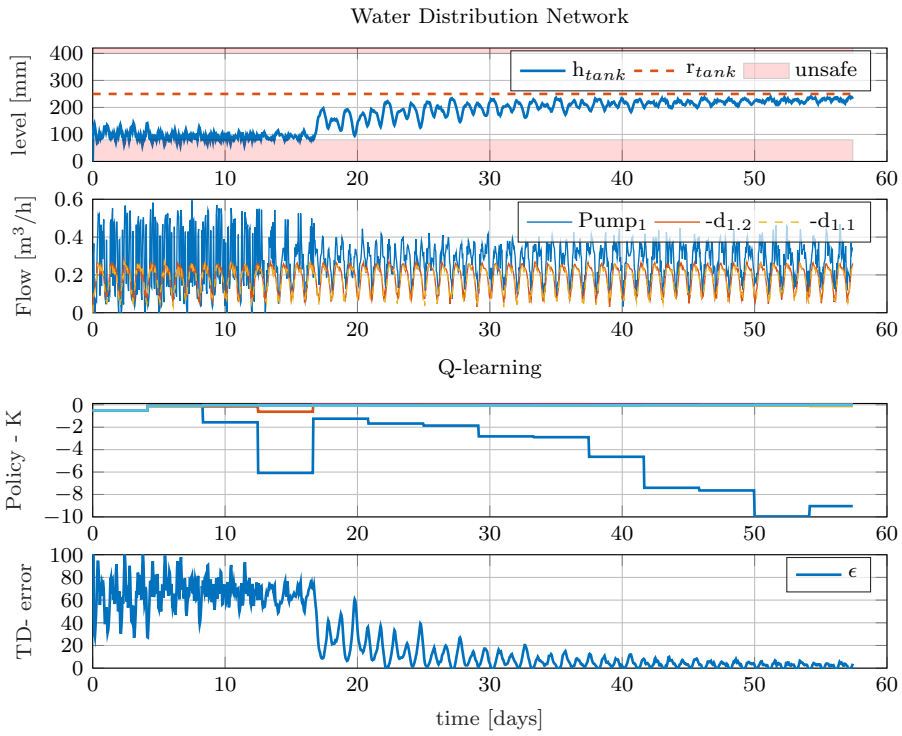


Figure E.16: Graphs of the experimental results with RL control applied to Bjerringbro's study case [42].

learning (RL) algorithm which finds the supervisory system policy based on a cost function criteria that minimises high control actions and energy consumption. The results of this work are shown in Figure E.16. The total end-user's demand ($d_{1.1}$ and $d_{1.2}$) is learned using a Fourier series basis. Additionally, the safety of the operation is guaranteed during the learning period with a policy supervisor.

This methodology is an AI control approach without stability proof, but it learns a satisfactory network management despite not having an extensive knowledge of the network, since the only information is the measured data. The safety boundaries are not violated and the end-user's water supply is guaranteed during the operation of the network. The laboratory experiment gives evidence of the robustness of the method, enabling the further investigation of AI techniques for control of real systems.

3.3 Study Cases for Waste Water Collection

Optimal Management of a Treatment Plant with Inlet Flow Variations in Fredericia

This study case summarises the work presented in [43], in this work the management of the waste water collection in Fredericia (Denmark) is studied. In this city, several industrial zones (red area), residential zones (blue area) and precipitations discharge waste water to a collection network that conveys to a treatment plant, see Figure E.17. The treatment plant opera-

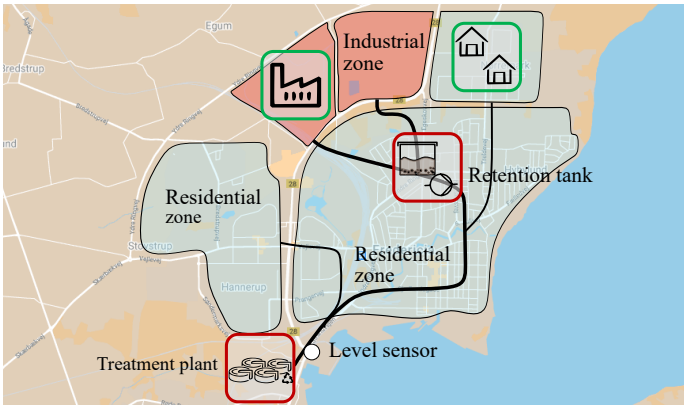


Figure E.17: Map of Fredericia and its waste water collection scheme. Green and red blocks represent the main elements of the network which are emulated with laboratory modules.

tion is based on a chemical process that increases the performance when the working conditions are stable. The waste water discharges from industry are stochastic disturbances that have a big impact on the treatment plant's performance. Therefore, the network management must regulate the inlet flow and pollutant concentration such that their variations are minimised at the treatment plant.

This work aims to minimise the inlet flow variation by controlling the industrial discharge, For this reason, the potential installation of a retention tank that regulates the varying discharge using MPC is studied. The controller considers an estimation of the household discharge via Kalman filter and takes into account the transport delay of the sewer network.

This scenario is reproduced in test-bed where the main components of the network are represented (main waste water sources, retention tank, and sewer scheme), see Figure E.18. The industry and residential discharge is locally controlled to reproduce the pattern extracted from real-data. Additionally, the only real-time measurement available is the sewer level at the inlet of the treatment plant, the controller in the experiments uses only one level sensor (72_L) to estimate the inlet flow.

The graph in Figure E.19 shows a clear minimisation of the flow variations

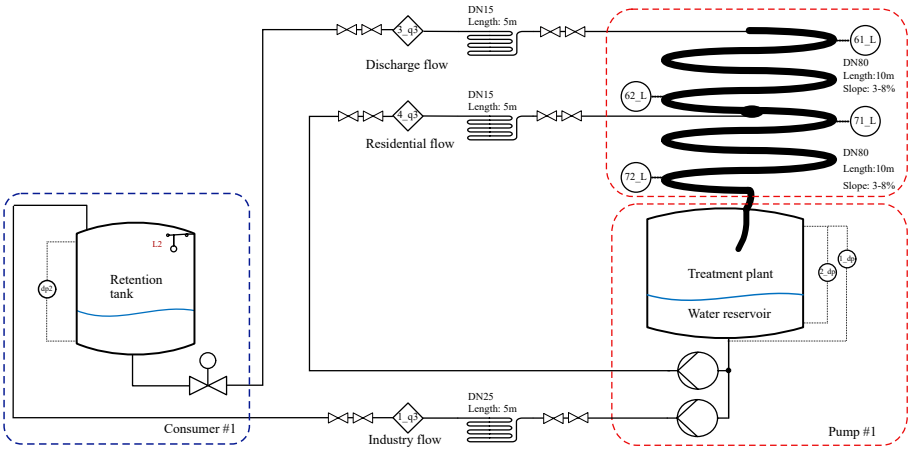


Figure E.18: Diagram of the emulated waste water collection at the SWIL.

with respect to the non-controlled operation. The experimental results show that the installation of a new control element in the network is feasible since it can considerably improve the operation of a real network.

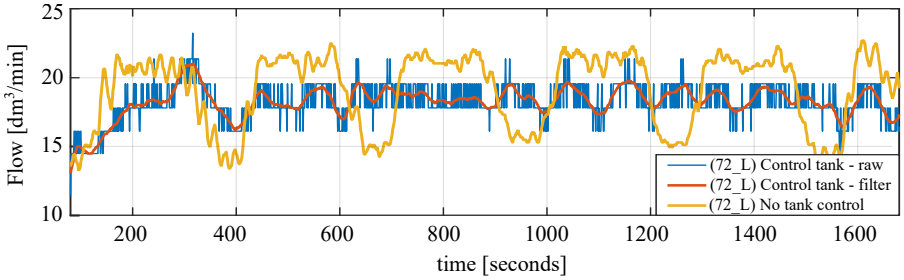


Figure E.19: Graph from the experimental results of Fredericia's study case [43]. The graph shows a comparison of the inlet flow at the treatment plant (sensor 72_L) between non-controlled industry discharge and a controlled one with MPC.

Fault Tolerant Control of a Sewer Network with the Backwater Effect in IshøJ

This study case is summarises the work presented in [44]. In this work a section of the waste water collection system in Ishøj (Denmark) is analysed, see Figure E.20.

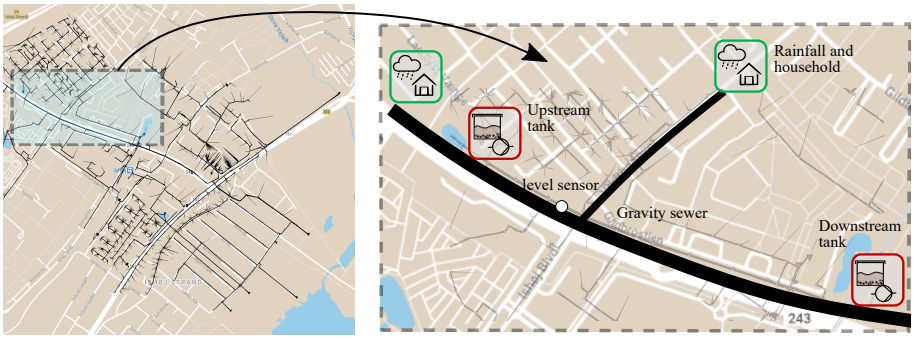


Figure E.20: Map of Ishøj and its waste water collection scheme. (**Right**) The complete network model. (**Left**) A section of the sewer with two tanks or open basins and a smaller sewer conveys to the main sewer. Green and red blocks represent the main elements of the network which are emulated with laboratory modules.

The structure of this section consists of a gravity sewer and two open basins used as retention tanks for rain drainage, one on the upstream and the other on the downstream. The knowledge of the tanks storing the volumes and the flow are needed to develop both a controller that optimises the system operation and a model which predicts overflows. This study proposes a data-driven model which accurately fits the network features, in particular, the back-water effect.

A test-bed with equivalent network features to the Ishøj's network section is emulated at the SWIL, the diagram of the test-bed is represented in Figure E.21. Although Ishøj´s scheme is a separate rain drainage collection, this laboratory experiments use disturbance profiles of combined household waste water and rainfall. The open basins are emulated with tanks placed at the upstream and downstream, the network disturbances are emulated by auxiliary modules equipped with a reservoir tank and a pump.

The results in Figure E.22 show a comparison between two model structures, kinematic wave (KW) and diffusion wave (DW). The level measurements are taken at different points of the sewer pipe, for simplicity, this paper only shows the measurements at the sensor (62_L) where the backwater effect is observed.

The effectiveness of the proposed models is presented, showing the capacity of each method for capturing back-flow inside the pipes. The experimental results show that the understanding of the back-water phenomena with a data-driven model can help the future development of fault tolerant controllers which consider this effect.

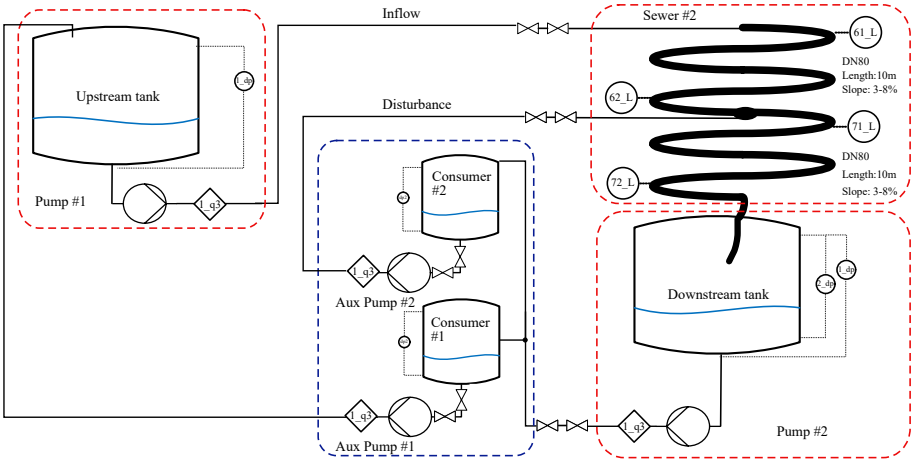


Figure E.21: Diagram of the Ishøj emulated waste water collection at the SWIL.

4 Discussion

This paper presents the development of a test facility for monitoring and real-time control of urban water networks, as well as several case studies where the laboratory contributed to the understanding and verification of the new control solutions.

The authors of this study consider that the test-beds configured at the SWIL meet the design criteria and the results support the design hypotheses: The facility at Aalborg University is able to emulate three kinds of water infrastructures. The modular properties of the test-beds allow to adapt the main features from different real networks, including datasets from the water utilities, thus increasing the realism of the laboratory test. Furthermore, the authors consider that replicating and testing certain management situations, that cannot be repeated in real infrastructures, can contribute to advance in the monitoring and real-time control of urban water networks. The verification of control methods in a customised test-bed allows quick prototyping or realisation of "proof of concepts".

On the scientific side, the four study cases analysed at the SWIL show that this facility enables the research of water infrastructures management in a novel and unique manner. The multiple configurations of the test-beds create a suitable environment to validate new control solutions on a specific real problem. The results show that the data collected on the emulated networks at the laboratory is qualitatively comparable to the real infrastructure, see Figure E.14 and Figure E.15.

This laboratory allows to test the reliability of a newly developed tech-

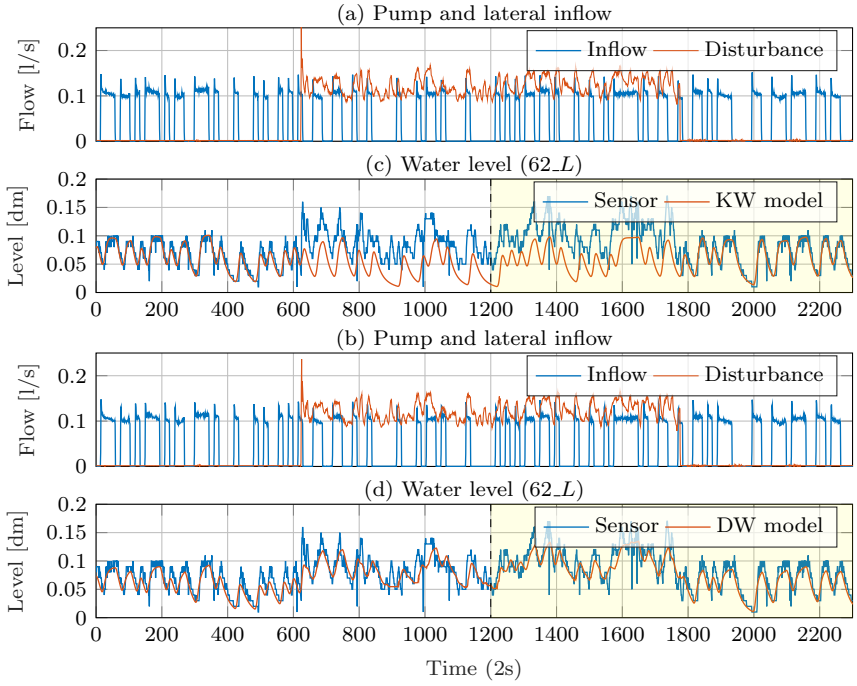


Figure E.22: Graph from the experimental results of Ishøj's study case [44], where the first part represents the system identification (training) and the second, in yellow, represents the validation.

nology and study its limitations with no consequences in case of failure. This type of validation is not feasible to perform in real infrastructures in which the impact of having a management failure can cause severe consequences such as environmental damages or damages of the infrastructure equipment, such as pumps or pipes, and discomfort of the end-users. For instance, in this safe environment, model-based controllers like MPC can be tested against model uncertainty, the propagation of pollutants in the network can be studied and develop methods to contain them, learning-based controllers can freely search for the equilibrium between optimal operation and resilient operation or network operators can evaluate the feasibility of a real network upgrade by connecting to the network additional retention tanks or pumping stations. Thus, the laboratory validation constitutes a safe and inexpensive method since the resources required to perform a test at the SWIL are relatively low: The preparation of the test-beds only requires of the assistance of a laboratory technician, and the energy and water usage can be considered negligible during the test.

Data-driven control solutions can particularly benefit from the laboratory

environment. The data collected during the laboratory experiments can support the study of certain physical phenomena like water leakages and backwater or train self-learning control algorithms.

Moreover, although fault-detection methods are not addressed in this paper, these methods can also be validated in the laboratory by recreating water leakage scenarios or water contamination propagation without water being wasted. This safe testing provides a sustainable manner of discovering new technology. The laboratory is equipped to study contamination, it uses conductivity sensors measuring salt concentration as a proxy to contamination sensors and a reverse osmosis unit to purify water.

Although the presented experiments satisfactorily reproduce the qualitative physical effects and the monitoring and management of a real water infrastructure, having scaled-down networks reduce the degree of realism. This means that the dimensionality of the tests is bounded by the number of modules available and the laboratory space, actuators are not ideally scaled-down and might introduce unwanted effects in the experiments and small networks can cause coupling between network elements.

In this work, the management scenarios are studied for each infrastructure individually. However, the infrastructure interconnection can and should also be studied. Various networks—water, heat, electricity—are no longer independent. Tons of water are used during electricity production. Vice-versa, electricity is needed for water distribution and heat production. This laboratory facility allows the study of different water networks and their interconnections, as well as links to the power supply and the internet. Research fields related with cyber-security and critical infrastructures can also be studied in this facility. The laboratory modules are already equipped with power meters at pumping and heating stations to study these problems.

5 Conclusion

The development of the smart water infrastructures laboratory has been presented. Here, the design process followed to reproduce a scaled-down water network with different modules is summarised. The main elements and features of the water infrastructures are represented in the laboratory modules. The configuration of these basic components such as pumps, tanks, or network topology are elements which characterise a network. Calculations based on component models are performed to adjust the sizing of the components to the water network properties, laboratory requirements and restrictions. The implemented DAQ system and communication interface recreate a real communication network with local smart-meters and controllers interconnected with a SCADA. This system has a modular architecture that facilitates the expansion of the test-bed and the integration of new technology or

Table E.1: List of sensors and actuators installed on the laboratory units.

Tag	Type	Model
#_*L*	Sensor: Level	Microsonic ZWS-15/CU/QS
#_*c*	Sensor: Conductivity	GF - Type 159001730
#_*pt*1	Sensor: Pressure&temp.	Grundfos DirectSensor RPI+T 0-1.6
#_*dp*	Sensor: Diff. pressure	JUMO 404382
#_*p*2	Sensor: Pressure	JUMO 404327
#_*q*1	Sensor: Volumetric flow	Festo SFAW-32
#_*q*2	Sensor: Volumetric flow	Festo SFAW-100
#_*q*3	Sensor: Volumetric flow	Endress+Hauser Proline Promag10
#_*V*1_15	Actuator: Valve DN 15	Belimo LQR24A-SR+R2015-1-S1
#_*V*3_25	Actuator: Valve DN 25	Bürkert 8804
*SV*1	Actuator: Valve	Danfoss EV210B+BE024DS
P#	Actuator: Pump	Grundfos UPM3 25-75-130
Aircontrol	Actuator: Air Control	Festo: VPPE

management solutions.

The aforementioned case studies are examples of the many possible configurations of the laboratory, where the SWIL demonstrates the capacity to replicate real problems in laboratory test-beds and reproduce real management scenarios in a scaled-down network.

A An appendix

This appendix shows the piping and instrumentation diagrams for each of the laboratory units with a complete list of installed components.

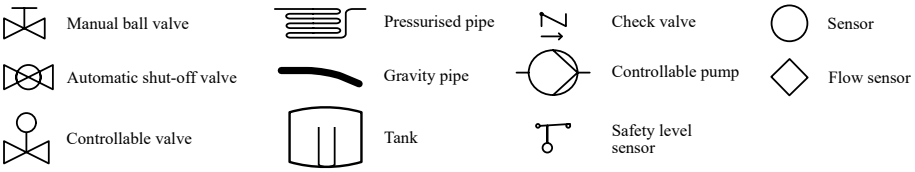


Figure E.23: Legend of the piping and instrumentation diagrams.

References

[1] OECD. *OECD Environmental Outlook to 2050*; OECD: Paris, France, 2012.

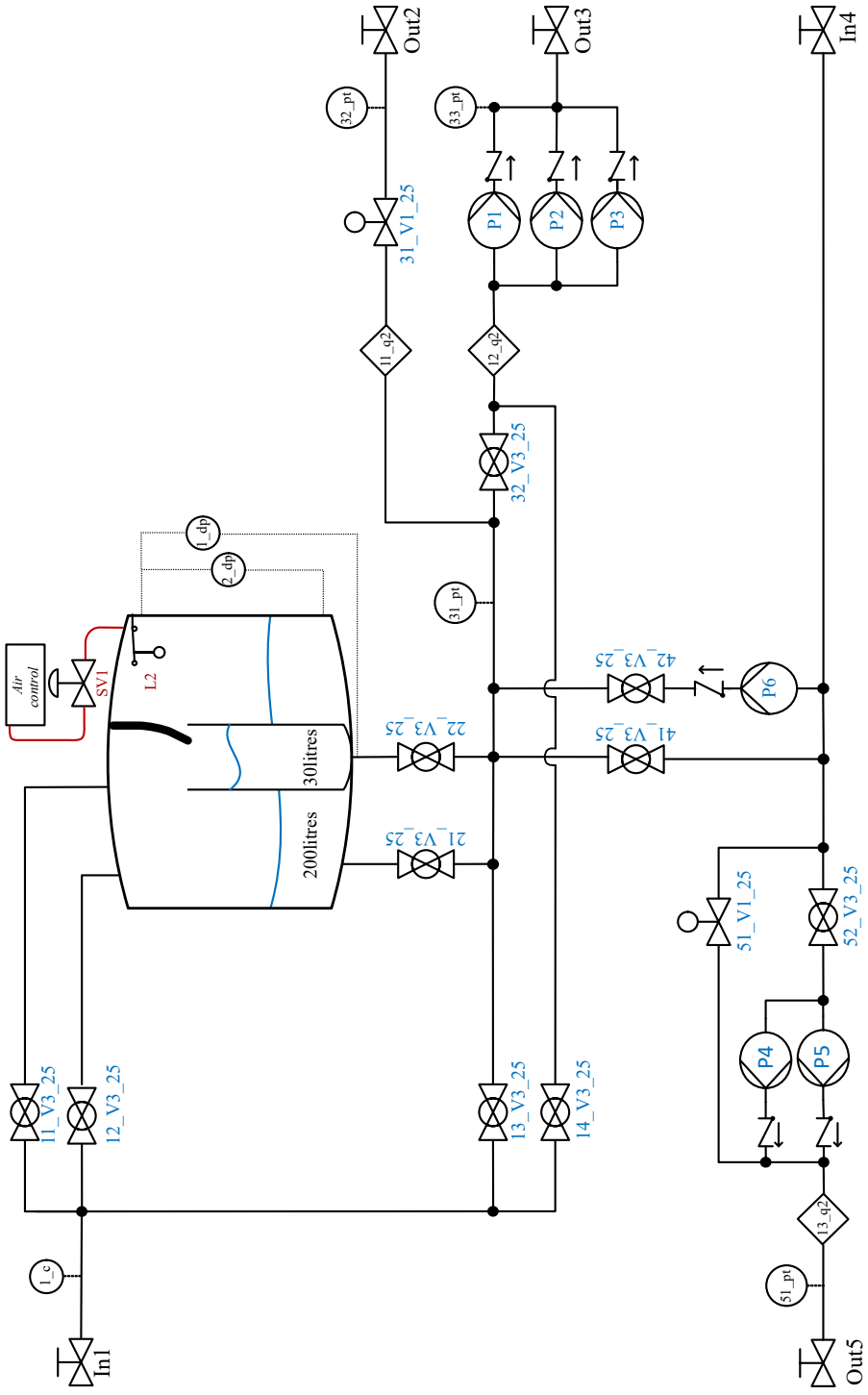


Figure E.24: Piping and instrumentation diagram of the pumping station unit. The legend is shown in Figure E.23 and the details for each component are listed in Table E.1. Shut-off valves are locally controlled to block or bypass different hydraulic circuits, thus enabling different features.

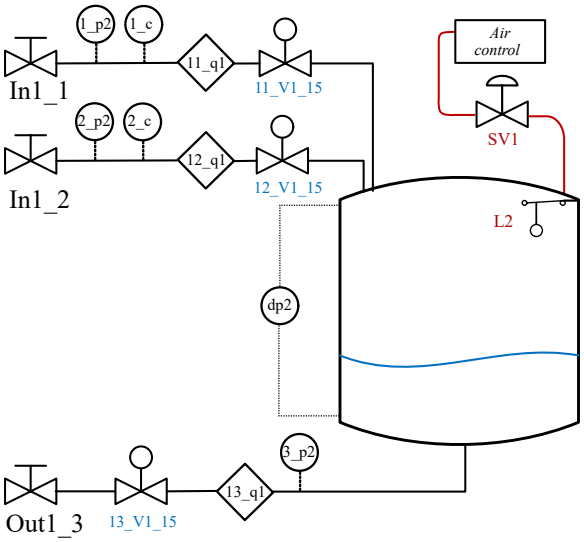


Figure E.25: Piping and instrumentation diagram of the water consumer unit. The legend is shown in Figure E.23 and the details for each component are listed in Table E.1.

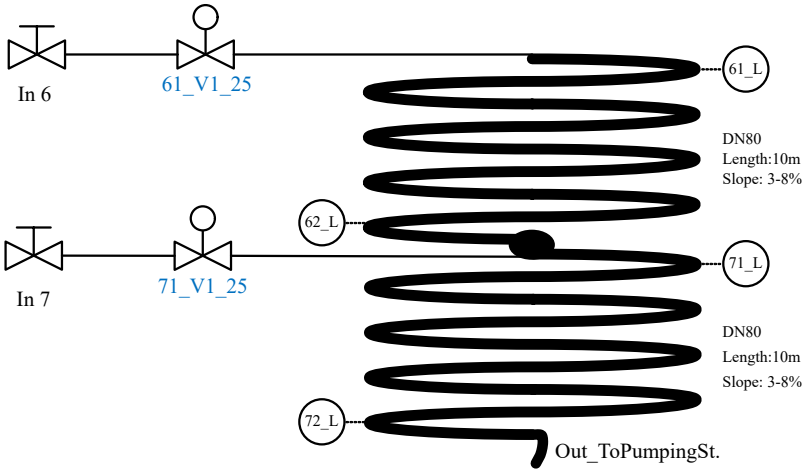


Figure E.26: Piping and instrumentation diagram of the sewer pipe unit. The legend is shown in Figure E.23 and the details for each component are listed in Table E.1.

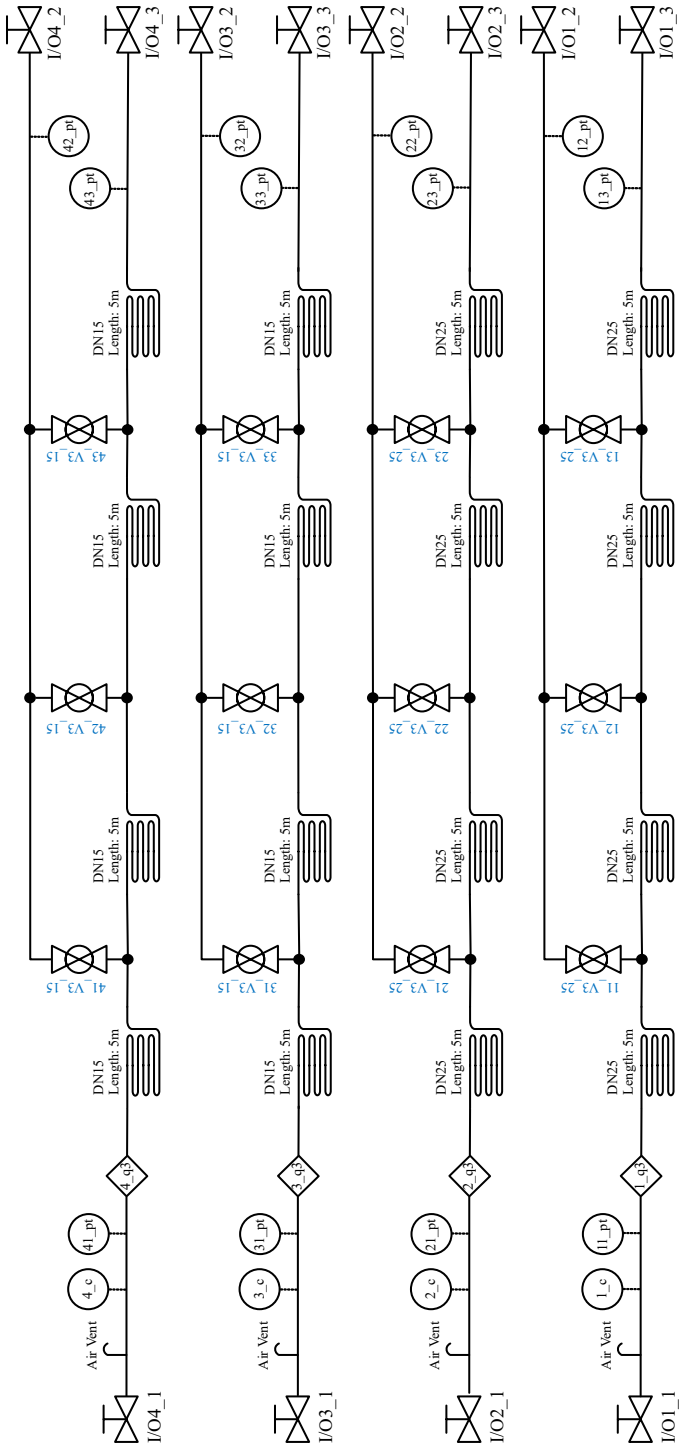


Figure E.27: Piping and instrumentation diagram of the pressurised pipe unit. . The legend is shown in Figure E.23 and the details for each component are listed in Table E.1. Shut-off valves are locally controlled to block or bypass different hydraulic circuits, thus enabling different features.

[2] CPSoS. *Analysis of the State-of-the-Art and Future Challenges in Cyber-Physical Systems of Systems*; CPSoS 611115; European Union: Brussels, Belgium, 2015.

[3] Lund, H. Renewable energy strategies for sustainable development. *Energy* **2007**, *32*, doi:10.1016/j.energy.2006.10.017.

[4] GIZ. *Guidelines for Water Loss Reduction—A Focus on Pressure Management*; GIZ: Bonn, Germany, 2011.

[5] Environmental Protection Agency. *Smart Data Infrastructure for Wet Weather Control and Decision Support*; Technical Report; EPA: Washington, DC, USA, 2021.

[6] Arnbjerg-Nielsen, K.; Willems, P.; Olsson, J.; Beecham, S.; Pathirana, A.; Bülow Gregersen, I.; Madsen, H.; Nguyen, V.T.V. Impacts of climate change on rainfall extremes and urban drainage systems: A review. *Water Sci. Technol.* **2013**, *68*, 16–28.

[7] OECD. *Diffuse Pollution, Degraded Waters*; OECD: Paris, France, 2017; p. 120, doi:10.1787/9789264269064-en.

[8] Environmental Protection Agency. *Effects of Water Age on Distribution System Water Quality*; Technical Report; EPA: Washington, DC, USA, 2007.

[9] Kowalska, B.; Kowalski, D.; Musz-Pomorska, A. Chlorine decay in water distribution systems. *Environ. Prot. Eng.* **2006**. *32*, 5–16.

[10] World Health Organization. *Water Safety in Distribution Systems*; World Health Organization: Geneva, Switzerland, 2014.

[11] Dadson, S.J.; Garrick, D.E.; Penning-Rowsell, E.C.; Hall, J.W.; Hope, R.; Hughes, J. (Eds.) *Water Science, Policy, and Management: A Global Challenge*, 1st ed.; Wiley: Hoboken, NJ, USA, 2019, doi:10.1002/9781119520627.

[12] Adedeji, K.B.; Hamam, Y.; Abu-Mahfouz, A.M. Impact of Pressure-Driven Demand on Background Leakage Estimation in Water Supply Networks. *Water* **2019**, *11*, 1600, doi:10.3390/w11081600.

[13] Bosco, C.; Campisano, A.; Modica, C.; Pezzinga, G. Application of Rehabilitation and Active Pressure Control Strategies for Leakage Reduction in a Case-Study Network. *Water* **2020**, *12*, 2215, doi:10.3390/w12082215.

[14] Wu, Z.; Sage, P. Pressure dependent demand optimization for leakage detection in water distribution systems. In *Water Management Challenges in Global Change*; Taylor & Francis: Oxfordshire, UK, 2007; pp. 353–361.

[15] Morosini, A.F.; Veltri, P.; Costanzo, F.; Savić, D. Identification of leakages by calibration of WDS models. *Procedia Eng.* **2014**, *70*, 660–667.

[16] Jensen, T.; Kallesøe, C. Application of a Novel Leakage Detection Framework for Municipal Water Supply on AAU Water Supply Lab. In Proceedings of the 2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol), Barcelona, Spain, 7–9 Septemer 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 428–433, doi:10.1109/SYSTOL.2016.7739787.

[17] Bendtsen, J.; Val, J.; Kallesøe, C.; Krstic, M. Control of District Heating System with Flow-dependent Delays. *IFAC-PapersOnLine* **2017**, *50*, 13612–13617, doi:10.1016/j.ifacol.2017.08.2385.

[18] Sakomoto, T.; Lutaaya, M.; Abraham, E. Managing Water Quality in Intermittent Supply Systems: The Case of Mukono Town, Uganda. *Water* **2020**, *12*, 806, doi:10.3390/w12030806.

[19] García, L.; Barreiro-Gomez, J.; Escobar, E.; Téllez, D.; Quijano, N.; Ocampo-Martinez, C. Modeling and real-time control of urban drainage systems: A review. *Adv. Water Resour.* **2015**, *85*, 120–132, doi:10.1016/j.advwatres.2015.08.007.

[20] Mollerup, A.; Mikkelsen, P.; Sin, G. A methodological approach to the design of optimising control strategies for sewer systems. *Environ. Model. Softw.* **2016**, *83*, 103–115, doi:10.1016/j.envsoft.2016.05.004.

[21] Lund, N.; Falk, A.K.; Borup, M.; Madsen, H.; Mikkelsen, P. Model predictive control of urban drainage systems: A review and perspective towards smart real-time water management. *Crit. Rev. Environ. Sci. Technol.* **2018**, *48*, 1–61, doi:10.1080/10643389.2018.1455484.

[22] Roche, S.; Nabian, N.; Kloeckl, K.; Ratti, C. Are 'Smart Cities' Smart Enough? In *Spatially Enabling Government, Industry and Citizens: Research Development and Perspectives*; GSDI Association Press: Needham, MA, USA, 2012; pp. 215–236.

[23] Eggimann, S.; Mutzner, L.; Wani, O.; Schneider, M.; Spuhler, D.; Moy de Vitry, M.; Beutler, P.; Maurer, M. The Potential of Knowing More: A Review of Data-Driven Urban Water Management. *Environ. Sci. Technol.* **2017**, *51*, doi:10.1021/acs.est.6b04267.

[24] Kerkez, B.; Gruden, C.; Lewis, M.; Montestruque, L.; Quigley, M.; Wong, B.; Bedig, A.; Kertesz, R.; Braun, T.; Cadwalader, O.; et al. Smarter Stormwater Systems. *Environ. Sci. Technol.* **2016**, *50*, doi:10.1021/acs.est.5b05870.

[25] Nikolopoulos, D.; Ostfeld, A.; Salomons, E.; Makropoulos, C. Resilience Assessment of Water Quality Sensor Designs under Cyber-Physical Attacks. *Water* **2021**, *13*, 647, doi:10.3390/w13050647.

[26] Tuptuk, N.; Hazell, P.; Watson, J.; Hailes, S. A Systematic Review of the State of Cyber-Security in Water Systems. *Water* **2021**, *13*, 81, doi:10.3390/w13010081.

[27] Madsen, O.; Møller, C. The AAU Smart Production Laboratory for Teaching and Research in Emerging Digital Manufacturing Technologies. *Procedia Manuf.* **2017**, *9*, 106–112, doi:10.1016/j.promfg.2017.04.036.

[28] Aicher, T.; Regulin, D.; Schütz, D.; Lieberoth-Leden, C.; Spindler, M.; Günthner, W.; Vogel-Heuser, B. Increasing flexibility of modular automated material flow systems: A meta model architecture. *IFAC-PapersOnLine* **2016**, *49*, 1543–1548, doi:10.1016/j.ifacol.2016.07.799.

[29] Eurac. Eurac Research, Ifrastructure Labs. 1992. Available online: http://www.eurac.edu (accessed on 14 April 2019).

[30] iTrust. Itrust—Singapore University of Technology and Design. 2017. Available online: https://itrust.sutd.edu.sg/ (accessed on 14 April 2019).

[31] Kallesøe, C. Fault Detection and Isolation in Centrifugal Pumps. Ph.D. Thesis, Aalborg University, Aalborg Øst, Denmark, 2005.

[32] Swamee, P.K.; Sharma, A.K. *Design of Water Supply Pipe Networks*; Wiley: Hoboken, NJ, USA, 2008, doi:10.1002/9780470225059.

[33] Schuetze, M.; Butler, D.; Beck, M. *Modelling, Simulation and Control of Urban Wastewater Systems*; Springer: Berlin/Heidelberg, Germany, 2002, doi:10.1007/978-1-4471-0157-4.

[34] Litrico, X.; Fromion, V. *Modeling and Control of Hydrosystems*; Springer: Berlin/Heidelberg, Germany, 2009.

[35] Te Chow, V. *Open Channel Hydraulics*; McGraw-Hill International Book Company: New York, NY, USA, 1982.

[36] Boysen, H. *kv: What, Why, How, Whence?*; Technical Paper; Danfoss A/S: Nordborg, Denmark, 2009. Available online: https://assets.danfoss.com/documents/90621/AC026186467824en-010201.pdf (accessed on 4 July 2021).

[37] Val, J. GitHub repository, SWIL. 2021. Available online: https://github.com/jvledesma/SWIL (accessed on 26 March 2021).

[38] Maschler, T.; Savic, D.A. *Simplification of Water Supply Network Models through Linearisation*; Technical Report 99/01; University of Exeter: Exeter, UK, 1999.

[39] Bizier, P. (Ed.) *Gravity Sanitary Sewer Design and Construction*; ASCE manuals and reports on engineering practice; American Society of Civil Engineers: Reston, VA, USA, 2007.

[40] 3S-Smart Software Solutions GmbH. CODESYS V3.5 SP14. Available online: https://www.codesys.com (accessed on 19 December 2018).

[41] Rathore, S.S. Nonlinear Optimal Control in Water Distribution Network. Master's Thesis, Aalborg University, Aalborg, Denmark, 2020.

[42] Val, J.; Wisniewski, R.; Kallesøe, C. Safe Reinforcement Learning Control for Water Distribution Networks. In Proceedings of the Conference on Control Technology and Applications, San Diego, CA, USA, 8–11 August 2021; IEEE: Piscataway, NJ, USA, 2021.

[43] Nielsen, K.; Pedersen, T.; Kallesøe, C.; Andersen, P.; Mestre, L.; Murigesan, P. Control of Sewer Flow Using a Buffer Tank. In Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics, Paris, France, 7–9 July 2020; pp. 63–70, doi:10.5220/0009871300630070.

[44] Balla, K.; Knudsen, C.; Hodzic, A.; Bendtsen, J.; Kallesøe, C. Nonlinear Grey-box Identification of Gravity-driven Sewer Networks with the Backwater Effect. In Proceedings of the Conference on Control Technology and Applications, San Diego, VA, USA, 8–11 August 2021; IEEE: Piscataway, NJ, USA, 2021.

Paper F

Water Age Control for Water Distribution Networks via Safe Reinforcement Learning

Jorge Val, Rafał Wisniewski, Carsten S. Kallesøe and Tsouvalas, Agisilaos.

The paper has been submitted for publication in the
Journal of IEEE Transactions on Control Systems Technology

The layout has been revised.

Abstract

Reinforcement Learning (RL) is a widely used control technique that finds an optimal policy without a system model using measured data. The search for the optimal policy requires that the system explores a broad region of the state space. This search puts at risk the safe operation since some of the explored areas might be near the physical system limits. Implementing learning methods in industrial applications is limited because of its uncertain behaviour when finding an optimal policy. This work proposes an RL control algorithm with a safety filter that supervises the exploration safety based on a nominal model, the performance of this safety filter is increased by modelling the uncertainty with a Gaussian Process (GP) regression. This method is applied to optimise the management of a Water Distribution Network (WDN) with an elevated reservoir; the management objectives are to regulate the tank filling while maintaining an adequate water turnover. The proposed methods are validated in a laboratory setup that emulates the hydraulic features of a WDN.

1 Introduction

Water Distribution Networks are urban infrastructures that transport drinking water from a water source to numerous end-users. The configuration of these infrastructures changes between regions, and the management adapts accordingly to network characteristics such as the topology of the terrain, the capacity of the water supply, the end-users demand or the water quality. Some of these networks can benefit from elevated reservoirs in their operation. The storage capacity of these elements can be utilised to relax the operation of the pumping stations during peak demands and guarantee a safe supply for emergencies or services [5]. Elevated reservoirs can also help to maintain adequate pressure management in the network. This pressure balancing can reduce the stress in the pipes and therefore reduce the risk of pipe burst and ensure supply in case of power failure [4]. Distribution infrastructures are designed such that the transportation time from the source to the consumers is low, avoiding stagnation in the pipes and subsequent deterioration of water quality. Chlorine avoids bacteria growth in the water, and consequently, the water quality issues are typically solved by maintaining an adequate chlorine concentration. However, its concentration decays in time [6]. Other types of distribution systems with different water sources can have chlorine-free water [20]. Nevertheless, high water age harms water quality in both types since a long residence time facilitates the growth of bacteria and microfilms. Other factors affect the water quality, such as temperature or mixing in the tank [9]. However, they are out of the scope of this work. This work studies the operation of small water utilities in Denmark, where the average temperature is below $20°C$, and the mixing inside the tank is sufficient, so its impact on the deterioration of the water quality is considered negligible [2]. Therefore, the main factor affecting the water quality in the studied network is the residence time in the reservoirs. The study presented in [13] suggests a 3 to 5 days complete water turnover as a basis; nevertheless, each utility must adapt the management to meet its own quality requirements.

1.1 Control of Water Distribution Networks

Adequate management of WDNs comprises four main objectives that [17] defines as: smoothness of the control actions, the safety of the supply, economy and water quality.
The management policies for WDN with elevated reservoirs are typically rule-based, where the main parameters considered for tank filling are the minimum and maximum tank levels. Thus, the pump is actuated based on these levels. The water age in the tank is not monitored in real-time; issues related to water age are typically solved by manually scheduling an emptying of the storage elements. This rule-based approach provides robust management that is easy to maintain. However, some important management objectives like smoothness or economics are disregarded. Furthermore, relying on a manual process for renovating the stored water increases the risk of having low water quality and lack of efficiency. The problem of efficiently controlling WDNs is a well-established research area that includes many different approaches, most of them in the framework of optimal control [18, 28]. However, the good performance of these techniques relies on a system model that requires continuous calibrations.

1.2 Motivation and Contribution of the Research

The maintenance of system models and controllers requires qualified personnel that is not always affordable, especially for small water utilities. The model dependence and high commissioning costs motivate the use of data-driven techniques that replace the reliance on detailed system knowledge and facilitate the implementation of optimal control techniques in a broader number of utilities.
Reinforcement Learning (RL) is a model-free technique that finds an optimal control policy from only measured data. RL algorithms are successfully applied for the control of dynamical systems in [15, 19]. However, its use in industrial applications is still scarce. The learning of an optimal control policy entails the exploration of a broad region of the state space. Therefore, learning while having an optimal operation are conflicting objectives that must be balanced. WDN are critical infrastructures that require a continuous operation; therefore, this approach will prioritise the robust operation of the system versus the optimal operation.
Reinforcement Learning optimises the policy of a system based on a Q-value function. This function represents an index of performance depending on state and action. This optimisation is performed over an infinite-dimensional state space where the system dynamics are unknown. Hence, the resulting optimal policy lacks state constraints.

Some studies have contributed to the improvement of safety in learning controllers [30] by combining a safe optimisation framework from MPC with the learning capabilities of RL. Gaussian Processes (GP) regressions are used to support the real-time control, [12] combines a nominal model with a GP regression. By having a learning-based MPC, the model calibration issues are compensated. Similarly, GP regression methods are used to provide safety guarantees to learning controllers [22]. Stochastic methods with MPC are also studied in the WDN domain [10],[26] and [27].
This paper presents an optimal control solution for WDN with an elevated reservoir without extensive system knowledge. This control strategy regulates the tank filling

and maintains the water age below safety limits. The proposed control structure combines reinforcement learning control and a safety filter; this method supervises the learned policy based on a deterministic nominal model and provides local robustness nearby the safety boundaries. Additionally, the safety filter performance is improved by including a GP regression that describes the uncertainty.

1.3 Organisation of This Article

The remainder of this article is organised as follows. Section 2 describes the system model of a water distribution network with an elevated reservoir. The model includes the tank dynamics and the turnover of the tank. Section 3 develops the model-free control strategy. Section 4 develops the safety filter that supervises the learning policy. Section 5 presents the case study and the validation with experimental results. Section 7 summarises the contribution of the work and introduces some ideas for future development of the proposed method.

2 System Model

This section describes a model of a small water distribution network with an elevated reservoir. This work uses the network configuration of Bjerringbro to validate the usability of the control scheme. This distribution topology is typically found in small water utilities; see Figure F.1. This paper considers a non-interrupted (continuous) operation where the pipe network is designed to avoid water stagnation. However, a significant water quality risk comes from the water age in the storage tanks. The turnover in the tank is used to monitor the water age. Therefore, a model of the daily turnover of the reservoir is developed based on available network measurements. The topology of this system consists of two pressure zones, and an elevated reservoir, a simplified map and configuration scheme of the network are illustrated in Figure F.1 and Figure F.2 respectively.

2.1 Water Distribution Network

Water Distribution Networks with elevated reservoirs are stiff systems where the network's flow dynamics are much faster than the elevated reservoir dynamics. To reduce the complexity of the model, this work assumes that the flows at the pumping stations and end-users are ideally regulated and the dominant system dynamics are the tank dynamics, given by

$$A_{er}\dot{h} = \sum_{j=1}^{n_{er}} q_j(t), \tag{F.1}$$

where A_{er} is the cross sectional area of the elevated reservoir, n_{er} is the number of inlets in a tank and q_j is the flow at the tank inlet j. Due to mass conservation in the network, the relation between the supply flows q_{p1}, q_{p2}, the demand flows d and tank

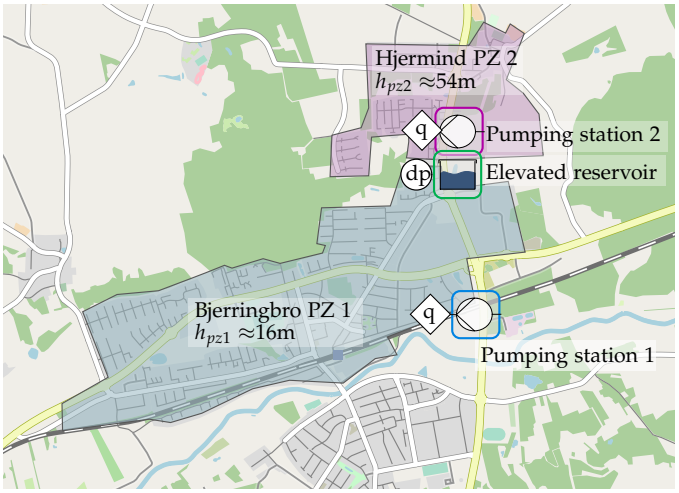


Figure F.1: Map of a water distribution network with an elevated reservoir in Denmark. Pressure Zone 1 (PZ 1) covers the Bjerringbro district, and Pressure Zone 2 (PZ 2) covers Hjermind that are located at a different elevation (h_{pz}).

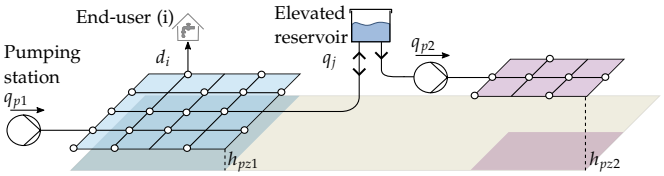


Figure F.2: Scheme of the study case network topology, a WDN with a single pumping station and an elevated reservoir. It is divided into PZ1 and PZ2 (blue and pink). The nodal demand d_i represents the consumption from the end-user i and q_j represents the inflow to the tank in connection j.

flows is described by,

$$\sum_{j=1}^{n_{er}} q_j(t) = q_{p1}(t) - \sum_{i=1}^{n_d} d_i(t) - q_{p2}(t), \tag{F.2}$$

where $d_i > 0$ represent demand of end-user in the pressure zone 1, n_d is the number of end-user demand and the flow to pressure zone 2. The demand pattern between the different end-users is similar, therefore the total demand $\bar{d}$ in a pressure zone is described by

$$d_i(t) = \beta_i \bar{d}(t) \qquad \forall i = 1, \ldots, n_d, \tag{F.3}$$

where β is a scaling factor. The total water demand in a WDN $\bar{d}$ includes the consumption of PZ1 and PZ2, this signal is a stochastic Wienner process that follows a daily pattern and it can be approximated with a Fourier Series (FS) of N^{th} order,

$$\bar{d}(t) = a_0 + \sum_{n=1}^{N} (a_n \cos(\omega_n t) + b_n \sin(\omega_n t)) + w, \tag{F.4}$$

where a_0, a_n and $b_n \in \mathbb{R}$ are the Fourier coefficients, $\omega_n = 2\pi n f_0$ and f_0 represents the fundamental frequency and w is Brownian noise.
Finally, the continuous-time models of the tank (F.1) and total demand (F.4) are transformed into discrete-time state space models, first, the tank model is defined as,

$$h_{k+1} = f(h_k, u_k, \bar{d}_k) = Ah_k + Bu_k + E\bar{d}_k, \tag{F.5}$$

where $h_k \in \mathbb{R}$ is the system state, $u_k \in \mathbb{R}$ is the controlled input flow and $\bar{d}_k \in \mathbb{R}$ represents the total system disturbances, and A,B and E are constant matrices with compatible dimensions. Then, the periodic disturbance model is also expressed in a state space discrete form,

$$\begin{aligned} s_{k+1} &= A_d s_k, \\ \bar{d}_k &= C_d s_k, \end{aligned} \tag{F.6}$$

where the system matrix $A_d = \text{diag}(1, F_1, \ldots, F_N)$, with $F_n = \begin{bmatrix} \cos(\omega_n \Delta t) & -\sin(\omega_n \Delta t) \\ \sin(\omega_n \Delta t) & \cos(\omega_n \Delta t) \end{bmatrix}$ where Δt is the sampling time, and the output matrix C_d includes the Fourier coefficients. The state vector $s_k \in \mathbb{R}^{n_s}$, with $n_s = 2N + 1$, is subject to the following initial condition

$$s_{i,t_0} = \begin{cases} c_0 & \text{if } i = 0 \\ \cos(\omega_n t_0) & \text{if } i > 0, \quad i \text{ odd} \\ \sin(\omega_n t_0) & \text{if } i > 0, \quad i \text{ even} \end{cases} \tag{F.7}$$

where c_0 is a constant, t_0 is the initial time value and the index vector $i \in \mathbb{Z}, [0, n_s]$.

2.2 Estimating the water age from measurements

The real-time monitoring of water quality and its regulation is complex since it is affected by multiple factors such as biological and chemical composition, mixing, and temperature [5]. This work aims to control the quality of the stored water by having adequate storage management. Therefore, it focuses only on those factors that the network management can control. Assuming a low water temperature and ideal mixing in the tank, the water quality is homogeneous for the stored volume. Thus, the water age, or Average Residence Time (ART), becomes the principal factor that impacts water quality. The ART in an elevated reservoir is defined as[7],

$$ART_k = \frac{v_{av,k}}{q_{av,k}}, \quad \text{for } q_{av,k} > 0, \tag{F.8}$$

where $v_{av,k}$ is the average volume, and q_{av} is the average flow entering the tank. The flow measurements in a water distribution network are limited, the proposed network have only flow measurements at the pumping stations. The flow measurements at the tank are not available but they can be inferred from its impact in the tank level due to mass conservation. Therefore, the ART is reformulated to be computed from level measurement as follows. Firstly, by discretising the tank dynamics (F.1), the average volume is defined with previous level measurements,

$$v_{av,k} = \frac{A_{er} \sum_{i=k-n_{av}}^{k} h_i}{n_{av}}, \tag{F.9}$$

where n_{av} represents the number of samples in a period. Secondly, the average inflow is denoted as,

$$q_{av,k} = \frac{A_{er} \sum_{i=k-n_{av}}^{k} \delta_i}{n_{av}}, \tag{F.10}$$

where δ represents only the positive level variations,

$$\delta_{i+1} = \max\{(h_{i+1} - h_i), 0\} \tag{F.11a}$$

$$= \max\{(A_{er}^{-1} \sum_{j=1}^{n_{er}} q_{j,i} \Delta t), 0\} \tag{F.11b}$$

where $q_{j,i}$ is the flow at inlet j at time i. Note that (F.11) has two expressions depending on the available measurement, for notational simplicity $\delta(q_k)$ is represented as function of the tank flows. Finally, the daily volume turnover [%] is denoted as a function of the flows in the network and level of the tank,

$$\tau_k = g(h_k, q_k) = 100 n_{av} \frac{q_{av,k}}{v_{av,k}}. \tag{F.12}$$

For numerical convenience, this paper uses the tank turnover $\hat{\tau}$ over of ART to regulate the ageing of the water in the tank.

Incremental average approximation

The turnover is defined in (F.12) as a non-linear equation that depends on past measurements (time-lag), this increases the computational effort and requires data storage.
This paper proposes an approximation of (F.12) by using a incremental average. First, denote the mean of the level variations δ,

$$m_k(q_k) = \frac{1}{n_{av}} \sum_{i=k-n_{av}}^{k} \delta_i(q_k) \tag{F.13}$$

Then, the new mean is computed by extending (F.13) and including the new input (F.11) as follows,

$$\hat{m}_{k+1} = \frac{1}{n_m}(\delta_{k+1}(q_k) + (n_m - 1)\hat{m}_k), \tag{F.14a}$$

$$= \hat{m}_k + \frac{\delta_{k+1}(q_k) - \hat{m}_k}{n_m}. \tag{F.14b}$$

where n_m is the window length of the moving average filter. Subsequently, the daily turnover output is approximated as follows,

$$\hat{\tau}_k = \hat{g}(q_k) = 100 \frac{\hat{m}_k}{h^*} \tag{F.15}$$

where h^* is a constant representing the average level during steady-state operation and q_k is the sum of tank inflows. The computational implementation of this model uses the expression (F.14b) to reduce the computation precision errors.

Topology considerations

This section transforms the computation of the water turnover in the tank for different hydraulic configurations. The model presented in (F.11) computes positive level variations in a tank. However, this model is only accurate for tanks with a single inlet. This issue motivates an alternative formulation of the turnover model that describes the in-feed of freshwater for multiple inlets.
Consider an example with the hydraulic configuration illustrated in Figure F.2, where the tank has two pipe connections, one bidirectional and one outflow. This configuration ensures a regular inflow of freshwater during steady-state conditions via *inlet 1*. However, the model (F.11), that utilises level measurements or the sum of the flows, returns zero level variations since $\sum_{j=1}^{n_{er}} q_j(t) = 0$.
This model aims to calculate the volume of freshwater introduced to the tank in a given time. For this, the expression (F.1) is expanded for different tank inlets and integrated over a time period T as follows,

$$A_{er} \int_T \dot{h} dt = \int_T \underbrace{A_{er}\dot{h}_1}_{inlet1} dt - \int_T \underbrace{A_{er}\dot{h}_2}_{inlet2} dt, \tag{F.16}$$

where $\dot{h}_1$ and $\dot{h}_2$ represent the level rate from *inlet 1* and *inlet 2* respectively. By analysing this particular study case, only *inlet 1* inflows water to the tank, *inlet 2* is a fixed outfeed but its magnitude is known (measured). The level rate $\dot{h}_1$ cannot be measured but it can be computed with the available measurements. With this system knowledge, the expression (F.16) is discretised and represented in terms of tank level and PZ2 flow,

$$\Delta h_1 = \Delta h + A_{er}^{-1} q_{p2} \Delta t, \tag{F.17}$$

where $\Delta h = h_{k+1} - h_k$. Then, the positive level variations are computed with $\dot{h}_1$,

$$\delta_{k+1} = \max\{\Delta h_1, 0\}. \tag{F.18}$$

Subsequently, the turnover (F.15) is similarly computed with *inlet 1* positive level variations (F.18).

3 Reinforcement Learning Control

This section presents the formulation of an optimal-adaptive controller which adapts to the hydraulic configuration and the consumption pattern of the network to achieve the utility management objectives. In the proposed control structure, the RL control aims to regulate the tank filling and the smoothness of the control. Water quality is later formulated in Section 4 as a safety problem.

3.1 Augmented state space

This work use the augmented state space model proposed in [23] to construct the learning scheme. This augmented model consist of the discrete-time tank model and

the periodic disturbances. Then, by combining (F.5) and (F.6) in an augmented state space,

$$\begin{bmatrix} h_{k+1} \\ s_{k+1} \end{bmatrix} = \begin{bmatrix} A & EC_d \\ \mathbf{0} & A_d \end{bmatrix} \begin{bmatrix} h_k \\ s_k \end{bmatrix} + \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix} u_k, \tag{F.19}$$

where u_k is the controlled input. This work assumes that all the states included in h and s are measurable. Finally, by defining the state vector $x_k = \begin{bmatrix} h_k, & s_k^T \end{bmatrix}^T$, the system (F.19) is represented in a compact form,

$$x_{k+1} = A_e x_k + B_e u_k, \tag{F.20}$$

where $x \in \mathbb{R}^{m_a}$, $u \in \mathbb{R}^{n_a}$. The feedback control policy is given by the following linear controller,

$$u_k = \pi(x_k) = -Kx_k. \tag{F.21}$$

3.2 Cost function and Bellman equation

A cost function that includes the system objectives is defined as follows,

$$\begin{aligned} V(x_k) &= \sum_{i=k}^{\infty} \gamma^{i-k} ((x_i - x^*)^T Q_e (x_i - x^*) + u_i^T R u_i) \\ &= \sum_{i=k}^{\infty} \gamma^{i-k} \rho(x_i, u_i), \end{aligned} \tag{F.22}$$

where Q_e and R are a constants that penalise the tracking error and high control actions respectively, $x^* \in \mathbb{R}^{m_a}$ includes the reference of the tank h^* and γ is a constant factor, $0 < \gamma < 1$, that discounts the rewards obtained in the future. The instant reward ρ_k is denoted

$$\rho(x_k, u_k) = (x_k - x^*)^T Q_e (x_k - x^*) + u_k^T R u_k. \tag{F.23}$$

By formulating the previous cost function with the Bellman's optimality principle, the optimal value function is expressed with the Bellman equation as presented in [16],

$$V^*(x_k) = \min_u (\rho(x_k, u_k) + \gamma V^*(x_{k+1})), \tag{F.24}$$

with the notation $(\cdot^*)$ representing the optimal value. Consider a candidate parametrisation to the value function (F.24) of the form [23],

$$V(x_k) = x_k^T P x_k + G x_k + c, \tag{F.25}$$

Then, combining (F.24) and (F.25) leads to,

$$V^*(x_k) = \min_u (\rho(x_k, u_k) + \gamma (x_{k+1}^T P x_{k+1} + G x_{k+1} + c). \tag{F.26}$$

Finally, the previous V-value function (F.26) is expressed as a Q-value function and it is represented in terms of state x and control action u, subsequently where the system dynamics of the augmented state-space model (F.19) are introduced,

$$\begin{aligned} Q(x_k, u_k) =& (x_k - x^*)^T Q_e (x_k - x^*) + u_k^T R u_k \\ &+ \gamma [(A x_k + B u_k)^T P (A x_k + B u_k) \\ &+ G(A x_k + B u_k) + c]. \end{aligned} \tag{F.27}$$

Then, by rearranging (F.27) into a matrix form

$$Q(x_k, u_k) = \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} M_{xx} & M_{xu} \\ M_{ux} & M_{uu} \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} N_x \\ N_u \end{bmatrix} + \begin{bmatrix} N_x \\ N_u \end{bmatrix}^T \begin{bmatrix} x_k \\ u_k \end{bmatrix} + d, \tag{F.28}$$

then, the optimal control policy for (F.28) is calculated as

$$u_k^* \in \underset{u}{\operatorname{argmin}}\, Q(x_k, u_k) = M_{uu}^{-1}(M_{ux}x_k + N_u) \tag{F.29}$$

3.3 Q-value using linear architecture

The Section 3.2 presents a Q-value function for a linear system that is built with the system matrices A_e, B_e. The following section presents an equivalent control approach where the Q-value function is approximated since the system matrices are unknown. This model-free approach proposes a linear parametric approximation for the Q-value function of the form,

$$\hat{Q}(x_k, u_k) = \phi^T(x_k, u_k)\theta, \tag{F.30}$$

where $\phi \in \mathbb{R}^{n_b}$ is a column vector with the BFs and $\theta \in \mathbb{R}^{n_b}$ is the coordinate vector with the number of bases $n_b = (m_a + n_a + 1)(m_a + n_a)/2$. The column vector is built with a finite set of polynomials of 2^{nd} degree, this polynomial approximation is inspired by the quadratic form of its model-based version (F.28) [23],

$$\phi(x_k, u_k) = [x_{1,k}^2, x_{1,k}x_{2,k}, \ldots, x_{m_a,k}^2, x_{m_a,k}u_k, u_k^2]^T. \tag{F.31}$$

Subsequently, the optimal control law for the approximated Q-value function is determined by

$$u_k \in \underset{u}{\operatorname{argmin}}\, \hat{Q}(x_k, u_k) = \underset{u}{\operatorname{argmin}}\, \phi^T(x_k, u_k)\theta \tag{F.32}$$

Then, the optimal control input is given by the following linear controller,

$$u_k = \hat{\pi}(\theta, x_k) = \hat{K}(\theta)x_k. \tag{F.33}$$

3.4 Parameter Update

The coordinate vector θ contains the parameters that characterise the Q-value function. This vector is initially unknown and its parameters have to be identified iteratively by using collected data. A Temporal Difference (TD) algorithm is used for approximating the Q-value function, this algorithm minimises the TD error between successive iterations [21, 29]. The TD with function approximators is formulated as

$$\phi^T(x_k, u_k)\theta_{k+1} = (1-\alpha)\phi^T(x_k, u_k)\theta_k + \alpha\left[\rho(x_k, u_k) + \gamma\phi^T(x_{k+1}, u_k')\theta_k\right] \tag{F.34}$$

where $0 < \alpha < 1$ is a constant learning rate. This method uses batch learning to minimise the TD, it consists of applying a control policy and collecting the measured data. When a batch of data is completed with m iterations, (F.34) is solved using Least Squares as follows

$$\theta_{l+1} = (1-\alpha)\theta_l + \alpha(\Phi_l \Phi_l^T)^{-1}\Phi_l \left[J_l + \gamma \Phi_l'^T \theta_l \right] \tag{F.35}$$

where l is the iteration number, $\Phi_l = [\phi_l, \ldots, \phi_{l+m}]$ and $J_l = [\rho_l, \ldots, \rho_{l+m}]^T$ are a matrix and a vector generated by evaluating the collected data into the polynomial BFs (F.31) and reward functions (F.23) respectively.

4 Policy Supervisor

The controller designed in Section 3 provides an optimal control policy for a continuous state-action space. This policy is obtained from a model-free optimisation where no constraints are considered. However, the operation of physical systems must be restricted to certain areas that guarantee a safe operation. Therefore, a policy supervisor module, also referred to as a safety filter, is introduced in the control structure, see Figure F.3. This module aims to assess the safety of the learned policy, first by

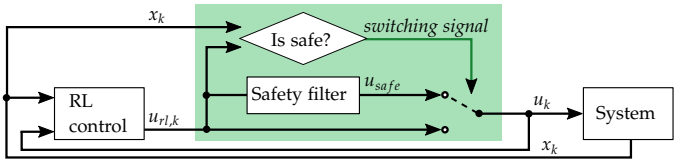


Figure F.3: Block diagram of the control architecture where the RL control is connected in series with a policy supervisor (green). The policy supervisor switches the control action based on a 1-step ahead prediction.

predicting a potential violation of the safety regions and then to provide a safe action. Moreover, this filter is used to restrict the water age in the tank, and it acts as a fallback control when the safety limits are violated.
This paper presents two approaches to evaluate the system's safety: one is based on a nominal model, and the other is based on a combined model that consists of a nominal model and a Gaussian Process (GP) regression.

4.1 Safety regions

In this work, the WDN's safety is built around two factors, the physical limitations of the tank and the water ageing in the tank. The goal of this filter is to repel the operation from these unsafe areas while taking into consideration the control policy (F.33) that now is denoted by u_{rl}. In this paper, the safety regions are defined as

follows,

$$h = \mathcal{H} \triangleq \{h \in \mathbb{R}^{n_h} | \underline{h} \leq h \leq \bar{h}\}, \tag{F.36a}$$

$$\tau = \mathcal{T} \triangleq \{\tau \in \mathbb{R}^{n_t} | \underline{\tau} \leq \tau \leq \bar{\tau}\}, \tag{F.36b}$$

$$u = \mathcal{U} \triangleq \{u \in \mathbb{R}^{n_u} | \underline{u} \leq u \leq \bar{u}\},. \tag{F.36c}$$

where the notation $(\underline{\cdot})$ and $(\bar{\cdot})$ define lower and upper bounds respectively.

4.2 Nominal filter

The filter approach uses a nominal model to predict potential risks in operation and provide a safe action if necessary. The nominal model is a linear model that is built with a priory system knowledge, and therefore it does not describe the exact behaviour of the system. However, it represents the essential system dynamics necessary to correct the path.

Nominal model

This paper refers as nominal model, an imperfect version of the model (F.5), a linear model of the form,

$$\hat{h}_{k+1} = \hat{f}(h_k, u_k) = \hat{A}h_k + \hat{B}u_k + \hat{E}d_{av}, \tag{F.37}$$

where d_{av} is a constant representing the average of the total demand and $\hat{A}$, $\hat{B}$ and $\hat{E}$ are system matrices of the nominal model [24]. The turnover signal is represented by the algebraic equation (F.15). This signal is computed with the flow model. Although the tank inflow is not directly measured, q_k is inferred from the network flows: the pump inflow 1, u_k(decision variable), the average demand, d_{av}(nominal guess) and the pump inflow 2, q_{p2} (measured).

Nominal safe policy

The safety filter is formulated as a constraint optimisation problem:

$$u_{safe} \in \underset{u_k}{\operatorname{argmin}} \quad ||u_{rl,k} - u_k||^2_{Q_1} + ||\tau^* - \hat{\tau}_k||^2_{Q_2} + ||\xi||^2_S \tag{F.38a}$$

$$\text{s.t.} \quad \hat{h}_{k+1} = \hat{f}(h_k, u_k) \tag{F.38b}$$

$$\hat{\tau}_k = \hat{g}(q_k) \tag{F.38c}$$

$$\hat{h}_k \geq \underline{h} - \xi_1 \tag{F.38d}$$

$$\hat{h}_k \leq \bar{h} + \xi_1 \tag{F.38e}$$

$$\hat{\tau}_k \geq \underline{\tau} - \xi_2 \tag{F.38f}$$

$$\underline{u} \leq u_k \leq \bar{u} \tag{F.38g}$$

$$\xi \geq 0 \tag{F.38h}$$

The objective function (F.38a) consists of two terms, the first penalises the difference between the safe and the learning action, u_{rl}, and the second term represents the tracking error between a constant reference value, τ^*, and the turnover of the tank.

Q_1 and Q_2 are constants penalising the aforementioned terms. ξ is a slack variable that relaxes some of the inequality constraints.

Algorithm 8 Safe LS-TD for Q-function - nominal filter.

Input: γ, α, n_s,
Initialisation: $l \leftarrow 0, x_0, \theta_0$ where $\hat{\pi}(\theta_0)$ must be an admissible policy.
repeat at every iteration k = 0,1,2, ...
  apply u_k and measure x_{k+1}
  $\Upsilon_{l_s} \leftarrow \rho(x_k, u_k) + \gamma \hat{Q}(x_{k+1}, \hat{K}_l x_{k+1})$
  if $k = (l+1)n_s$ **then** ▷ Policy update
    $\theta_{l+1} \leftarrow (1-\alpha)\theta_l + \alpha(\Phi_l \Phi_l^T)^{-1}\Phi_l \Upsilon_l$
    $\hat{\pi}(\theta_{l+1}, x) \leftarrow \text{argmin}_u \, \phi(x,u)^T \theta_{l+1}$
    $l \leftarrow l+1$
  end if
  if $\hat{h}_{k+1} \in H$ and $\hat{\tau}_k \in \mathcal{T}$ **then** ▷ Policy supervisor
    $u_k = \hat{K}(\theta_l)x_k + \epsilon_k$
  else
    $u_k = u_{safe} + \epsilon'_k$
  end if
until

The Algorithm 8 is executed in real-time, and it comprises two main objectives, updating the RL policy and rectifying the policy during potential unsafe operation [24]. Remark that the control action in the algorithm consists of the policy and a persistence of excitation signal ϵ_k. This signal facilitates that the collected data contains sufficient information to identify the Q-value function. Safety and learning are conflicting since a high gain in the persistence excitation signal might break the safety guarantees or, on the other hand, strict safety policies might lower the quality of the collected data. Note that a different persistence of excitation signal ϵ'_k is added to the safety actuation. The magnitude of this signal is adjusted to have a low impact on safety but is sufficient to avoid numerical issues during the parameter identification. This work tolerates the risk of unsafe operation around the boundaries for the benefit of the overall performance. In this way, the disruptive effect of the safety actions in the learning is reduced.

4.3 Combined filter

The knowledge of the system is limited in industrial applications, and the filter can become either conservative or reckless depending on how accurate is the system information. The combined filter aims to improve the performance of the previous safety method by reducing the reliance on the calibration of the nominal model. This method for compensating the model uncertainty consists of an imperfect linear model (nominal) and a GP regression. The GP regression term is added to the system dynamics to capture the deviation of the nominal model from the real system.

Combined model

Consider that the discrete system model is formed by two terms as follows,

$$h_{k+1} = \hat{f}(h_k, u_k) + B_r(r(z_k) + w_k), \tag{F.39}$$

where $\hat{f}$ represents the known nominal model and r the non-modelled dynamics of the system, both functions $\hat{f}$ and r are assumed to be differentiable. The observed input vector is built with the augmented system states and control action, $z_k = [x_k^T, u_k^T]^T$, B_r is an index matrix and the random variable and $w_k \sim \mathcal{N}(0, \Sigma^w)$ is i.i.d. process noise.
The residual function $r(z)$ is unknown, however, by using a GP regression, $r(z)$ can be inferred. Then, by combining the nominal model with the GP approximation $\hat{r}(z)$, the system model (F.39) is transformed into,

$$h_{k+1} \approx \hat{f}(h_k, u_k) + B_r\hat{r}(z_k). \tag{F.40}$$

The derivation of the GP approximation $\hat{r}(z)$ with predictive mean and variance equations is given in the Appx.A, then the implementation of the GP training is briefly described.
Subsequently, to compensate the uncertainty in the turnover, this approach considers the level model $\tilde{h}$ to compute (F.42c).

Combined safe policy

To introduce the residual function (F.40) into the safety filter, a reformulation of the safe-optimal control problem is required. In this way, the constraints are represented as chance constraints as follows,

$$Pr\{h \in \mathcal{H}\} \geq p_h, \quad \forall k \tag{F.41a}$$
$$Pr\{\tau \in \mathcal{T}\} \geq p_\tau, \quad \forall k \tag{F.41b}$$

where p_h and p_τ are the satisfaction probabilities. Finding an algebraic solution to the problem is difficult when working with chance constraints. Therefore, a transformation of the chance constraint (F.41) into deterministic equivalents (F.42d), (F.42e), and (F.42f) is performed. Then, by utilising the mean and variance models from (F.50), the optimisation problem is formulated as:

$$u_{safe} \in \underset{u_k}{\text{argmin}} \quad ||u_{rl,k} - u_k||^2_{Q_1} + ||\tau^* - \hat{\tau}_k||^2_{Q_2} + ||\xi||^2_S \tag{F.42a}$$
$$\text{s.t.} \quad \tilde{h}_{k+1} = \hat{f}(h_k, u_k) + B_r\mu_k^r(z_k) \tag{F.42b}$$
$$\hat{\tau}_k = \hat{g}(\tilde{h}) \tag{F.42c}$$
$$\tilde{h}_{k+1} \geq \underline{h} + K_c\sigma^r(z_k) - \xi_1 \tag{F.42d}$$
$$\tilde{h}_{k+1} \leq \bar{h} - K_c\sigma^r(z_k) + \xi_1 \tag{F.42e}$$
$$\hat{\tau}(\tilde{h}_{k+1}) \leq \underline{\tau} - \xi_2 \tag{F.42f}$$
$$\underline{u} \leq u_k \leq \bar{u} \tag{F.42g}$$
$$\xi \geq 0 \tag{F.42h}$$

where the standard deviation $\sigma^r(z_k) = \sqrt{\Sigma^r}(z_k)$ is computed with the variance model (F.50c), K_c represents the confidence gain and ξ is a slack variable that relaxes some of the inequality constraints. Remark that the mean $\mu_k^r(z_k)$ and variance function $\Sigma_k^r(z_k)$ include the decision variable u_k in its input vector z_k.
The Algorithm 8 is executed in real-time with the combined safety filter. Then a prediction is considered safe if meets the following criteria,

$$\underline{h} + K_c\sigma^r(z'_k) \leq \tilde{h}_{k+1}(z'_k) \leq \bar{h} - K_c\sigma^r(z'_k) \tag{F.43a}$$

$$\underline{\tau} \leq \hat{\tau}(\tilde{h}_{k+1}) \tag{F.43b}$$

where the observed input vector z'_k is built with the augmented system states and the RL control action, $z'_k = [x_k^T, u_{rl,k}^T]^T$. Figure F.4 illustrates a normal distribution and how the deviation from the mean value is adjusted with K_c. Remark that this method must fulfil

$$|K_c\sigma^r(z_k)| \leq |\bar{h} - \underline{h}|/2, \tag{F.44}$$

this condition limits the width of the confidence interval based on the distance between upper and lower bounds. The standard deviation, σ^r, is time-variant and its regression model is trained online. This challenges the selection of a suitable gain. Therefore, this paper proposes the use of a *naive* prediction [14] as a protection mechanism that makes the safety problem less sensitive to the GP model. It consists on cancelling the variance term in (F.42d) and (F.42e) by setting $K_c = 0$ when the condition (F.44) is not satisfied.

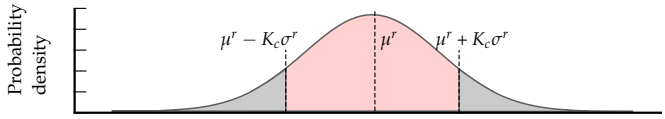


Figure F.4: Scheme of a normal distribution where red represents the area inside the confidence interval of the residual function.

Figure F.5 shows a period of a simulation where the safety filter is intermittently rectifying the system trajectory. The top graph shows the free-exploration prediction $\tilde{h}_{k+1}^{rl}$ with Confidence Interval CI^{rl}, they are computed with the measured state h_k and learning control input u_{rl}, this signal is used for detecting unsafe trajectories. The safe-exploration prediction is represented by $\tilde{h}_{k+1}^{sf}$ and it computed with the state h_k and the safe control input u_{safe}, when no-risk of constraint violation is predicted both signals are aligned. The bottom graph shows the control signals and how the safety filter corrects the control action u_{rl} when the prediction crosses the boundaries. Furthermore, a clear deviation of $\hat{h}_{k+1}^{rl}$ is observed with respect to $\tilde{h}_{k+1}^{rl}$, this is due to the poor calibration of the nominal model. The GP model of the uncertainty is trained online, and the confidence intervals are gradually reduced.

5 Results

The proposed control architecture is validated in a laboratory testbed, first with the nominal and then with the combined safety. The emulated network represents a WDN

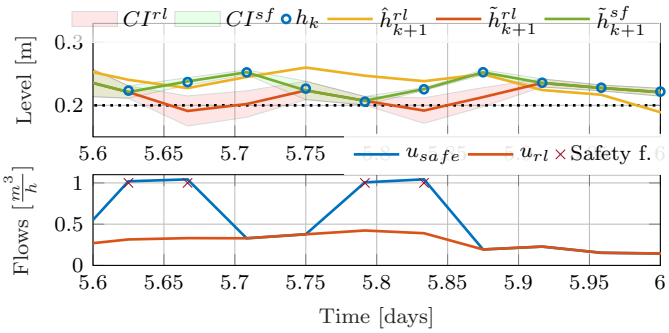


Figure F.5: Simulation of the WDN with a RL controller and a combined filter. (Top) Blue dots represent the real level measurement and the dotted line represents the lower boundary. The nominal model signal $\hat{h}^{rl}_{k+1} = \hat{f}(h_k, u_{rl})$ and the combined model $\tilde{h}^{rl}_{k+1} = \hat{f}(h_k, u_{rl}) + B_r\hat{r}(z'_k)$ are 1-step ahead predictions, and $\tilde{h}^{sf}_{k+1} = \hat{f}(h_k, u_k) + B_r\hat{r}(z_k)$ is the safe prediction. (Bottom) The RL control input u_{rl} and the control input for safe exploration u_{safe}.

with an elevated reservoir and two pressure zones, reproducing the network features of the distribution system shown in Figure F.1.

5.1 Study-case: Bjerringbro

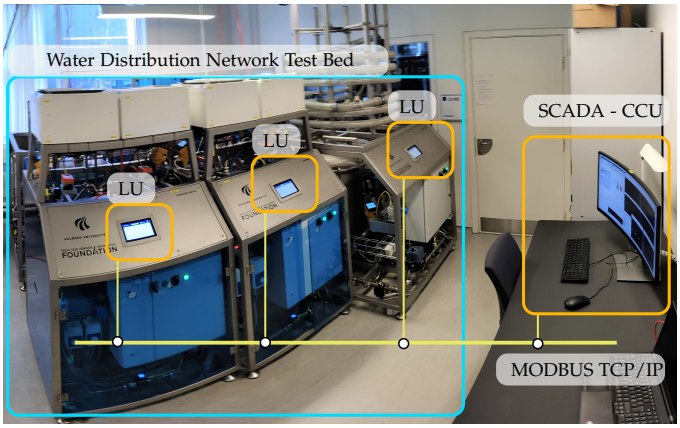


Figure F.6: Picture of the testbed at the SWIL. (Left) The laboratory modules emulating a WDN are placed on the left. (Right) SCADA PC for monitoring and control of the testbed. The communication architecture between Central Control Unit (CCU) and Local Units (LUs) is represented in yellow.

A modular laboratory testbed is built at the Smart Water Infrastructures Laboratory (SWIL) at Aalborg University, where the hydraulic configuration of the study case is emulated.

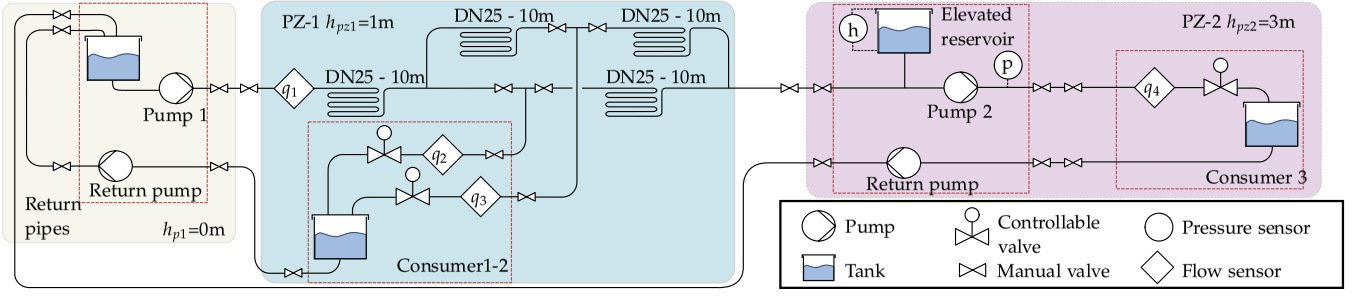


Figure F.7: Scheme of the laboratory setup with a main pumping station (grey), two Pressure Zones (PZ1-blue and PZ2-pink) and an elevated reservoir with a booster station (pump 2). The water is collected at the consumer tanks and recirculated to the supply reservoir. A legend is shown on the bottom-right side.

The actual network and testbed are divided into two Pressure Zones (PZs), PZ-1 is supplied directly from the the pumping station 1 and elevated reservoir, and PZ-2 is located at a higher elevation and is supplied from the elevated reservoir with the pumping station 2, see Figure F.7.
In this testbed, the elevation at the different network areas is emulated by locally controlling the air pressure at the consumption nodes (collection tanks). Note that the hydraulic configuration of the laboratory elevated reservoir, with one inlet, differs from reality, which has two separated inlets (see Figure F.2). This connection makes a major difference in a real network. However, for this experiment, the impact of this connection variation is negligible since the quality of the water is not monitored, and the hydraulic properties are equivalent.
The consumption profiles are emulated with a Fourier series of 2^{nd} order which approximates the consumption pattern from the entire PZs. The magnitude of the signals is reduced to fit the testbed scale, the consumption ratio between PZ1 (q_2 and q_3) and PZ2 (q_4) is kept between 0.73 - 0.27, respectively, reproducing the consumption between areas. Pump 2 has a local pressure control that boosts the pressure at the PZ-2 consumer. Finally, the water is transported to the supply reservoir for re-circulation.
Additionally, local PI controllers regulate the pump inflow at the supply nodes and the valve outflow at the demand nodes. To reduce the impact of these local controllers in the tests, the sampling time for local controllers is 1 second, and for supervisory control (RL) is 60 seconds.
The testbed is equipped with multiple sensors. However, only three measurements are available for the supervisory control, the flow at the pumping units (q_{p1}, q_{p2}) and the differential pressure (h) at the elevated reservoir. The rest of the sensors are used for hardware protection and for monitoring the testbed. The data from the testbed is locally collected at the LUs with Codesys Runtime [1], and it is interfaced with the CCU via TCP/IP Modbus. The proposed control strategy is implemented in Simulink at the CCU. A simple representation of the communication architecture is shown in Figure F.6, a detailed explanation of the laboratory system is presented in [25].
The optimization problems (F.38) and (F.42) are solved with the symbolic framework `CasADI` and a primal-dual interior point solver `IPOPT` is selected to solve the non-linear optimisation problem [3].

5.2 Experimental Results: Nominal safety

The RL control is tested together with a nominal safety filter in this experiment. The nominal model used in the filter is calibrated to fit the laboratory specifications and considers an average consumption, $d_{av} = 0.5$. The turnover computation is computed with the flow model (F.11b), with an average level of h^*=0.3[m] and a filter size n_m=24. The whole control structure has multiple objectives, which in some operational scenarios can be conflicting objectives. This work has relaxed the constraints that represent the water age boundaries (F.38f) to prioritise safety at the tank level boundaries (F.38d)(F.38e) during the transient.
During the first 5 hours, the learning transient is observed in Figure F.8. The operation during this period is driven by the safety filter policy that often corrects the system trajectory. Thus, maintaining the safe operation of the system. During this first part,

the identification of the Q-value function, and subsequently, the optimal policy K, continues despite the system chattering around the safety boundaries, see Figure F.9. After 5 hours of experiment, the RL controller improves and the RL policy becomes dominant. Finally, the operation of the system is driven by a near optimal policy that slowly approached to a steady-state.

5.3 Experimental Results: Combined safety

In this experiment, the RL control is tested with a combined safety filter. In this case, the safety filter is challenged by assuming an inaccurate calibration of the nominal model which considers 0.4 times the tank size A_{er} size and 0.5 times the actual average demand d_{av}. The combined safety filter aims to compensate this poor model calibration. The GP model is initialised with a prior random input and output batches and the confidence gain $K_c = 1$.
The turnover model is computed with level measurements. The graph in Figure F.10 shows turnover signals computed with different models, an error between the approximated flow and level models is observed. This error in the dynamics is especially patent in steady-state conditions where $\hat{m}(q)$, that is computed with a constant d_{av}, fails to predict the periodicity of the disturbance.

Figure F.11 shows the experimental results of a RL controller and a combined safety filter. The graph can be divided into two parts: a learning transient and a steady-state operation. The learning transient comprises a GP model learning and a Q-value function learning. During the first 4 hours, when the GP model is not entirely trained, the predicted model differs from reality. See the bottom graph of Figure F.12. Therefore, the GP prediction provides a wide Confidence Interval (CI) which indicates high uncertainty in the prediction. This lack of confidence makes the safety filter provide a more conservative actuation. This effect is especially noticeable between hours 0 and 2, where the system path is corrected far from the safety boundaries. Once the GP model is improved, hours 3.5-6, the system can safely operate near the boundaries, hence enlarging the exploration limits. During the steady-state operation, after hour 6, the RL controller takes over system policy and the safety filter is only triggered for correcting the approximated turnover signal $\hat{\tau}$. A scaling error is observed between the approximated and actual turnover signal $\hat{\tau}$ and τ. However, this error is considered a minor safety risk since the approximated signal represents a worst-case scenario from a model calibration.
The graph in Figure F.12 shows the convergence of the Algorithm 8 to an optimal policy where the TD error is minimised. Likewise, the training of the GP model with Algorithm 9 shows a reduction of the prediction error between the nominal and combined models.

6 Discussion

Based on the collected results, a brief discussion of the strengths and weaknesses of the presented method is given.

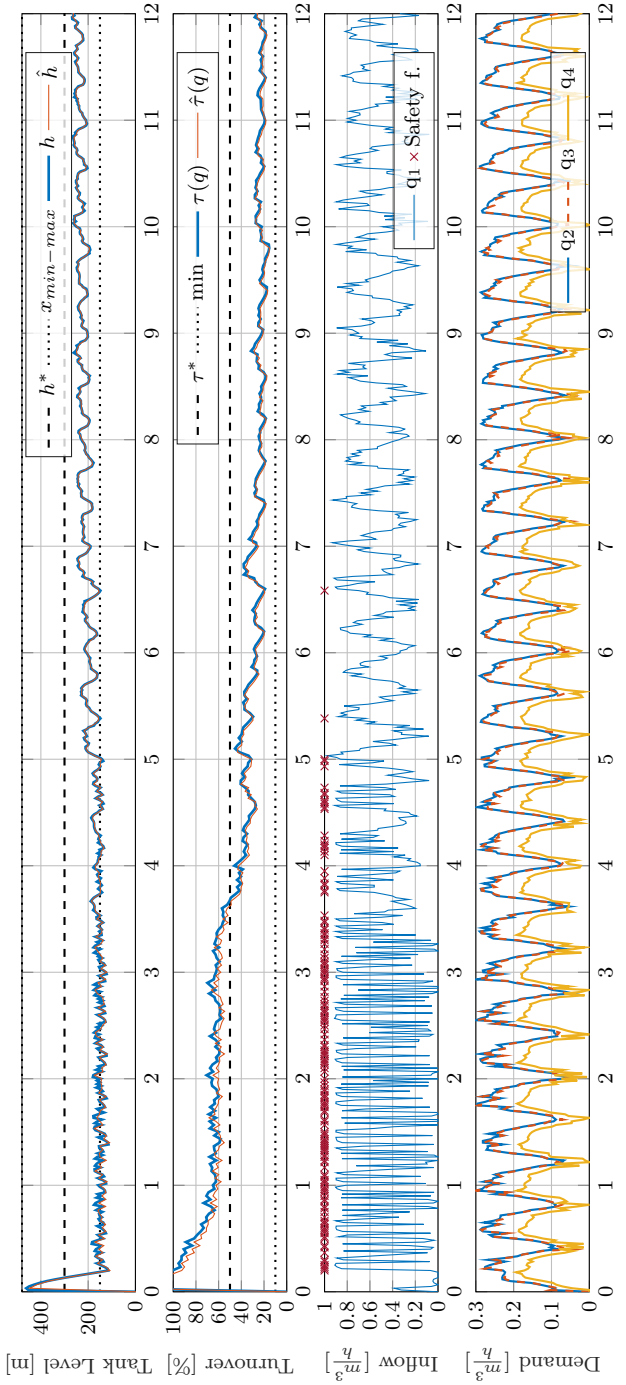


Figure F.8: Tank level regulation. Daily turn over of the tank, real signal τ. The inflow of the pumping station q_1 where the red crosses mark the actuations that are corrected by the safety filter. Demands from PZ1 (q_2 and q_3) and PZ2 (q_4).

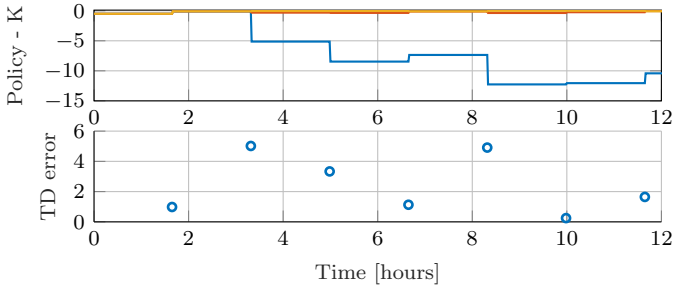


Figure F.9: Experimental results with RL control and nominal safety filter. Top: Transient of the learned policy. Bottom: Temporal Difference.

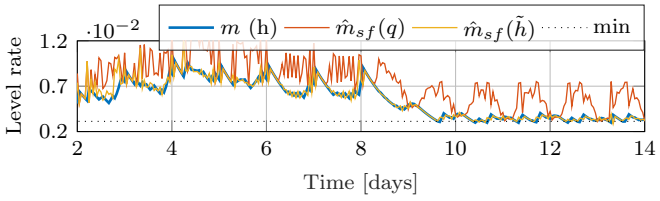


Figure F.10: Simulation of the positive level rate mean: Blue signal (blue) is computed with model (F.13), Red and yellow are computed with the approximated model (F.14), with a nominal flow model (F.11b) and a predicted level (F.11a) respectively.

6.1 Learning Controller

The presented RL method shows to be a suitable solution for obtaining an optimal-adaptive policy. The linear learning structure with polynomial bases provides a preliminary structure to the Q-value function that suits the linear behaviour of the system and quadratic rewards to be learned. To facilitate learning an optimal controller and the quick adaptation to a changing environment. The design criteria in this controller prioritise the simplicity of the solution over performance. Having such a simple learning structure increases the interpretability of the results compared to other black-box methods, thus reducing the commissioning time and cost. Nevertheless, the performance of this linear controller relies on the assumption that linear dynamics describe the system's behaviour. This becomes a strong assumption when the controller is validated against a real setup where, in addition to the linear tank dynamics, the actuator dynamics and communication delays affect the control loop. The convergence of the learning method requires quadratic rewards (objectives) and the operation around an operating point. Additionally, this method provides a continuous adaptation (learning) to the environment. This includes learning during steady-state and safety operation. The method is based on a Least Squares estimate which is sensitive to the quality of the collected data. Therefore, a persistence of excitation signal must ensure a sufficient exploration; this also includes safety saturation periods where the exploration of new areas is limited.

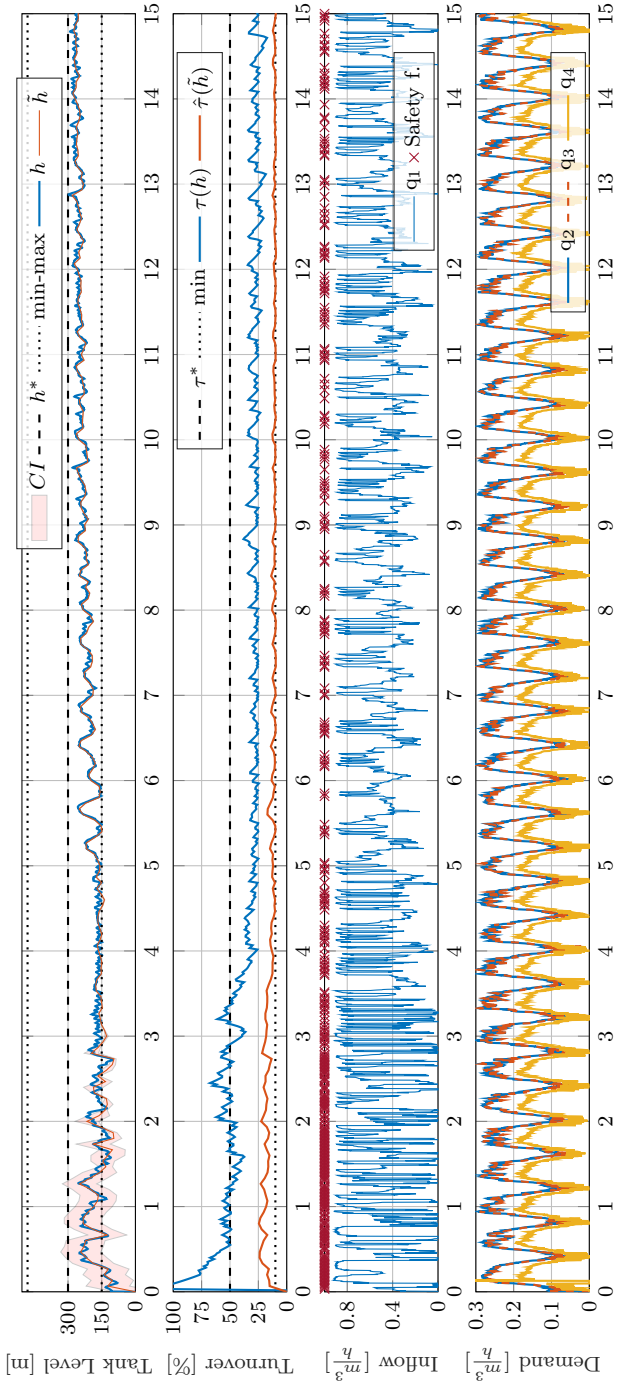


Figure F.11: Tank level regulation. Daily turnover of the tank, real signal τ and its approximation $\hat{\tau}$. Inflow of the pumping station q_1 where the red marks point the actuations that are corrected by the safety filter. Demands from PZ1 (q_2 and q_3) and PZ2 (q_4).

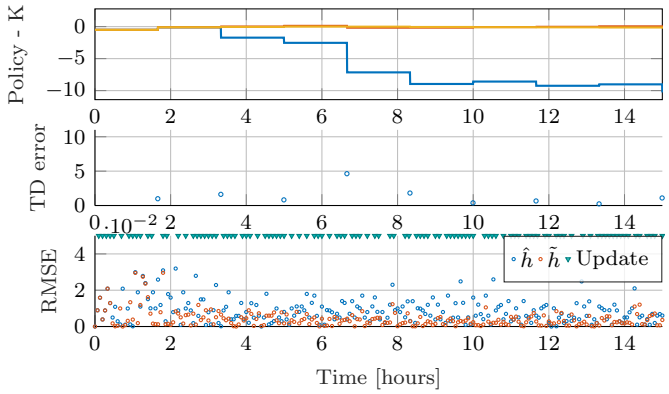


Figure F.12: Experimental results with RL control and combined safety filter. Top: Transient of the learned policy.Middle: The temporal difference computed as $|\hat{K}_l - \hat{K}_{l+1}|$. Bottom: The Root-Mean Square Error (RMSE) of the residual generated by the estimated $\hat{h}$ and $\tilde{h}$ compared with the measured level h, The triangles show the GP model updates during the execution of Algorithm 9.

6.2 Safety Filter

One of the main challenges for deploying learning controllers in real systems is the uncertainty of their policies, especially during the learning transients. This RL optimisation framework does not consider the physical limitations of the system. These safety guarantees are essential when deploying data-driven/learning controllers in large scale systems that require a robust operation like water infrastructures.

The experimental results show the importance of including a safety filter for assisting a learning controller or for any existing controller with uncertain policy. The filter actions maintain the system's operation within a safe level region when the operation is approaching unsafe areas. Additionally, this work includes a management objective in the safety criteria to limit the water age.

A GP regression combined with a nominal model provides a simple method to improve safety in an uncertain environment. The risk assessment of the RL policy is performed, including a variable confidence interval. The confidence gain K_c modifies the learning transient on the application convenience. A high gain represents a lack of confidence in the RL policy and reduces the RL exploration area. However, a poor state-space exploration might lead to sub-optimal RL policies. On the other hand, a low gain relaxes the safety condition and allows the RL policies to explore a wider state-space region. The feasibility of the safety control problem is an important factor to be considered when selecting a suitable K_c, high gain increases the influence of the GP regression model, in particular the variance model. When the GP model is not completely trained, this influence might be an issue, and the safety problem becomes unfeasible.

One of the purposes of having real-time training for the GP, Algorithm 9, is to decrease the confidence interval gradually. Thus, removing the unnecessary safety assistance and entrusting the RL to drive the system's policy. In case of reaching large confi-

dence intervals that conflict with the safety optimisation, the algorithm neglects the variance to find a safe-robust solution. The experimental results conclude that providing robust safety while learning the GP model is challenging. Insufficient training of a GP model brings a high uncertainty in the predicted mean and variance. This study reduces the variance dynamics (Confidence Interval) impact by reducing the confidence gain K_c to facilitate the computation of a feasible solution. This safety method is a 1-step ahead optimisation that does not imply significant computational effort. Moreover, the GP model utilises the same collected data for learning the Q-value function x_k and u_k to facilitate the implementation.

Finally, a brief discussion analysing the joint performance of the learning controller and safety block is given. The use of a safety filter allows the inclusion of management objectives that cannot be incorporated in a linear learning architecture. Moreover, by strategically formulating RL objectives in the safety objective function, the filter can be utilised as a learning guideline for an RL controller that does not have any knowledge of the system. However, RL and safety are decoupled blocks that have different control objectives. Therefore, the formulation of these objectives must be balanced such that the filter is not dominant with respect to the existing controller and vice versa. The control actions' smoothness also conflicts with the learning performance that requires persistent variations to identify the Q-value function.

7 Conclusion and future work

This paper proposes an optimal-adaptive control solution for a WDN with an elevated reservoir. The management of this network comprises two main operation objectives: controlling the network pressure and water age by regulating the tank's level and turnover. The control method consists of an RL linear controller that provides the primary policy and a policy supervisor that predicts and corrects potential violations of the safety boundaries. The predictions are based on a linear nominal model. The performance of this filter is subject to the calibration of the model. Therefore, a GP regression is included in the filter to compensate for incorrect model calibration and include non-linearities that a linear model cannot capture. The proposed control strategy is validated in a laboratory setup that emulates the WDN dynamics. The experimental results reflect the benefits and the limitations of the method.
Although this paper presents promising results that support the implementation of safe learning techniques in industrial applications, this method is challenged by several factors. Future development of this method must consider a non-linear learning architecture that allows the formulation of non-linear objectives in the reward. It allows learning an optimal policy that includes important management objectives such as water age or operational costs.

Acknowledgment

Financial support from Poul Due Jensen Foundation (Grundfos Foundation) for this research is gratefully acknowledged.

A Gaussian Process Regression

This model is inspired by the formulation presented in [11] where the uncertainty between an imperfect nominal model and the real system is modelled with a Gaussian Process (GP) regression. Then, by applying a GP regression, this method aims to learn the uncertainty term $r(z)$ represented in (F.39) as follows. First, the residual is expressed as,

$$y_k = r(x_k, u_k) + w_k = B_r^{\dagger}(\underbrace{h_{k+1}}_{measure} - \underbrace{\hat{f}(x_k, u_k)}_{nominal}), \tag{F.45}$$

where $B_r^{\dagger}$ is the Moore-Penrose pseudo-inverse of B_r. Then, consider a training data set $\mathcal{D}$ that consists of M observations,

$$\begin{aligned} \mathcal{D} = \{ \mathbf{y} = [y_1, \dots, y_M]^T \in \mathbb{R}^M \\ \mathbf{z} = [z_1, \dots, z_M]^T \in \mathbb{R}^{M \times n_z} \} \end{aligned} \tag{F.46}$$

where z denotes an input vector and y a scalar output (target). This definition is performed by assuming that each of elements y of the output vector are independent for a given input data z_k. Then, by giving a GP prior on r with kernel $k(\cdot, \cdot)$ and prior mean function is zero,

$$y \sim \mathcal{N}(0, K_{\mathbf{zz}} + I\sigma^2). \tag{F.47}$$

The result is a normally distributed measurement where $K_{\mathbf{zz}}$ is the Gram matrix of the data points such that $K_{ij} = k(z_i, z_j)$, the selection of the kernel k structure and its parameterisation determines the distribution of the predicted output. In this paper, the selected kernel is the squared exponential kernel function,

$$k(z_i, z_j) = \sigma_f^2 \exp(-\frac{1}{2}(z_i - z_j)^T L^{-1}(z_i - z_j)), \tag{F.48}$$

where L is a positive diagonal length scale matrix and σ_f^2 the signal variance. This kernel is selected based on the domain knowledge since the dynamics of the system present a continuous and smooth behaviour.
The joint distribution of the training data $\mathbf{z}$ and the test data z_* is

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K_{\mathbf{zz}} + I\sigma^2 & K_{\mathbf{z}z_*} \\ K_{z_*\mathbf{z}} & K_{z_*z_*} \end{bmatrix} \right), \tag{F.49}$$

where $[K_{\mathbf{z}z_*}]_j = k(\mathbf{z}_j, z_*)$, $K_{z_*\mathbf{z}} = K_{\mathbf{z}z_*}^T$, and similarly $K_{z_*z_*} = k(z_*, z_*)$. The resulting conditional distribution of the uncertainty residual is Gaussian [8],

$$\hat{r}(y_*|\mathbf{y}) = \mathcal{N}(\mu^r(z_*), \Sigma^r(z_*)), \tag{F.50a}$$

$$\mu^r(z_*) = K_{z_*\mathbf{z}}(K_{\mathbf{zz}} + I\sigma^2)^{-1}\mathbf{y}, \tag{F.50b}$$

$$\Sigma^r(z_*) = K_{z_*z_*} - K_{z_*\mathbf{z}}(K_{\mathbf{zz}} + I\sigma^2)^{-1}K_{\mathbf{z}z_*} \tag{F.50c}$$

where $\mu^p(z_*)$ and $\Sigma^p(z_*)$ are mean and variances of the GP.
The GP model is trained with measured data, the collected data batch shares the same vector structure as the RL batch, it is built with the pair augmented state vector x and control action u. The training of the GP is performed by executing Algorithm 9 in real-time, and the hyper-parameters of the GP are obtained with the MATLAB function `fitrgp()`. The algorithm needs to be initialised with prior data $\mathbf{z}_0$ and $\mathbf{y}_0$, the number of new samples per update n_{gp} and the threshold e^*. This threshold represents the minimum deviation between the measured variable h_k and estimated $\tilde{h}_k(z_k)$, with $z_k = [x_k^T, u_k^T]^T$. The collected data is stored between updates in a Last In First Out stack (LIFO).

Algorithm 9 Training of the GP model.

Input: n_{gp}, e^*
Initialisation: $[\sigma_{f0}, L_0, \sigma_0] \leftarrow$ `fitrgp` $(\mathbf{z}_0, \mathbf{y}_0)$
repeat at every iteration k = 1,2, ...
    collect data $\hat{h}_k$, z_k and y_k
    e_k = `RMSE`$(h_k$ - $\tilde{h}_k)$
    if $e_k \geq e^*$ **then** ▷ Collect data
        save z_k and y_k in stack $\mathbf{z}_j$ and $\mathbf{y}_j$
        if $k = (j+1)n_{gp}$ **then** ▷ GP update
            $[\sigma_f, L, \sigma] \leftarrow$ `fitrgp` $(\mathbf{z}, \mathbf{y})$
            j = j+1
        end if
    end if
until

References

[1] 3S-Smart Software Solutions GmbH, "Codesys." [Online]. Available: https://www.codesys.com

[2] C. Agudelo-Vera, S. Avvedimento, J. Boxall, E. Creaco, H. de Kater, A. Di Nardo, A. Djukic, I. Douterelo, K. E. Fish, P. L. Iglesias Rey, N. Jacimovic, H. E. Jacobs, Z. Kapelan, J. Martinez Solano, C. Montoya Pachongo, O. Piller, C. Quintiliani, J. Ručka, L. Tuhovčák, and M. Blokker, "Drinking water temperature around the globe: Understanding, policies, challenges and opportunities," *Water*, vol. 12, no. 4, 2020.

[3] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl, "CasADi – A software framework for nonlinear optimization and optimal control," *Mathematical Programming Computation*, In Press, 2018.

[4] A. Bahadori, "Chapter 8 - water supply and distribution systems," in *Essentials of Oil and Gas Utilities*, A. Bahadori, Ed. Gulf Professional Publishing, 2016, pp. 225–328.

[5] U. S. Environmental Protection Agency, "Finished water storage facilities," EPA, Technical Report, 2002. [Online]. Available: https://www.epa.gov/

[6] F. García-Ávila, C. Sánchez-Alvarracín, M. Cadme-Galabay, J. Conchado-Martínez, G. García-Mera, and C. Zhindón-Arévalo, "Relationship between chlorine decay and temperature in the drinking water," *MethodsX*, vol. 7, p. 101002, 2020.

[7] V. Gauthier, M.-C. Besner, B. Barbeau, R. Millette, and M. Prévost, "Storage tank management to improve drinking water quality: case study," *Journal of water resources planning and management*, vol. 126, no. 4, pp. 221–228, 2000.

[8] A. Girard, C. E. Rasmussen, J. Q. n. Candela, and R. Murray-Smith, "Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, ser. NIPS'02. Cambridge, MA, USA: MIT Press, 2002, p. 545–552.

[9] W. M. Grayman, L. A. Rossman, R. A. Deininger, C. D. Smith, C. N. Arnold, and J. F. Smith, "Mixing and aging of water in distribution system storage facilities," *Journal-American Water Works Association*, vol. 96, no. 9, pp. 70–80, 2004.

[10] J. Grosso, C. Ocampo-Martínez, V. Puig, and B. Joseph, "Chance-constrained model predictive control for drinking water networks," *Journal of process control*, vol. 24, no. 5, pp. 504–516, 2014.

[11] L. Hewing, A. Liniger, and M. N. Zeilinger, "Cautious nmpc with gaussian process dynamics for autonomous miniature race cars," in *2018 European Control Conference (ECC)*, 2018, pp. 1341–1348.

[12] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 269–296, 2020.

[13] G. Kirmeyer, L. Kirby, B. Murphy, P. Noran, K. Martel, T. Lund, J. Anderson, and R. Medhurst, "Maintaining and operating finished water storage facilities," *AWWA and AwwaRF, Denver*, 1999.

[14] J. Kocijan, *Modelling and control of dynamic systems using Gaussian process models*. Springer, 2016.

[15] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[16] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement Learning for Partially Observable Dynamic Processes: Adaptive Dynamic Programming Using Measured Output Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, 2011.

[17] C. Ocampo-Martinez, V. Puig, G. Cembrano, and J. Quevedo, "Application of Predictive Control Strategies to the Management of Complex Networks in the Urban Water Cycle," *Control Systems, IEEE*, 2013.

[18] G. S. Sankar, S. P. M. Kumar, S. Narasimhan, S. Narasimhan, and S. M. Bhallamudi, "Optimal control of water distribution networks with storage facilities," *Journal of Process Control*, vol. 32, pp. 127–137, 2015.

[19] J. Shin, T. A. Badgwell, K.-H. Liu, and J. H. Lee, "Reinforcement learning – overview of recent progress and implications for process control," *Computers & Chemical Engineering*, 2019.

[20] J. Stockmarr, "Groundwater quality monitoring in denmark," *GEUS Bulletin*, vol. 7, pp. 33–36, 2005.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

[22] B. Tearle, K. P. Wabersich, A. Carron, and M. N. Zeilinger, "A predictive safety filter for learning-based racing control," *arXiv preprint arXiv:2102.11907*, 2021.

[23] J. Val, R. Wisniewski, and C. S. Kallesøe, "Reinforcement learning control for water distribution networks with periodic disturbances," in *2021 American Control Conference (ACC)*, 2021, pp. 1010–1015.

[24] J. Val, R. Wisniewski, and C. Kallesøe, "Safe reinforcement learning control for water distribution networks," in *Conference on Control Technology and Applications*. United States: IEEE, 2021.

[25] J. Val Ledesma, R. Wisniewski, and C. S. Kallesøe, "Smart water infrastructures laboratory: Reconfigurable test-beds for research in water infrastructures management," *Water*, vol. 13, no. 13, 2021. [Online]. Available: https://www.mdpi.com/2073-4441/13/13/1875

[26] Y. Wang, C. Ocampo-Martinez, and V. Puig, "Robust model predictive control based on gaussian processes: Application to drinking water networks," in *2015 European Control Conference (ECC)*, 2015, pp. 3292–3297.

[27] ——, "Stochastic model predictive control based on gaussian processes applied to drinking water networks," *IET Control Theory and Applications*, vol. 10, pp. 947 – 955, 05 2016.

[28] Y. Wang, V. Puig, and G. Cembraño, "Non-linear economic model predictive control of water distribution networks," *Journal of Process Control*, vol. 56, pp. 23–34, 2017.

[29] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[30] M. Zanon and S. Gros, "Safe reinforcement learning using robust mpc," *IEEE Transactions on Automatic Control*, 2020.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-894-6

AALBORG UNIVERSITY PRESS