Applied Intelligence
https://doi.org/10.1007/s10489-024-05452-8

Optimal operation of reverse osmosis desalination process with deep reinforcement learning methods

Arash Golabi[1] · Abdelkarim Erradi[1] · Hazim Qiblawey[2] · Ashraf Tantawy[3] · Ahmed Bensaid[1] · Khaled Shaban[1]

Accepted: 7 April 2024
© The Author(s) 2024

Abstract

The reverse osmosis (RO) process is a well-established desalination technology, wherein energy-efficient techniques and advanced process control methods significantly reduce production costs. This study proposes an optimal real-time management method to minimize the total daily operation cost of an RO desalination plant, integrating a storage tank system to meet varying daily freshwater demand. Utilizing the dynamic model of the RO process, a cascade structure with two reinforcement learning (RL) agents, namely the deep deterministic policy gradient (DDPG) and deep Q-Network (DQN), is developed to optimize the operation of the RO plant. The DDPG agent, manipulating the high-pressure pump, controls the permeate flow rate to track a reference setpoint value. Simultaneously, the DQN agent selects the optimal setpoint value and communicates it to the DDPG controller to minimize the plant's operation cost. Monitoring storage tanks, permeate flow rates, and water demand enables the DQN agent to determine the required amount of permeate water, optimizing water quality and energy consumption. Additionally, the DQN agent monitors the storage tank's water level to prevent overflow or underflow of permeate water. Simulation results demonstrate the effectiveness and practicality of the designed RL agents.

Keywords Reinforcement learning · Reverse osmosis · Desalination process · Dynamic modeling · Deep deterministic policy gradient · Deep Q-Network · Optimal management · Data-driven controller

1 Introduction

One of the well-established and widely used desalination processes for brackish and salty water is reverse osmosis (RO). RO aims to reduce the production costs of clean water through energy-efficient techniques and advanced control methods [1–3]. Representing 65% of the total installed capacity of desalination technologies globally, RO plays a significant role in the world's water desalination industry [4].

In the RO process, energy consumption significantly contributes to the cost of freshwater production. Optimizing the operation of the RO process to enhance performance and reduce energy consumption has recently garnered considerable attention [5–7]. Employing methodologies from process systems engineering, researchers have developed technologies to improve membrane processes, implement optimal designs, and reduce power consumption in seawater desalination systems [8, 9]. In a comprehensive review [10], the impact of RO membrane element performance on

✉ Arash Golabi
arash.golabi@qu.edu.qa

Abdelkarim Erradi
erradi@qu.edu.qa

Hazim Qiblawey
hazim@qu.edu.qa

Ashraf Tantawy
ashraf.tantavy@dmu.ac.uk

Ahmed Bensaid
abensaid@qu.edu.qa

Khaled Shaban
khaled.shaban@qu.edu.qa

[1] Department of Computer Science and Engineering, Qatar University, Doha, Qatar

[2] Department of Chemical Engineering, Qatar University, Doha, Qatar

[3] School of Computer Science and Informatics, De Montfort University, Leicester, UK

Published online: 13 May 2024

specific energy consumption in the RO process is thoroughly investigated. Other studies, such as [11, 12], focus on reducing energy costs in seawater RO, adjusting the relationship between water production and demand according to load and electricity price fluctuations. In [13], a fuzzy logic-based control system optimizes RO operational costs based on feedwater electrical conductivity, temperature as input values, and the RO recovery setpoint as a control action. Similarly, [14] designs and operates an RO system according to daily water demands, adjusting to changes in seawater temperature to achieve an optimal operational policy. Experimental examination of an RO desalination plant under different working conditions is presented in [15], focusing on specific energy consumption and water production costs. Additionally, [16] introduces an optimal control objective for energy management in the RO process using a hybrid energy system and a deep reinforcement learning algorithm. Finally, [17] proposes a modeling and optimization strategy for achieving optimal operation in an industrial RO plant.

For the purpose of energy optimization, understanding the RO process is crucial, and developing an RO model is essential to facilitate optimization and reduce energy consumption [18]. Creating an RO model involves establishing correlations between key operational conditions and performance indicators to comprehend the mechanism and evaluate RO membrane performance. With the developed RO process model as a foundation, simulation and optimization can be performed. There are three methods for obtaining an RO process model: membrane transport, lumped parameters, and data-driven models [19]. The solution-diffusion transport mechanism is widely accepted as a mathematical model for solute and solvent transport in the RO process [20] and can be employed to construct the RO model. In recent years, several studies have utilized transport membrane modeling for the optimization and performance assessment of the RO process [21–23]. In [24], mathematical models are explained to describe the steady-state and transient behavior of the RO desalination process.

Linear and nonlinear modeling of the RO process provides the foundation for developing an efficient controller to maintain freshwater production while minimizing operating costs. Both linear and nonlinear dynamic models for RO processes have been introduced in previous studies [25]. The design of a dynamic model for RO desalination, focusing on spiral-wound membrane modules and the corresponding controller, is detailed in [26]. Using a simplified functional decomposition method, [27] investigates the servo and regulatory performance of different PID loops for controlling the permeate flow rate in the RO desalination process. The literature features various contributions to RO system process control. In [28], for instance, a control system based on optimization is designed and implemented to optimize energy efficiency in an experimental RO membrane water desalination process. Two PID controllers are proposed in [29] for controlling the flux and conductivity of an RO plant using controllers based on the whale optimization algorithm. In [18], control strategies such as internal multi-loop model control and proportional-integral control are implemented to simulate the RO desalination process for both servo and regulatory purposes.

Data-driven methods, based on machine learning (ML) techniques, are increasingly becoming flexible tools in RO process systems [30]. In [22, 31], a review focuses on recent trends and developments, primarily emphasizing the modeling and simulation of RO plants using Artificial Neural Networks (ANN) to address challenges in membrane-based desalination systems. Another study, [32], employs an ANN to predict and forecast water resource variables. The review in [33] elucidates various membrane-based water treatment designs, along with plant performances, utilizing Artificial Intelligence (AI) methods to reduce waste generation and enable cleaner production. [34] explores an NN-based method to predict the dynamic water permeability constant for an RO desalination plant under fouling conditions. For small-scale prototype operation in a seawater RO desalination plant with fluctuating power input, [35] incorporates ANN models into the control system.

In RO desalination systems, dynamics are highly nonlinear, constrained, and subject to uncertainties such as membrane fouling and varying feed water parameters. These factors contribute to the complexity of creating a mathematical model for an RO system. Consequently, designing an optimal controller for managing RO desalination systems poses a significant challenge [36]. Considering this, a data-driven approach for control and optimal management of an RO process based on the available data emerges as a promising solution. Reinforcement Learning (RL) offers a unique approach, as it leverages the concept of learning controllers and can acquire high-quality control behavior by learning from scratch through interactions with the process [37]. The application of RL, which uses data to learn optimum control policies, holds potential for addressing the complexities of RO systems [38]. In RL, agents undergo training as they interact with their environment, performing different actions that lead to positive or negative rewards based on the states they reach. This principle can be applied to complex processes like the RO process, allowing the system to learn intricate behaviors and optimal control policies through experience. RL has gained significant attention for its effectiveness and widespread use in solving problems with discrete and continuous states and action spaces, including real-world scenarios like chemical process control problems [39–42].

The methods proposed in [7, 12, 43, 44] focus on optimizing the daily operation of the RO plant. They employ a nonlinear solver to address the discretized large-scale nonlinear programming. However, when faced with new situations, such as uncertain water demand, these methods need to resolve nonlinear problems, making them less suitable for real-time management and control issues. It's crucial to highlight that fully discretizing both differential and algebraic variables results in a large-scale problem, posing a significant challenge [7]. Additionally, these studies rely on the steady-state model of the RO process and overlook the controller's impact during the transition regime from one steady-state to another.

In this study, we propose the development of a data-driven framework using RL methods to control and optimize the daily operation of an RO desalination plant in real-time. The main objectives of the proposed framework include controlling permeate flow rate, improving energy efficiency, ensuring permeate water quality, and maximizing plant availability during real-time management of the RO plant. Initially, a data-driven controller based on DDPG is designed to regulate the permeate flow rate of the RO plant. Notably, the DDPG controller developed in this study adopts a multi-step tracking approach for controlling the permeate flow rate in a dynamic model of an RO plant, distinguishing it from [40]. In the subsequent step, a deep Q-Network (DQN) is designed to monitor and optimize the RO plant in real-time by providing setpoints for the controller. Specifically, the DQN agent is structured to complement existing control systems without substantial modification, generating optimized control setpoints. Consequently, the proposed approach offers a flexible and practical solution that can effectively enhance the performance of existing RO plants. Moreover, the performance of the DDPG controller is compared with that of a PID controller, and the performance of the DQN is assessed against various RL algorithms in the simulation and discussion section.

The main objective of this work is to develop an integrated framework for real-time control, management, and optimization using RL methods to minimize the total daily operation cost of a simulated RO desalination plant with a water storage tank system, meeting daily variable freshwater demand. Two RL agents, based on DDPG and DQN algorithms, are designed to optimize the RO plant's real-time operation. The DDPG method controls the permeate flow rate by adjusting a high-pressure pump to reach a reference setpoint determined by a decision-maker. Trained through a reward function that minimizes the error between the reference value and output permeate water, the DDPG agent regulates the flow rate effectively. In the cascade structure, the DQN agent selects optimal setpoint values, minimizing operational costs by determining the permeate water amount while considering water quality in terms of permeate concentration and monitoring the storage tank's water level to prevent overflow or underflow. The reward function, focusing on minimizing daily operating costs, preventing underflow or overflow, and maintaining water quality, guides the DQN agent in learning an optimal policy during training. Significantly, the flexibility of the DQN agent and its compatibility with existing control systems make it a practical solution to enhance the performance of established RO plants without requiring substantial modifications.

The remainder of the article is structured as follows. In Section 2, the desalination process and problem description are presented. The modeling discussion is provided in Section 3, where the mathematical model of desalination process components is explained. Section 4 discusses the optimal operation of the desalination plant and the design of the RL agents. Simulation and discussion are elaborated in Section 5, where the daily operation of the RO plant with the designed RL agents has been investigated. In Section 6, the concluding remarks are provided. A summary of the abbreviations and the mathematical symbols used in this paper is presented in Table A1 in Appendix A.

2 Desalination Process and problem description

This section explores the model structure of the RO desalination process, addressing the efficient operation of the RO plant to meet water demand and optimize energy consumption.

The schematic diagram of the considered RO plant is depicted in Fig. 1, illustrating the desalination process. The system comprises an RO system, a high-pressure pump with a controller regulating feed pressure for the RO system, a storage tank system, and an energy recovery device.

The RO system utilizes high-pressure pumps to convert saline water into freshwater by overcoming osmotic pressure through a semi-permeable membrane [45]. The high-pressure pump generates the necessary pressure to force saline water against the membrane, separating freshwater from dissolved materials like salt. The resulting desalinated water, known as permeate or product water, is demineralized. Meanwhile, the brine water, containing concentrated contaminants, remains after the filtration process. Energy recovery devices (ERDs) play a crucial role in recovering energy from the brine stream and transferring it to a high-pressure pump, leading to significantly reduced energy consumption.

One of the key components of the RO plant is the controller. It should be meticulously designed to regulate the permeate flow rate of the RO system and adjust system pressure for optimal permeate water production. While a traditional controller like a PID controller can be employed for permeate water regulation [1], a data-driven controller holds promise due to the highly nonlinear and complex nature of the RO process, coupled with uncertainties such as membrane fouling. Moreover, the controller often struggles to handle significant variations in water demand, especially when the

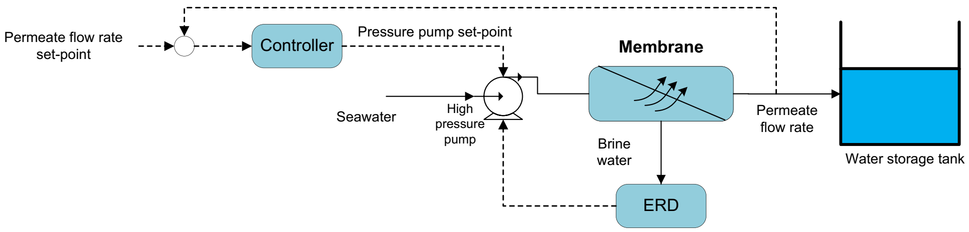


Fig. 1 Schematic diagram of RO desalination system

demand exceeds the desalination plant's capacity. To address this issue, a storage tank system is incorporated into the system layout, enabling the system to manage substantial variations in water demand effectively [44].

On the other hand, the control part can maintain a constant permeate production rate based on a reference tracking value. However, the controller itself cannot directly optimize the operation cost and energy usage of the RO process. The controller module is specifically designed to track commands for generating permeate flow rates issued by a designated command controller. However, it lacks autonomy in generating setpoints independently. Therefore, a supervisory tool is needed for efficiently monitoring the operation of the RO system by utilizing observational data and facilitating the provision of setpoints to the controller. To this goal, a distinct optimizer or intelligent agent based on AI algorithms has been developed to manage the RO plant using real-time information for determining permeate flow rates. The design of this agent considers not only energy consumption but also addresses significant variations in freshwater demand to prevent overflow or underflow in the storage tank system. The optimal management of the daily operation of the RO plant is structured according to the hierarchical framework depicted in Fig. 2. To achieve this, both the controller and the AI-based optimizer are meticulously designed and tuned to meet the specified demands. The cascade structure of the proposed framework shown in Fig. 2 gives the flexibility to develop the AI-based optimizer and the controller independently. By designing the optimizer itself, it is possible to provide optimal setpoints for existing controllers, making it an effective solution for improving the performance of existing RO plants.Moreover, by changing the storage tank system or water demand data, the controller does not need to be altered or redesigned as they only require observations of the current system state, not explicit dependencies on system parameters.

3 Mathematical model of desalination process components

Establishing a model for the RO plant is essential for designing the controller and optimizer. This section outlines the mathematical models for the components of the RO desalination process. Initially, a mathematical model for the RO membrane is presented, followed by data on the daily demand for freshwater. Then, a mathematical model for the storage tank system is explained.

3.1 RO membrane model

A model that adequately describes the operation of RO membranes is an essential step in designing the controller and optimizer for the RO process. In the following, the equations for describing the RO membrane are provided.

Balance equations for dynamic variable brine mass, brine concentration, and permeate concentration are obtained as follows [46]:

$$\frac{dM_{\mathrm{b}}}{dt} = \mathcal{Q}_{\mathrm{f}} - \mathcal{Q}_{\mathrm{b}} - \mathcal{Q}_{\mathrm{m}}$$

$$\frac{dC_{\mathrm{b}}}{dt} = \frac{1}{M_{\mathrm{b}}}\left[\mathcal{Q}_{\mathrm{f}}\left(\mathcal{Q}_{\mathrm{f}} - \mathcal{Q}_{\mathrm{b}}\right) - \mathcal{Q}_{\mathrm{m}}\left(C_{\mathrm{m}} - C_{\mathrm{b}}\right)\right]$$

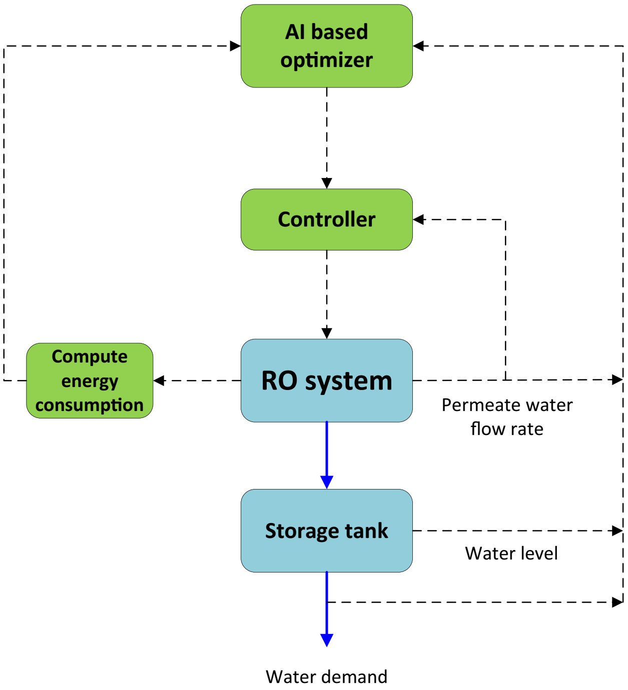


Fig. 2 Hierarchical structure of RO desalination system with controller and optimizer

$$\frac{dC_\mathrm{p}}{dt} = \frac{1}{M_\mathrm{p}}\left[\mathcal{Q}_\mathrm{m}C_\mathrm{m} - \mathcal{Q}_\mathrm{p}C_\mathrm{p}\right] \tag{1}$$

where $M_b[kg]$ and $M_p[kg]$ are brine and permeate mass, $\mathcal{Q}_f[kg/s]$, $\mathcal{Q}_b[kg/s]$ and $\mathcal{Q}_m[kg/s]$ are feed, brine and membrane mass flow rate. $C_b[kg/m^3]$, $C_p[kg/m^3]$ and $C_m[kg/m^3]$ are brine, permeate and membrane concentration and mass flow rate at permeate side is $\mathcal{Q}_p = \mathcal{Q}_m$. With a reject valve, the brine mass flow rate can be computed by the following equation [25].

$$\mathcal{Q}_b = \mathcal{Q}_\mathrm{b}^{\max} - \left(\frac{\mathcal{Q}_\mathrm{b}^{\max} - \mathcal{Q}_\mathrm{b}^{\min}}{a_\mathrm{v}^{\max} - a_\mathrm{v}^{\min}}\right) a_\mathrm{v}^{\max} + \left(\frac{\mathcal{Q}_\mathrm{b}^{\max} - \mathcal{Q}_\mathrm{b}^{\min}}{a_\mathrm{v}^{\max} - a_\mathrm{v}^{\min}}\right) a_\mathrm{v} \tag{2}$$

where a_v is the valve opening, $a_\mathrm{v}^{\min}$ and $a_\mathrm{v}^{\max}$ are the minimum and maximum percentage of the reject valve opening and:

$$\mathcal{Q}_b^{\min} = \mathcal{Q}_f - \mathcal{Q}_p^{\max},\ \ \mathcal{Q}_b^{\max} = \mathcal{Q}_f - \mathcal{Q}_p^{\min}$$
$$\mathcal{Q}_p^{\max} = \frac{a_{rec}^{\max}}{100}\mathcal{Q}_f,\ \ \mathcal{Q}_p^{\min} = \frac{a_{rec}^{\min}}{100}\mathcal{Q}_f$$

with $a_{rec}^{\max}$ and $a_{rec}^{\min}$ as the maximum and minimum recovery rate, respectively. The brine pressure $P_b[kPa]$ is calculated with valve reject valve rangeability R based on the following equation [25]:

$$P_b = R^{2(1-(\frac{a_v}{100}))}\mathcal{Q}_b^2 + P_{bo} \tag{3}$$

The permeate flow rate $\mathcal{Q}_p[kg/s]$ is a function of the difference between trans-membrane pressure and net osmotic pressure and is computed as follows:

$$\mathcal{Q}_p = A_w A_{em} T_{cp}(\Delta P - \beta\Delta\Pi),$$
$$\mathcal{Q}_s = B_s A_{em} T_{cs}\left(\beta\bar{C} - C_p\right) \tag{4}$$

where A_w is permeability of membrane and B_s is the membrane salt permeability, $A_{em} = n_v n_e A_m$ with n_v as pressure vessel number, n_e elements number in a pressure vessel, $A_m[m^2]$ as membrane active surface area and β is concentration polarization factor. T_{cp} and T_{cs} are temperature correction factor. Trans-membrane pressure and Osmotic pressure are obtained by the following equation [45]:

$$\Delta P = \frac{P_f + P_b}{2} - P_p,$$
$$\Delta\Pi = \frac{\Pi_f + \Pi_b}{2} - \Pi_p, \tag{5}$$

with $\Pi_i = 75.84C_i$ for $i \in \{f, b, p\}$. $\bar{C}[kg/m^3]$ is average of feedwater and brine concentration obtained by the following equation [45]:

$$\bar{C} = \frac{C_\mathrm{f} + C_\mathrm{b}}{2} \tag{6}$$

The temperature coefficient factors T_cp and T_cs are obtained as below:

$$T_\mathrm{cp} = \exp\left(a_T\frac{T_\mathrm{f} - T_\mathrm{ref}}{T_\mathrm{f}}\right),$$
$$T_\mathrm{cs} = \exp\left(b_T\frac{T_\mathrm{f} - T_\mathrm{ref}}{T_\mathrm{f}}\right) \tag{7}$$

where T_ref is the reference temperature, a_T and b_T is membrane water passage temperature constant and membrane salt passage temperature constant. Membrane surface concentration C_m is obtained by the following equation:

$$\frac{C_m - C_p}{\bar{C} - C_p} = \exp\left(\frac{J_w}{K_m}\right) \tag{8}$$

where $J_w = \mathcal{Q}_p/A_{em}$ and K_m is mass transfer coefficient and is given by the following equations [47]:

$$K_m = 0.065\left(\frac{D_b}{d_h}\right)(N_{Re}^{0.875})(N_{Sc}^{0.25}) \tag{9}$$

where $\rho_b \approx 10^3(kg/m^3)$

$$N_{Re} = \frac{\rho_b d_h \mathcal{Q}_b}{d_f W \eta_b},\ N_{Sc} = \frac{\eta_b}{\rho_b D_b}$$
$$D_b = 6.725 \times 10^{-6}\exp\left(0.1546 \times C_b - \frac{2513}{T_f}\right)$$
$$\eta_b = 1.234 \times 10^{-6}\left(0.0212 \times C_b + \frac{1965}{T_f}\right).$$

Note that 1 Kg (water mass) per second is 3.6 cubic meters per hour. Therefore, we use $F_i[m^3/h] = 3.6\mathcal{Q}_i[kg/s]$ for $i \in \{f, p\}$ to show the feed and permeate flow rate based on the cubic meters per hour. In the long run, membrane decay and fouling are unavoidable. To determine long-term RO plant performance accurately, fouling effects should be considered. Fouling in membrane systems can be evaluated using mathematical predictive models [48–50]. These models can estimate the fouling effect by calculating the permeate flux decline over time due to long-term variation of the water permeability coefficient. Although, the membrane module parameter can be assumed to remain constant when considering the operation and optimization of the RO process over a short period. However, the fouling effect can be used to check the robustness of the long-term performance of the controller. The following mathematical model is considered for the fouling [50]:

$$A_w = A_{w0} e^{(-\frac{t_1}{\tau_{w1}})},$$
$$B_s = B_{s0} e^{(\frac{t_1}{\tau_{w2}})}. \qquad (10)$$

where τ_{w_1} and τ_{w_2} denote membrane performance decay constants and t_1 and t_2 are the time spent since the last cleaning and the last replacement, respectively. A_{w0} and B_{s0} are the initial membrane coefficients.

The energy consumption in the RO system in Kwh can be computed by the following equation:

$$E_c = \frac{0.036 Q_f P_f}{\xi_{\text{HP}}} - 0.036 Q_b P_b \xi_{\text{E}} \qquad (11)$$

where ξ_{E} is efficiency of ERD, and $\xi_{\text{HP}} = \xi_{\text{M}} \xi_{\text{P}}$ with ξ_{P} shows the efficiency of pump and ξ_{M} denotes the efficiency of motor.

3.2 Demand of daily freshwater

A daily demand for freshwater needs to be calculated so that an optimal RO system operation can be conducted based on the information provided. In this study, the information about the demand for daily freshwater in studies [51, 52] is utilized to create the following equation for water demand $F_d[m^3/h]$:

$$F_d(t) = \begin{cases} 41.60 - 2.54t - 3.22t^2 + 0.86t^3 & t < 7.5 \\ 829.9 - 150.43t + 9.87t^2 - 0.203t^3 & t > 7.5 \\ \& \ t < 23 & \end{cases} \qquad (12)$$

It’s important to highlight that the demand curve can be adjusted based on the size of the RO plant and the freshwater requirements, allowing for scaling up or down. Figure 3 illustrates the assumed daily demand for freshwater.

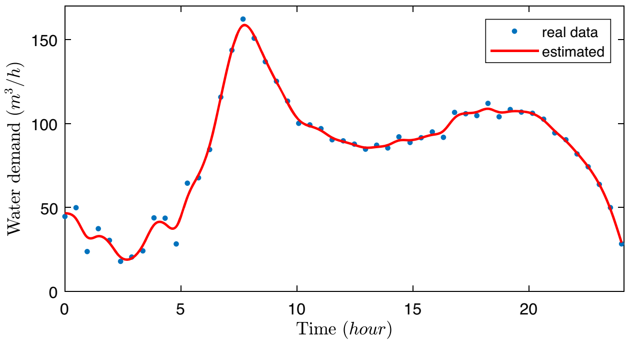


Fig. 3 The data for the demand of water

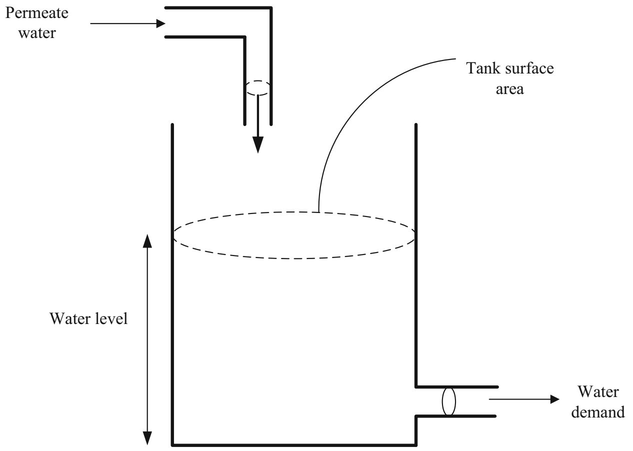


Fig. 4 Storage tank system schematic

3.3 Storage tank system

The mathematical model for a storage tank system whose schematic is shown in Fig. 4 can be expressed in terms of water level and the output flow rate determined by the demand and the concentration of output water.

$$\frac{dH_{st}}{dt} = \frac{F_p - F_d}{A_{st}},$$
$$\frac{dC_{st}}{dt} = \frac{F_p \left(C_p - C_{st}\right)}{A_{st} H_{st}}, \qquad (13)$$

Here, $A_{st}[m^2]$ and $H_{st}[m]$ represent the area and water level of the storage tank system, respectively. $F_p[m^3/h]$ and $C_p[kg/m^3]$ are the permeate water flow rate and salt concentration of the RO process. C_{st} is the concentration of outlet water of the storage tank system, $F_d[m^3/h]$ is the flow rate of output fresh water to guarantee the request of freshwater user demand which is scheduled one day earlier and can be obtained from a field data regression. To ensure the system operational safety, the reservoir level should be $H_{st}^{(min)} \leq H_{st} \leq H_{st}^{(max)}$.

4 Optimal operation of desalination plant

In this section, two RL agents are designed for control and optimal management of the desalination process with mathematical models described in Section 3. First, a DDPG agent is designed to regulate permeate water flow rate based on a given setpoint in the RO desalination process by manipulating a high-pressure pump. Then, a DQN is trained and employed to determine the setpoint for the controller by considering the demand for fresh water and optimizing the energy usage for producing freshwater.

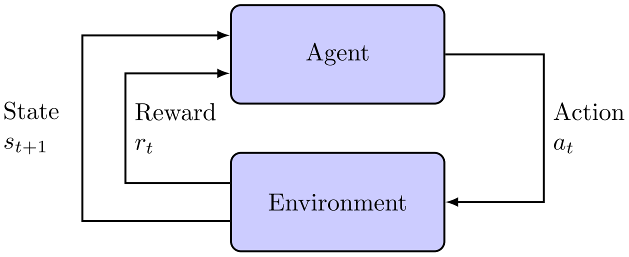


Fig. 5 The interaction between environment and agent [38]

The following notations and definitions have been utilized during the design of two RL agents, as illustrated in Fig. 5. The state space is shown with $\mathcal{S}$ and state s is $s \in \mathcal{S}$. The action space is determined by $\mathcal{A}$ and the action a is $a \in \mathcal{A}$. The Q-function $Q(s, a)$ determines the action-value function for the action a and the state s. The actor function is shown by $\pi(s)$ for $s \in \mathcal{S}$, which deterministically maps states to a specific action. A reward function provides feedback to the agent about what is correct and wrong with rewards and penalties and is shown with $r(.)$. Moreover, usually, several episodes are needed to train the RL agents. An episode is a group of states, actions, and rewards culminating in a terminal state.

4.1 Design the RL-DDPG controller

Here, the primary objective is to design a data-driven controller using the DDPG method to regulate the permeate water from the RO process by manipulating the high-pressure pump. DDPG is adept at handling a wide range of control problems and represents a model-free reinforcement learning strategy that integrates DQN and Deterministic Policy Gradients (DPGs), enabling an actor-critical RL agent to find the optimal policy and maximize expected cumulative long-term returns [53].

In this approach, the actor predicts an action based on the state input, while the critic predicts the value of the current state-action. As illustrated in Fig. 6, the state inputs are the reference tracking error and permeate water flow rate of the RO system, the action is the feed flow pressure, and the reward function is a function of the reference tracking error. To estimate the Q-function for the critic network, DQN employs a deep neural network, following an ϵ-greedy policy in a discrete action space. For an actor-network, DPG maps the state to a specific action deterministically. DPG achieves this by parameterizing the actor function and updating its parameters based on the policy's performance gradient [38].

Following the DDPG algorithm in [53], the subsequent steps can be taken to train the DDPG agent for one episode.

1. The parameterized critic function $Q(s, a \mid \phi_Q)$ with weights ϕ_Q, and actor function $\pi(s \mid \phi_\pi)$ with weights ϕ_π is randomly initialized
2. The target networks $Q'(s, a \mid \phi_{Q'})$ and $\pi'(s \mid \phi_{\pi'})$ with $\phi_{\pi'} \leftarrow \phi_\pi$ and $\phi_{Q'} \leftarrow \phi_Q$ are initialized
3. To explore the action space, a random process $\mathcal{N}$ is initialized and then an initial state s_1 is observed
4. For each training time step $t = 1 : T$, the following steps are taken
   (a) Based on the current state s_t, the action a_t is selected as $a_t = \pi(s_t \mid \phi_\pi) + \mathcal{N}_t$
   (b) Upon the execution action a_t, the reward r_t and next state s_{t+1} is received
   (c) The experience (s_t, a_t, r_t, s_{t+1}) is stored in experience buffer $\mathcal{R}$
   (d) A random minibatch with size N from experiences (s_i, a_i, r_i, s_{i+1}) are selected from experience buffer R
   (e) $y_i = r_i + \gamma_1 Q'\left(s_{i+1}, \pi'\left(s_{i+1} \mid \phi_{\pi'}\right) \mid \phi_{Q'}\right)$ is computed with t factor γ_1
   (f) The critic network weights is updated by minimizing the loss function $L_f = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - Q(s_i, a_i \mid \phi_Q)\right)^2$
   (g) Using a sampled policy gradient, the actor network is updated to maximize the discounted expected rewards $\nabla_{\phi_\pi} J \approx \frac{1}{N}\sum_{i=1}^{N} \mathcal{G}_{Q_i}\mathcal{G}_{\pi_i}$ with

$$\begin{aligned}\mathcal{G}_{Q_i} &= \nabla_a Q\left(s, a \mid \phi_Q\right) |_{s=s_i, a=\pi(s_i)}\\ \mathcal{G}_{\pi_i} &= \nabla_{\phi_\pi} \pi\left(s \mid \phi_\pi\right) |_{s=s_i}\end{aligned}$$

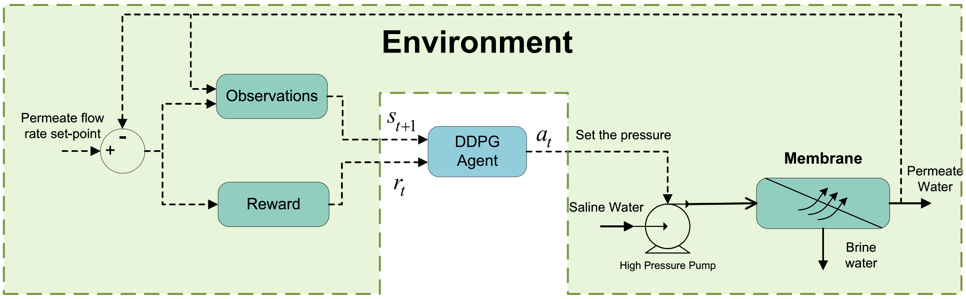


Fig. 6 Interaction between environment including RO system and agent as a controller

(h) Now the parameters target networks with smoothing factor α are update as follows:

$$\phi_{Q'} \leftarrow \alpha\phi_Q + (1-\alpha)\phi_{Q'}$$
$$\phi_{\pi'} \leftarrow \alpha\phi_\pi + (1-\alpha)\phi_{\pi'}$$

A DDPG agent will undergo training based on the outlined steps, with its primary goal being to control the permeate flow rate by manipulating a high-pressure pump. In this context, it's crucial to define the observations and actions within the environment. Figure 6 illustrates how the agent sends actions to the environment and observes the next state and rewards from the environment in the structure of agent-environment interaction depicted in Fig. 5. At time t, the DDPG agent takes the action a_t to set the value for feed pressure (P_f) in (5) as follows:

$$P_f \leftarrow a_t \tag{14}$$

Then, it receives the reward r_t as well as the observation s_{t+1} as follows:

$$s_{t+1} = \left[\int e d\tau, e, F_p\right]^\top, \tag{15}$$

where $F_\mathrm{p}[m^3/h]$ is permeate flow rate obtained by (4) and the error is defined $e = F_\mathrm{ref} - F_\mathrm{p}$ with F_ref as a reference value for the permeate flow rate of the RO process.

Remark 1 The DDPG agent training uses a reset function to randomize the reference signal for the controller at the beginning of each episode. Therefore, the agent is trained to follow the setpoint $\tilde{F}_\mathrm{ref}$ where $\tilde{F}_\mathrm{ref}$ is drawn from a uniform distribution interval $(F_\mathrm{ref}^\mathrm{min}, F_\mathrm{ref}^\mathrm{max})$. So, it assures that the agent can track a range of values for the reference signal.

4.1.1 The reward function design for DDPG

The reward function for training the DDPG agent has a remarkable impact on the controller performance and is defined as follows:

$$r_t = \begin{cases} \beta_1 & |e| < \eta_1 \\ -\beta_2 & |e| \geq \eta_1 \end{cases} \tag{16}$$

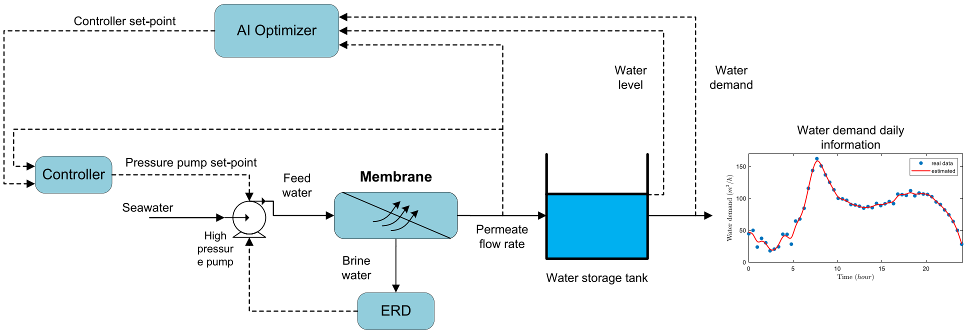
where β_1 and β_2 are the positive values that determine the reward (if $|e| < \eta_1$) or penalty (if $|e| \geq \eta_1$) that agent receives during the training phase. All the values β_1, β_2 and η_1 are specified before the training of the agent. This assure that the agent during the training try to keep the error in boundary region $|e| < \eta_1$ to maximize the cumulative rewards. It is important to note that η_1 determines the threshold for the absolute error $|e|$. If the absolute error is less than η_1, the agent receives a reward of β_1; otherwise, it incurs a penalty of $-\beta_2$. In this work, the main goal of agent is to track the reference setpoints between values of 12 and 30 as it will be shown in the simulation results. An absolute error of 0.2 corresponds to a percentage error between 0.67% and 2%, which is acceptable for our purposes.

4.2 Design optimizer agent based RL-DQN algorithm

The primary goal is to design a DQN agent tasked with determining the required amount of permeate water to be produced by the RO plant based on specified demands and a cost function related to energy efficiency, as shown in Fig. 7. Essentially, the DQN sends the reference value for the permeate flow rate to the controller. The controller's primary function is to manipulate the feed pressure to produce the permeate water as requested by the reference value. As depicted in Fig. 8, the input states for the DQN agent are derived from the water demand, the water level in the storage tank system, and the permeate flow rate of the RO system. The action of the DQN agent serves as the setpoint for the DDPG controller. The reward function, explained subsequently, is a function of energy consumption, water quality, and preventing overflow and underflow in the storage tank system.

The DQN algorithm, a model-free, off-policy RL methodology in discrete action space, is a variant of Q-learning. Unlike standard Q-learning, which is inefficient for large Markov Decision Processes (MDPs) with numerous states and actions, DQN can handle high-dimensional observation

Fig. 7 Schematic of RO desalination system with an optimizer and controller for daily management of water demand



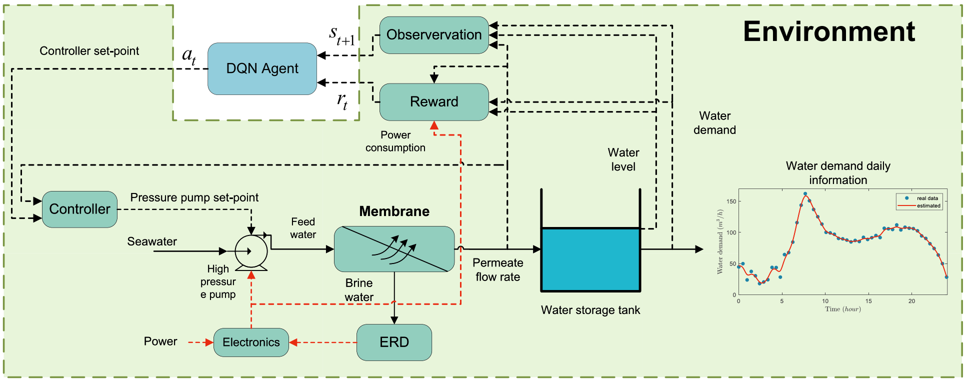


Fig. 8 Interaction between environment and DQN agent

and action spaces by using a neural network to estimate the Q-value function [54]. Based on the DQN algorithm in [54, 55], the following steps are employed for one episode during the training of a DQN agent.

1. The action-value function $Q(s, a \mid \phi_Q)$ with random weights ϕ_Q is initialized
2. The target action-value function $Q'(s, a \mid \phi_{Q'})$ is initialized with $\phi_{Q'} \leftarrow \phi_Q$
3. To explore the action space, a random process $\mathcal{N}$ is initialized and then an initial state s_1 is observed
4. For each training time step $t = 1 : T$, the following steps are taken

   (a) An action is selected based on the following rule:

$$a_t = \begin{cases} \text{random action} & \text{probability of } \epsilon \\ \arg\max_a Q(s_t, a \mid \phi_Q) & \text{probability of } 1 - \epsilon \end{cases} \tag{17}$$

   (b) Upon implementing the action a_t, the reward r_t and next state s_{t+1} are received

   (c) The experience (s_t, a_t, r_t, s_{t+1}) is stored in experience buffer $\mathcal{R}$

   (d) A random minibatch with size N from experiences (s_i, a_i, r_i, s_{i+1}) are selected from experience buffer $\mathcal{R}$

   (e) The value function target y_i is computed with a discount factor γ_2 by following equation:

$$y_i = \begin{cases} r_i \text{ if } s_{i+1} \text{ is terminal state} \\ r_i' \text{ otherwise} \end{cases} \tag{18}$$

   with $r_i' = r_i + \gamma_2 \max_{a'} Q'\left(s_{i+1}, a' \mid \phi_{Q'}\right)$

   (f) The action-value function parameters ϕ_Q is updated by minimizing the following loss function

$$L_f = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - Q(s_i, a_i \mid \phi_Q)\right)^2$$

   (g) Now the parameters of the target action-value function with smoothing factor α are update as follows:

$$\phi_{Q'} \leftarrow \alpha \phi_Q + (1 - \alpha) \phi_{Q'}$$

Interaction between environment and DQN agent has been shown in Fig. 8.

A DQN agent will undergo training following the outlined steps, with its primary goal being to provide setpoints for the controller. To accomplish this, it is crucial to identify observations and actions within the environment. Figure 7 shows how the agent sends actions to the environment and observes the rewards from the environment. The DQN agent in time t takes the action a_t to set the reference value (F_{ref}) for the DQN agent as follows:

$$F_{\text{ref}} \leftarrow a_t \tag{19}$$

Then, it receives the reward r_t as well as the observation s_{t+1} as follows:

$$s_{t+1} = \left[H_{st}, F_d, \Delta F_d, F_p, \Delta F_p\right]^{\top}, \tag{20}$$

where $H_{st}[m]$ is tank water level, $F_d[m^3/h]$ is demand of water and $F_{\text{p}}[m^3/h]$ is permeate flow rate obtained by (4) and ΔF_d and ΔF_p are $\Delta F_d = F_d(t) - F_d(t-1)$ and $\Delta F_p = F_p(t) - F_p(t-1)$ respectively.

Remark 2 The initial water level value for the episode is an essential factor. DQN issues different setpoints for the DDPG controller based on the initial water level value. A reset function has been used to accommodate the randomized initial value of water level in the storage tank system, with the value of $\tilde{H}_0$ drawn from a uniform distribution interval ($H_0^{\min}$, $H_0^{\max}$), thereby ensuring that the DQN agent can manage the desalinization process with different values of initial water level in the tank system.

Remark 3 The water demand data provided in (12) is constant for a specific time t_s. To consider a more realistic scenario and to make the smart optimizer agent able to handle the variation in water demand in each time instant, at the beginning of each episode, a reset function sets the demand of water in each time instant t_s as $\tilde{F}_d(t_s)$ with the following equation:

$$\tilde{F}_d(t_s) = F_d(t_s) + \tilde{f}(t_s), \tag{21}$$

where $\tilde{f}(t_s)$ is drawn from a normal distribution $\mathcal{N}(0, \alpha F_d(t_s))$ for $t_s \in (0, 24)$. Here, α is a parameter that determines the standard deviation of the normal distribution from which $\tilde{f}(t_s)$ is drawn. The larger the value of α, the wider the spread of the distribution, indicating higher uncertainty or variability in water demand.

4.2.1 Reward function design for the DQN agent

In the training phase, the reward function acts as an essential mechanism, influencing the agent's behavior for optimal performance in RO operations. Training the DQN agent with this reward function aims to derive an optimal policy that selects setpoints for the DDPG controller, minimizing the total operation cost of freshwater production and satisfying specified constraints. The primary cost consideration involves the energy consumption of the high-pressure pump, with some ERDs mitigating this consumption. Operational constraints include maintaining the desired permeate concentration, avoiding tank overflow or underflow, and ensuring the water quality aligns with defined standards.

The main objective of the RO plant is to supply freshwater with a suitable permeate concentration, stored in a tank to meet varying user demand. Selection of setpoint values must carefully consider the permeate water flow rate to prevent tank overflow or underflow. Additionally, water quality, measured in terms of concentration, becomes a decision variable for optimal setpoint selection. Therefore, optimizing the daily operation of the RO plant involves training the DQN agent to minimize operation costs or maximize profit in delivering freshwater while upholding water quality, meeting demand, and preventing tank overflow or underflow.

To achieve this, the reward function comprises three integral parts, as follows:

$$\begin{aligned} r_1(t) &= \bar{F}_d \times \mathcal{P}_{\text{fw}}\left(\bar{C}_{st}\right) \times \frac{T_s}{h_s}, \\ r_2(t) &= -T_s \times \int_0^{T_s} \frac{E_c \mathcal{P}_{\text{ec}}}{h_s} dt, \\ r_3(t) &= -\bar{F}_p \times (T_{\text{final}} - t) \times \frac{\upsilon_{eb}}{h_s}. \end{aligned} \tag{22}$$

where T_s is the sample time for the DQN agent, $\mathcal{P}_{\text{fw}}(.)$ is the price based on the quality of the water (concentration of freshwater from the tank system $\bar{C}_{st}$), $\mathcal{P}_{\text{ec}}$ is the price for 1 Kwh, h_s is the time unit for one hour for example if unit time is second $h_s = 3600$, $E_c[kwh]$ is the energy consumption by the RO system. υ_{eb} is 1 when the level of water exceeds the predefined bounds (overflow or underflow) and 0 otherwise. Suppose there is an overflow or underflow in the operation of the RO plant. In that case, the training of the agent is stopped, and there is a penalty proportional to $(T_{\text{final}} - t)$, which is the remaining time to complete the daily operation of the RO plant. $\bar{F}_d$, $\bar{C}_{\text{st}}$ and $\bar{F}_p$ are the moving average with time windows length T_s of water demand, output concentration of tank system, and permeate water flow rate, respectively. The total reward function is defined as follows:

$$r_t = \lambda_1 r_1(t) + \lambda_2 r_2(t) + \lambda_3 r_3(t). \tag{23}$$

where the λ_1, λ_2 and λ_3 are the scaling weights to determine the importance $r_1(t)$, $r_2(t)$ and $r_3(t)$. The DQN agent, during the training, learn the optimal policy to transmit the optimal setpoint values for the RO plant by maximizing the reward

Table 1 Membrane specification of element-FilmTec SW30HR-380 and feedwater parameter values [25]

Feed parameters	Value	Unit
feed flow rate Q_f	50	kg/s
feed temperature T_f	303	K
feed concentration C_f	42	kg/m^3
RO Parameters	Value	Unit
A_m	35.3	m^2
n_v	6	
n_e	56	
A_w	2.05×10^{-6}	$m/(s, kPa)$
B_s	2.03×10^{-5}	m/s
$a_{rec}^{\max}$	10	%
$a_{rec}^{\min}$	40	%
a_T	9	
b_T	8.08	
R	50	
ρ_w	1000	kg/m^3
b_f	7986	$(kPa.s)/m^4$
d_f	7×10^{-3}	m
d_h	1	m
Tank Parameters	Value	Unit
A_{st}	100	m^2
H_T	6	m

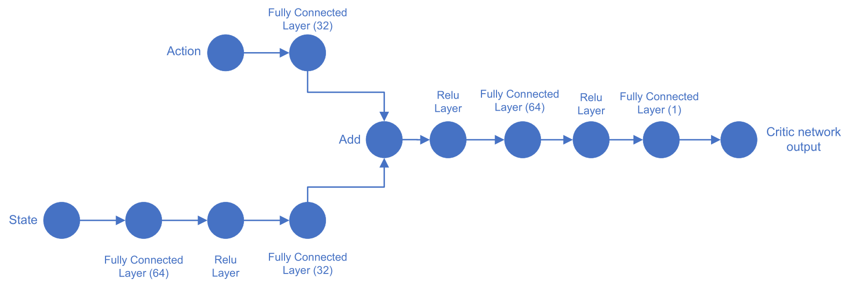

Fully Connected Layer (32)
Action
Relu Layer
Fully Connected Layer (64)
Relu Layer
Fully Connected Layer (1)
Add
Critic network output
State
Fully Connected Layer (64)
Relu Layer
Fully Connected Layer (32)

Fig. 9 The topology of critic network for the DDPG controller

function in (23). Therefore, maximizing the reward function means that the optimal policy by the agent is implemented to keep the quality of freshwater from the tanks system at an appropriate value (with a reward term of $r_1(t)$), reduce the cost of permeate water or in other words, reduce the energy consumption by the high-pressure pump (with penalty term of $r_2(t)$) and satisfy the water demand and avoid the underflow and overflow of permeate water in the storage tank system (with penalty term of $r_3(t)$).

5 Simulation and discussion

DDPG and DQN agents are trained to provide daily operational support for an RO desalination plant with membrane specifications outlined in Table 1 and incorporating a storage tank system based on water demand data. In the initial step, a DDPG controller is trained for the RO process to regulate the permeate water flow rate based on a reference value issued by the higher-level optimizer, namely the DQN agent. Subsequently, the DQN agent is designed using information related to energy consumption cost, storage tank water level, and freshwater price. The DQN agent communicates the setpoint value as an action to the controller, determining the required water production for the desalination process system. The schematic of the discussed RO desalination process with the controller and optimizer is illustrated in Fig. 7.

5.1 RL-DDPG controller training

In the initial stage, a data-driven controller, utilizing the DDPG method discussed in Section 4.1, is developed to regulate the permeate water flow rate. The observations comprise the integrator error, the error between the reference value and the output value of the permeate flow rate, and the actual permeate flow rate of the RO process. The action involves setting the feed pressure for the high-pressure pump.

The reward function in (16) is calculated by taking the values $\eta_1 = 0.2$, $\beta_1 = 10$ and $\beta_2 = 1$. By maximizing the reward function during training, the agent learns to map the observation in (15) to the action in (14) to keep the magnitude of the error less than the value of η_1. The critic and actor-network typologies for the DDPG agent are illustrated in Figs. 9 and 10.

The discount factor and sampling time for the DDPG agent are set as $\gamma_1 = 1$ and $T_s = 4$ seconds, respectively, with a learning rate of 0.001 for neural network training. A discount factor $\gamma_1 = 1$ implies that the RL controller treats immediate and future rewards equally. This is advantageous in tracking problems where maintaining consistent performance over time is crucial. A discount factor of 1 helps balance short-term accuracy with long-term stability, ensuring the DDPG agent considers the entire future trajectory and makes informed decisions for the reference tracking problem.

During agent training, the setpoint for the permeate flow rate of the RO plant is randomly selected from a uniform distribution interval of (40, 108) in terms of m^3/h to ensure that the agent can effectively manage the pressure, allowing the permeate flow rate to follow the setpoints for regulation. The training process spans 15000 episodes, each consisting of 25 steps, with a total duration of 100 seconds. If the tracking error magnitude is within the specified bound, the agent receives a reward of 10. Consequently, the maximum reward with a discount factor of 1 is 250. The average reward of 200 in Fig. 11 indicates that, on average, the controller maintains the error between the reference signal value and the output permeate flow rate within the requested bound for 20 out of the total 25 steps. The variance in episode rewards in Fig. 11 arises from tracking the reference point within the range of 40 to 110 m^3/h for permeate flow rate. The DDPG RL agent is trained to follow the reference point within this range, leading to varying cumulative reference tracking errors and, consequently, different episode rewards. Figure 12 illustrates how the permeate flow rate tracks various reference setpoints by employing the DDPG agent.

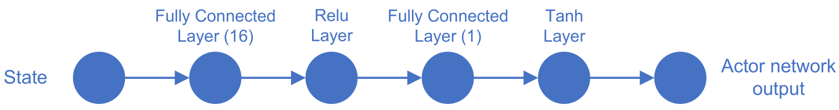


Fig. 10 The topology of actor network for the DDPG controller

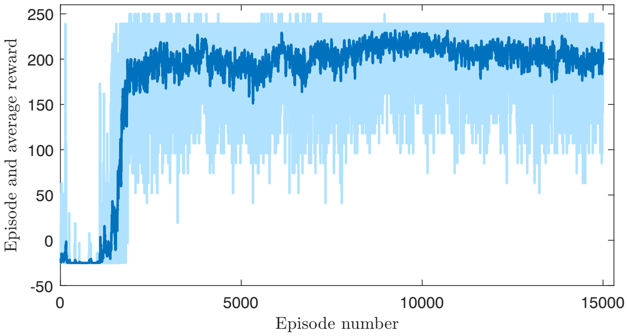


Fig. 11 Average of episode reward

Now the RO model is run for 100 days to check the performance of the DDPG controller in control of permeate flow rate by considering the fouling effect shown in (10). Because the permeate flow rate declines over time due to a long-term decrease in the water permeability coefficient, the controller increases the pressure to keep the permeate flow rate around the requested reference value. As it is shown in Fig. 13, the RL controller increases the feed pressure by almost 10% to keep the permeate flow rate around 60 $[m^3/h]$.

To compare the performance of the presented DDPG controller with a PID controller, a PID controller is fine-tuned using MATLAB with frequency-based approaches, targeting a setpoint of 50 (m^3/h). A PID controller with following structure is considered:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d \frac{de(t)}{dt} \tag{24}$$

where the error e is defined in (15). The gains of the PID controller are obtained as $K_p = 2.23$, $K_i = 40.12$, and $K_d = 1.58$. Subsequently, the cumulative reference tracking error of the PID controller is compared with that of the DDPG controller for a range of setpoints between $43 - 108$ (m^3/h). The results are presented in Fig. 14. The PID controller demonstrates adequate reference tracking between $40 - 65$ (m^3/h), aligning with its tuning range. However, for higher setpoints ranging from $65 - 110$ (m^3/h), the DDPG controller consistently outperforms PID control. This underscores the advantage of DDPG in learning complex nonlinear policies compared to traditional linear controllers like PID. This distinction arises because PID controllers, effective in systems with linear dynamics, may struggle to maintain optimal performance in highly nonlinear environments such as the RO system. PID controllers rely on linearized models and tuning procedures, which do not generalize well to nonlinear systems. In contrast, DDPG controllers can continuously learn and refine their control policy throughout training by interacting with the true nonlinear system dynamics. However, it is important to note that It is important to note that, DDPG requires more computational resources than PID controllers due to the use of neural networks and the need for training data. PID controllers are computationally simpler and can be implemented more easily in real-time applications.

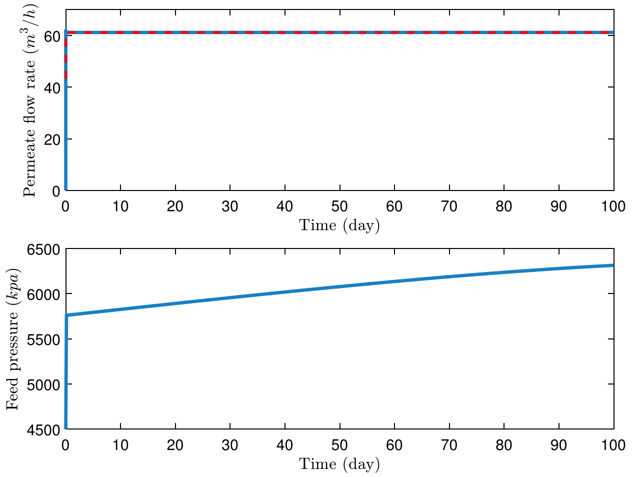


Fig. 13 The performance of controller with the fouling

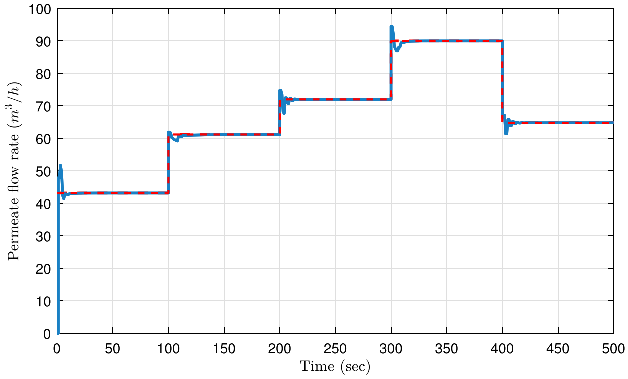


Fig. 12 Permeate flow rate tracking of multi step setpoints

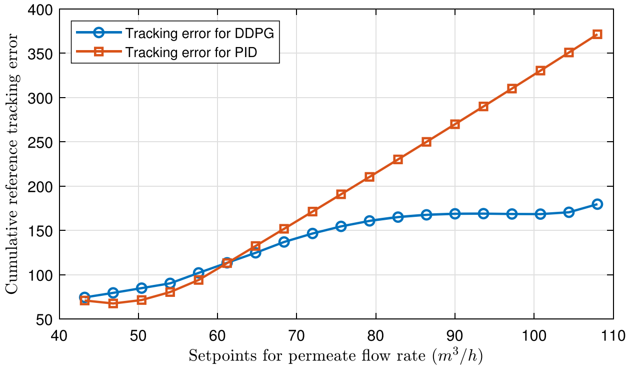


Fig. 14 Comparison of cumulative reference tracking error for PID and DDPG controllers

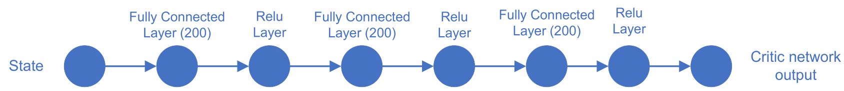

State
Fully Connected Layer (200)
Relu Layer
Fully Connected Layer (200)
Relu Layer
Fully Connected Layer (200)
Relu Layer
Critic network output

Fig. 15 The topology of critic network for the DQN agent

5.2 RL-DQN optimizer training

In the second step, an RL agent, utilizing the DQN approach as discussed in Section 4.2, is devised for the optimal operation and management of the RO desalination plant. The observations are derived from the tank water level, water demand flow rate, and permeate flow rate, as depicted in (20). The action involves setting the setpoints for the permeate flow rate, which are then transmitted to the DDPG controller. The reward function in (23) relies on price data for permeate water, contingent on the concentration of permeate water and the electricity price.

To incorporate the quality of stored freshwater in the tank into the reward function, we assume that freshwater is delivered from the storage tank system to the end-user at a price proportional to the concentration of the permeate water. Specifically, lower concentrations are delivered to the end-user at a higher price. Assuming prices ρ_2 and ρ_1 for one cubic meter of freshwater with concentrations C_{p_1} and C_{p_2}, the following equation determines the value of freshwater in terms of water for a given concentration C_{p_x}:

$$\mathcal{P}_{\mathrm{fw}}(C_{px}) = \frac{\rho_2 - \rho_1}{C_{p_2} - C_{p_1}} \times \left(C_{px} - C_{p_1}\right) + \rho_1 \tag{25}$$

Therefore based on the operational cost and maintenance cost of a RO desalination plant [56], $\mathcal{P}_{\mathrm{fw}}(C_{px})$ is obtained as follows:

$$\mathcal{P}_{\mathrm{fw}}(C_{px}) = -10.4651\left(C_{px} - 0.17\right) + 5 \tag{26}$$

The price of electricity is assumed as $\mathcal{P}_{\mathrm{ec}} = 0.08\$$ per kWh [14].

By maximizing the reward function during training, the agent learns to map the observation in (20) to the action in (19) to satisfy the designed requirements for optimal managing of the RO plant. The critic network topology of the DQN agent is illustrated in Fig. 15.

The discount factor and sampling time for the DQN agent are selected as $\gamma_1 = 0.9$ and $T_s = 15$ minutes with the learning rate for the neural network training as 0.0001. In this section, three scenarios are explored for the training of the DQN agent. In the first scenario, a constant initial value is assumed for the storage tank water level. Subsequently, a DQN agent is trained to regulate permeate water production, taking into account a randomized initial value for the tank level. Finally, the DQN agent undergoes training while considering stochastic water demand in addition to the randomized initial water level.

5.2.1 DQN agent training with a constant level tank initial value

In this section, the DQN agent is trained to operate the desalination RO process in real-time with a constant initial value for the water level in the tank storage system. Therefore, the agent for a predefined initial value for the water level in the storage tank and demand water data tries to maximize the reward function defined in (23). The agent is trained for 1000 episodes, and the reward for each episode and average reward with a time window length of 20 is shown in Fig. 16. At the end of the training, the average reward is about 3454.9. The negative and small episode reward values in Fig. 16 show the early stages of training in which the agent fails to complete the water management to fulfill the water demand, so an overflow or underflow happens. By training the agent, it learns to select the optimal setpoint by maximizing the reward function in (23) for the optimal usage of the energy consumption by the high-pressure pump, satisfy the water demand, and maintain the freshwater quality in the tank system.

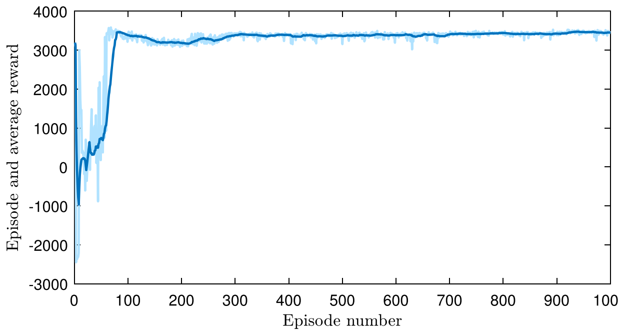


Fig. 16 The reward function for the DQN agent

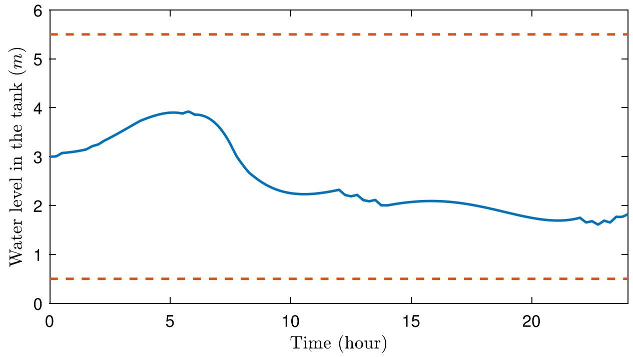


Fig. 17 Water level variation in the storage tank system

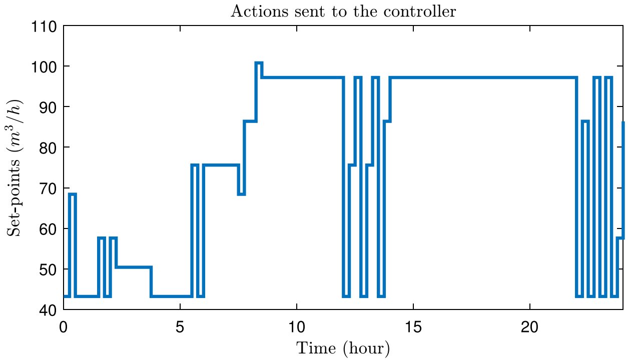


Fig. 18 Determining setpoints for the controller

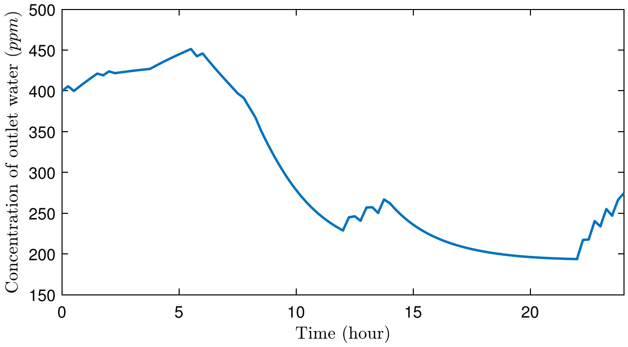


Fig. 19 Concentration of outlet water of storage tank system

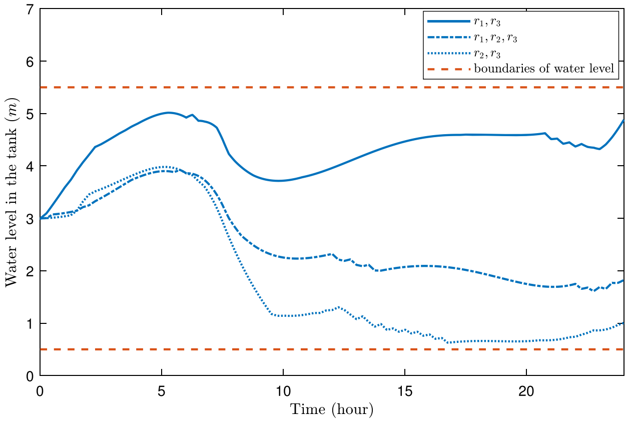


Fig. 20 The water level variation in storage tank system

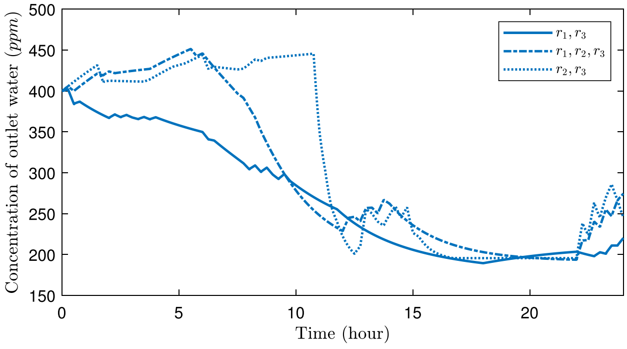


Fig. 21 The outlet water concentration of the storage tank system

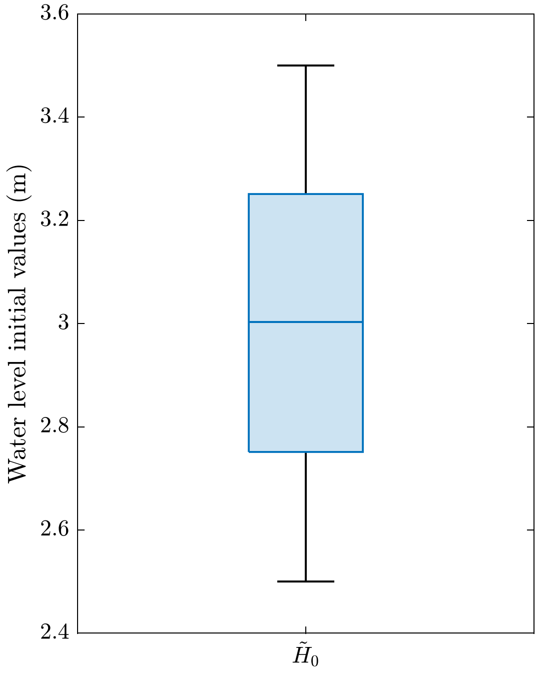


Fig. 22 Distribution of water level initial value

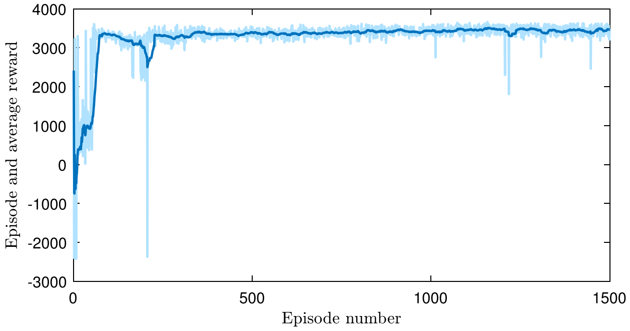


Fig. 23 The reward function for the DQN agent

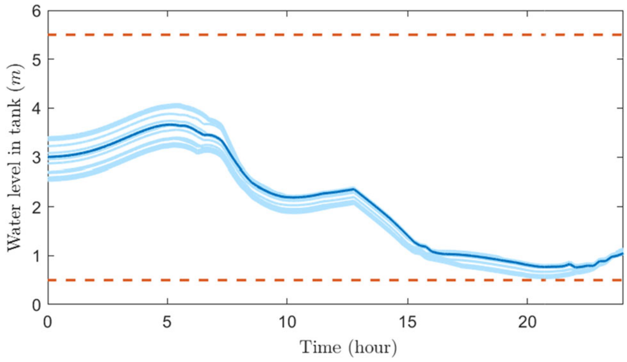


Fig. 24 Water level variation in the storage tank system for different value of water level initial conditions

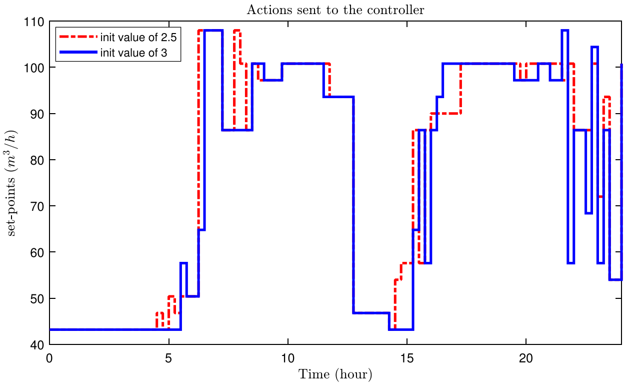


Fig. 25 Determining setpoints for the controller for different water level initial values

The variation in water level in the storage tank system is shown in Fig. 17. As is demonstrated in this figure, the DQN agent manages and stores the permeate water produced by the RO plant in the tank system to meet the required water demand and avoid the overflow and underflow specified with the red dashed line in the figure.

The sent optimal setpoints are shown in Fig. 18. As it is shown in this figure, the duration of the times with a high demand for freshwater shown in Fig. 3), the sent setpoints values are higher than other times.

The concentration of outlet permeate water from the storage tank system has been shown in Fig. 19 where the quality of outlet water keeps improving by just adjusting the pressure of the high-pressure pump. The concentration's initial value is 400[ppm]. Quality of outlet water in terms of concentration is considered in the reward function defined in (23) in terms of $r_1(t)$. By maximizing the reward function, the DQN agent learns the optimal policy to improve water quality by reducing the concentration of outlet water. It is imperative to note that the concentration of permeate water depends on the feed pressure, which means that with increasing the feed pressure, the permeate concentration is decreased.

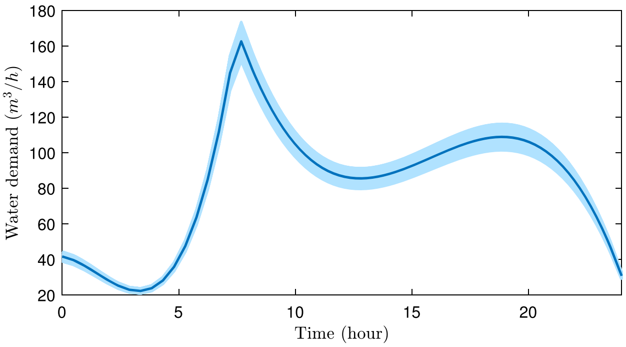


Fig. 26 The stochastic water demand data

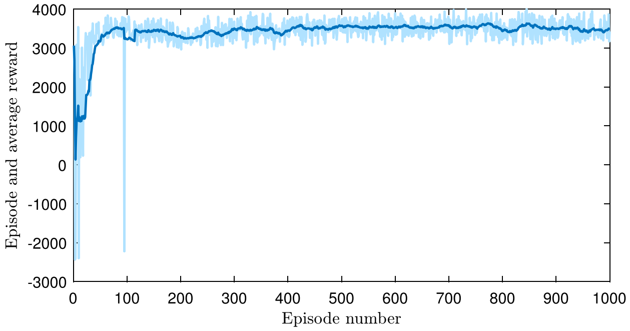


Fig. 27 The reward function for the DQN agent

As mentioned, increasing the feed pressure not only enhances water quality but also boosts the permeate flow rate, albeit at the cost of higher energy usage. The agent must carefully consider the trade-off between improving the quality of the outlet water in the storage tank system and the energy consumption of the high-pressure pump. To address this, the DQN agent undergoes training for two scenarios.

In the first scenario, the reward for water quality in (23) is replaced with a constant term, meaning only the terms $r_2(t)$ and $r_3(t)$ are considered in (23). In this case, the agent aims to minimize energy consumption by maintaining the water level in the tank near the defined lower threshold, resulting in higher permeate water concentration. In the second scenario, energy consumption is not factored into the reward function, meaning the terms $r_1(t)$ and $r_3(t)$ are considered in (23). In this case, the DQN agent learns an optimal policy that maintains a higher feed flow pressure compared to the previous scenario to improve water quality and keep the tank consistently full. Figures 20 and 21 depict the quality of permeate water and the water level in the storage tank system, illustrating the impact of considering energy consumption and permeate water quality.

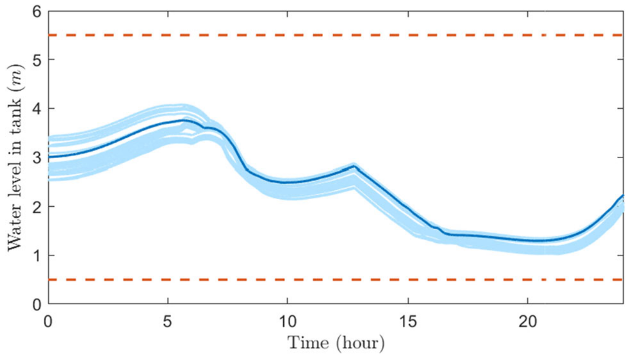


Fig. 28 Water level variation in the storage tank system by considering randomness in water level initial value and stochastic water demand

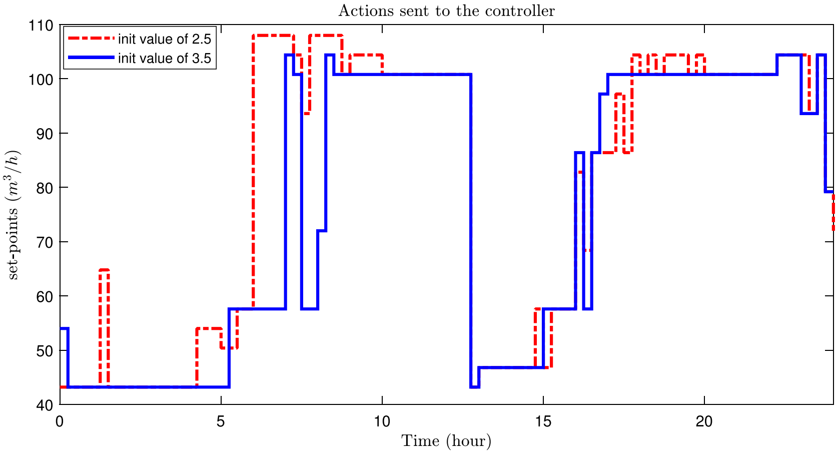


Fig. 29 Setpoints issued by the RL agent to manage water demand

5.2.2 DQN agent training with randomized tank level initial value

In this section, the training process for the DQN agent incorporates the variability in the initial value of the tank level. The fluctuation in the initial tank water level significantly influences the training of the DQN agent. For instance, when dealing with a high initial value, the agent needs to reduce the production of permeate water from the RO plant to prevent tank overflow. To enhance the agent's ability to adapt to the randomness in the initial value of the water level in the storage tank system, the initial value is randomly drawn from a uniform distribution of (2.5, 3.5) in each training episode, as depicted in Fig. 22.

The DQN agent is trained for 1000 episodes, and the rewards for each episode, as well as the average episode reward, are displayed in Fig. 23. At the conclusion of the training, the average reward is approximately 3452.62. The average reward for this scenario is nearly identical to the average reward for the scenario with a constant initial water level. However, there is an increased variation around the average reward. Training the agent with a randomized initial value for the water level in the storage tank system equips it to handle real-time randomness in the initial water level for optimal management of the RO plant.

Figure 24 demonstrates the variation of water level during one day in the storage tank system for different values of initial water level in the tank. This figure illustrates that the DQN agent for different initial water levels optimizes the RO plant to produce enough permeate water to avoid overflow or underflow.

Figure 25 shows the setpoints sent to determine the production of permeate water in the RO plant for two random initial water levels in the tank system. During the period prior to the pick-demand for freshwater, the DQN agent increases the high-pressure pump, so the tank has enough permeate water to satisfy the end-user's demand. At other times, it reduces the pressure of feed water to avoid overflowing and the energy consumption the high-pressure pump uses.

5.2.3 DQN agent training with stochastic water demand and randomized water level initial value

Finally, it is assumed that the water demand data is not a deterministic time-varying function, in addition to the initial value of the tank water level. The water demand at each time

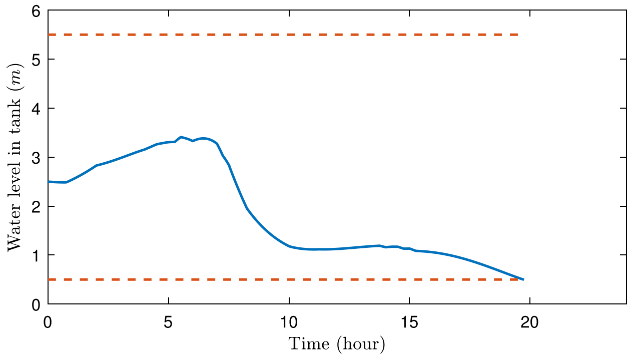


Fig. 30 Failing in managing the water by the DQN by increasing α in (21) for water water demand

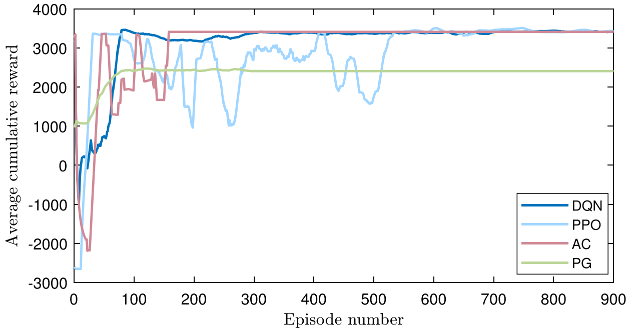


Fig. 31 Comparative Analysis of Cumulative Average Reward: DQN, PPO, AC, and PG Algorithms

Table 2 Hyper-parameter settings for various RL algorithms

Hyper-parameters	DQN	PPO	AC	PG
Critic network layer structure	[200,200,200]	[256,256]	[128,128]	[256,256]
Actor network layer structure	–	[80,64]	[96,64]	[80,64]
Learning rate for critic layer	0.001	0.0001	0.001	0.001
Learning rate for actor layer	–	0.001	0.001	0.001
Discount factor	0.9	0.9	0.9	0.9

instant is drawn from a normal distribution, as explained in (21) with $\alpha = 0.04$. Both variations in the initial tank water level and demand water are considered during the training of the DQN agent. The initial value for the water level is drawn from a uniform distribution (2.5, 3.5), as shown in Fig. 22. The water demand is provided to the DQN agent with a stochastic term specified in (21). Figure 26 illustrates the water demand, where the light blue area indicates the section where the water demand can change.

The DQN agent is trained for 1000 episodes, and the reward for each episode, along with the average episode reward, is depicted in Fig. 27. At the conclusion of the training, the average reward is approximately 3450.2. In comparison with previous scenarios, the agent demonstrates an ability to manage the randomness in tank water level and uncertainty in water demand in real-time, with the average reward showing minimal change. However, there is an increased variability around the average episode reward compared to previous scenarios.

Figure 28 illustrates the variation in water level within the storage tank system over the course of one day. As depicted in this figure, the DQN agent adeptly handles the storage of permeate water in the tank system, avoiding both overflow and underflow. Consequently, the DQN agent maintains the water level at an appropriate range, ensuring an ample supply of freshwater for delivery to end-users and, consequently, reducing energy consumption by the high-pressure pump.

Figure 29 shows the setpoints provided by the DQN agent for the DDPG controller. This figure illustrates how the DQN agent increases the high-pressure pump during the period before high water demand, so the tank has enough permeate water to meet demand. At other times, it reduces the pressure of feed water to avoid overflowing and the energy consumption the high-pressure pump uses.

Figure 30 explains that by selecting $\alpha = 0.08$, the DQN agent fails to succeed in managing the RO plant, and an underflow happens.

5.2.4 Comparative analysis of RL algorithms for optimal RO operation

This section compares DQN with several other prominent RL algorithms, namely Proximal Policy Optimization (PPO), Policy Gradient (PG), and Actor-Critic (AC), regarding their performance for the optimal operation of RO, using the parameters defined in Table 2. For brevity, we primarily compare the average cumulative reward for each algorithm, as shown in Fig. 31, while Table 3 presents a detailed discussion on max reward, min reward, and average reward values.

DQN is specifically designed for environments featuring high-dimensional state spaces, making it well-suited for complex tasks. Its success in handling large state spaces and demonstrating stability in learning has made it a popular choice across various applications. PPO, as a policy optimization algorithm, aims to maximize expected rewards while preventing large policy updates to maintain stability during training. Commonly used for continuous action spaces, PPO is recognized for its sample efficiency. PG methods directly parameterize the policy and optimize it through gradient ascent, proving to be model-free and suitable for both discrete and continuous action spaces. AC methods exhibit higher sample efficiency than DQN, particularly in continuous action spaces. Both DQN and AC methods typically demand fewer samples for training in continuous action spaces. DQN is renowned for its stability, especially in tasks with discrete action spaces. Based on the average cumulative reward shown in Fig. 31 and the results presented in Table 3, it is shown that the performance of DQN, PPO, and AC is

Table 3 Performance metrics of RL algorithms

Metric	DQN	PPO	AC	PG
Average Cumulative Reward	3452.38	3459.23	3454.71	2406.02
Max Reward	3624.85	3640.27	3590.13	2505.70
Min Reward	-2435.01	-2549.03	-2422.31	952.59

comparable, whereas the performance of PG is notably inferior compared to the other approaches. Furthermore, it should be noted that while DQN agents require tuning for the critic layer, PPO, AC, and PG agents also demand tuning for both the actor and critic layers to achieve optimal performance. Therefore, retraining of DQN is likely to take less time compared to other approaches.

6 Conclusion

This study aims to minimize the total daily operation cost of an RO desalination plant, meeting the variable daily freshwater demand through the implementation of an optimal real-time management method using RL techniques. Utilizing DDPG and DQN, a hierarchical structure with two RL agents was developed to optimize the RO plant, taking into account the dynamic model of the RO process. The primary role of the DDPG agent is to control the permeate flow rate by adjusting the high-pressure pump's pressure. Considering factors such as the water level in the storage tank system, permeate flow rate, and water demand, the DQN agent calculates the required amount of permeate water, aiming to maintain water quality in terms of permeate concentration.

The simulation results for the DDPG agent demonstrate its capability to control the permeate flow rate by manipulating the high-pressure pump within the complex RO system. However, it is noteworthy that training DDPG agents requires nearly 48 hours on a PC with an Intel Core i7-3770 and 8GB RAM. Additionally, the DDPG agent undergoes testing for the long-term operation of the RO plant to observe the impact of fouling. The DDPG controller adjusts the pressure for the extended operation of the RO system, maintaining the permeate flow rate at the required level to compensate for fouling effects. Comparing the performance of DDPG controllers with PID controllers in the nonlinear dynamics of a RO system showcases the superior adaptability of DDPG across a broad spectrum of reference setpoints.

Concerning DQN simulation results, three scenarios were examined: one with no initial water level randomness in the tank system, another with randomness in the storage tank system's initial value, and the third with stochastic water demand and initial water level randomness. With increased uncertainty in the environment and RO system parameters, the average episode reward for the DQN agent remains relatively consistent. However, heightened system uncertainty leads to greater variability in the episode reward for DQN agents. Also, comparing DQN with PPO and AC shows comparable performance, while PG performs notably worse; it's important to note that tuning is necessary for both actor and critic layers in PPO, AC, and PG, making DQN retraining likely less time-consuming.

The agent effectively oversees daily RO plant operations, optimizing energy use and improving delivered freshwater quality using a well-structured DQN critic network. It balances the trade-off between enhancing outlet water quality and reducing the high-pressure pump's energy consumption. While not prioritizing permeate water quality, the focus on energy efficiency includes maintaining the tank water level near the specified lower threshold.

The future research direction is to develop the RL agents to manage the energy consumption in an RO plant with solar panels and a battery storage system.

Appendix A: Description of most frequent symbols

Table 4 Summary of abbreviations and mathematical notation

Abbreviations	Description
RO	Reverse osmosis
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q-Network
DPG	Deterministic Policy Gradient
RL	Reinforcement Learning
ANN	Artificial Neural Network
PID	Proportional Integral Derivative
ML	Machine Learning
AI	Artificial Intelligence
ERD	Energy Recovery Device
MDP	Markov Decision Process
NN	Neural Network
Parameters	**Description**
M	Mass[Kg]
C	Concentration[Kg/m^3]
$\mathcal{Q}$	Mass Flow rate[Kg/s]
P	Pressure[kPa]
Π	Osmotic Pressure[kPa]
R	Reject valve rangeability
A_w	Water Permeability[$m/(s, kPa)$]
B_s	Salt Permeability[$m/(s)$]
T_{cp}	Temperature Coefficient Factor[K]
T_{cs}	Temperature Coefficient Factor[K]
T_{ref}	Reference temperature [K]
β	Concentration polarization factor
ΔP	Trans-Membrane Pressure[kPa]
$\Delta\Pi$	Net osmotic pressure [kPa]
a_{T}	Membrane water passage temperature constant
b_{T}	Membrane salt passage temperature constant

Table 4 continued

Abbreviations	Description
K_{m}	Mass transfer coefficient $[m/s]$
J_{w}	Water flux $[Kg/(m^2, s)]$
N_{Re}	Reynolds number
N_{Sc}	Schmidt number
ρ_{b}	Water density $[Kg/m^3]$
D_{b}	Diffusivity of the brine
F_{d}	Water demand $[m^3/h]$
H_{st}	Tank water level $[m]$
A_{st}	Tank cross sectional area $[m^2]$
C_{st}	Tank outlet water concentration $[Kg/m^3]$
$\bar{C}$	feedwater and brine concentration average $[Kg/m^3]$
d_{f}	is feed spacer thickness $[m]$
d_{h}	brine channel's spacer thickens
η_{b}	viscosity of the brine
n_{v}	Pressure vessel number
n_{e}	Elements number in a pressure vessel
A_{m}	Membrane active surface area $[m^2]$
τ_{w}	Membrane performance decay constants
E_{c}	Energy consumption $[Kwh]$
ξ_{P}	Pump efficiency
ξ_{M}	Motor efficiency
ξ_{E}	ERD efficiency
$\mathcal{P}_{\mathrm{ec}}$	Price of electricity $[\$/Kwh]$
$\mathcal{P}_{\mathrm{fw}}(.)$	Price of freshwater based on water quality $[\$/m^3]$
$\tilde{F}_{\mathrm{d}}$	Stochastic Water demand $[m^3/h]$
$\mathcal{N}(.,.)$	Normal distribution
$Q(.,.)$	Q function
$\pi(.)$	Actor function
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$r(.)$	Reward function
F_{ref}	Reference value for permeate water $[m^3/h]$
Subscripts	**Description**
b	Brine side
f	Feed side
m	Membrane side
p	Permeate side

Author Contributions All authors - Arash Golabi, Abdelkarim Erradi, Hazim Qiblawey, Ashraf tantawy, Ahmed Bensaid and Khaled Shaban contributed to this work in an appropriate way.

Funding Open Access funding provided by the Qatar National Library. This work was made possible by the grant #QUHI-CENG-21/22-2 from Qatar University. The statements made herein are solely the responsibility of the authors.

Data Availability Based on the mathematical models explained, all data used in this study has been generated by them.

Code Availability On agreement of all sides, the code will be made publicly available upon publication.

Declarations

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

Consent to participate Yes.

Consent for publication Yes.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Alatiqi I, Ettouney H, El-Dessouky H (1999) Process control in water desalination industry: An overview. Desalination 126(1–3):15–32. https://doi.org/10.1016/S0011-9164(99)00151-4
2. Majali F, Ettouney H, Abdel-Jabbar N, Qiblawey H (2008) Design and operating characteristics of pilot scale reverse osmosis plants. Desalination 222(1–3):441–450. https://doi.org/10.1016/j.desal.2007.01.169
3. Qiblawey H, Banat F, Al-Nasser Q (2011) Performance of reverse osmosis pilot plant powered by Photovoltaic in Jordan. Renew Energy 36(12):3452–3460. https://doi.org/10.1016/j.renene.2011.05.026
4. Toth AJ (2020) Modelling and optimisation of multi-stage flash distillation and reverse osmosis for desalination of saline process wastewater sources. Membranes 10(10):1–18. https://doi.org/10.3390/membranes10100265
5. Wang Z, Zhang Y, Wang T, Zhang B, Ma H (2021) Design and energy consumption analysis of small reverse osmosis seawater desalination equipment. Energies 14(8):1–18. https://doi.org/10.3390/en14082275
6. Bartholomew TV, Siefert NS, Mauter MS (2018) Cost Optimization of Osmotically Assisted Reverse Osmosis. Environl Sci & Technol, pp 8–02771. https://doi.org/10.1021/acs.est.8b02771
7. Jiang A, Wang J, Biegler LT, Cheng W, Xing C, Jiang Z (2015) Operational cost optimization of a full-scale SWRO system under multi-parameter variable conditions. Desalination 355:124–140. https://doi.org/10.1016/j.desal.2014.10.016
8. Lu YY, Hu YD, Zhang XL, Wu LY, Liu QZ (2007) Optimum design of reverse osmosis system under different feed concentration and product specification. J Membr Sci 287(2):219–229. https://doi.org/10.1016/j.memsci.2006.10.037

9. Khan MAM, Rehman S, Al-Sulaiman FA (2018) A hybrid renewable energy system as a potential energy source for water desalination using reverse osmosis: A review. Renewable and Sustainable Energy Reviews 97(January):456–477. https://doi.org/10.1016/j.rser.2018.08.049
10. Okamoto Y, Lienhard JH (2019) How RO membrane permeability and other performance factors affect process cost and energy use: A review. Desalination 470. https://doi.org/10.1016/J.DESAL.2019.07.004
11. Jiang A, Jiangzhou S, Cheng W, Wang J, Ding Q, Xing C (2015) Operational optimization of SWRO process with the consideration of load fluctuation and electricity price. IFAC-PapersOnLine 28(8):598–604. https://doi.org/10.1016/j.ifacol.2015.09.033
12. Jiang A, Biegler LT, Wang J, Cheng W, Ding Q, Jiangzhou S (2015) Optimal operations for large-scale seawater reverse osmosis networks. J Membr Sci 476:508–524. https://doi.org/10.1016/j.memsci.2014.12.005
13. Galizia A, Mamo J, Blandin G, Verdaguer M, Comas J, Rodríguez-Roda I, Monclús H (2021) Advanced control system for reverse osmosis optimization in water reuse systems. Desalination 518. https://doi.org/10.1016/j.desal.2021.115284
14. Sassi KM, Mujtaba IM (2013) Optimal operation of RO system with daily variation of freshwater demand and seawater temperature. Comput Chem Eng 59:101–110. https://doi.org/10.1016/j.compchemeng.2013.03.020
15. Hossam-Eldin A, Abed K, Youssef K, Kotb H (2019) Experimental investigation of energy consumption and model identification of reverse osmosis desalination system fed by hybrid renewable energy source under different operating conditions. IEEJ Transactions on Electrical and Electronic Engineering, pp 1409–1415. https://doi.org/10.1002/tee.22943
16. Zhang G, Hu W, Cao D, Liu W, Huang R, Huang Q, Chen Z, Blaabjerg F (2021) Data-driven optimal energy management for a wind-solar-diesel-battery-reverse osmosis hybrid energy system using a deep reinforcement learning approach. Energy Conversion and Management 227(October 2020):113608. https://doi.org/10.1016/j.enconman.2020.113608
17. Di Martino M, Avraamidou S, Pistikopoulos EN (2022) A neural network based superstructure optimization approach to reverse osmosis desalination plants. Membranes 12(2):1–26. https://doi.org/10.3390/membranes12020199
18. Sobana S, Panda RC (2014) Modeling and control of reverse osmosis desalination process using centralized and decentralized techniques. Desalination 344:243–251. https://doi.org/10.1016/j.desal.2014.03.014
19. Pascual X, Gu H, Bartman AR, Zhu A, Rahardianto A, Giralt J, Rallo R, Christofides PD, Cohen Y (2013) Data-driven models of steady state and transient operations of spiral-wound RO plant. Desalination 316(November):154–161. https://doi.org/10.1016/j.desal.2013.02.006
20. Jiang A, Ding Q, Wang J, Jiangzhou S, Cheng W, Xing C (2014) Mathematical modeling and simulation of SWRO process based on simultaneous method. Journal of Applied Mathematics 2014. https://doi.org/10.1155/2014/908569
21. Alsarayreh AA, Al-Obaidi MA, Patel R, Mujtaba IM (2020) Scope and limitations of modelling, simulation, and optimisation of a spiral wound reverse osmosis process-basedwater desalination. Processes 8(5):1–33. https://doi.org/10.3390/PR8050573
22. Mahadeva R, Manik G, Goel A, Dhakal N (2019) A review of the artificial neural network based modelling and simulation approaches applied to optimize reverse osmosis desalination techniques. Desalination and Water Treatment 156(April 2018):245–256. https://doi.org/10.5004/dwt.2019.23999
23. Ghoneim AA, Alabdulali HA (2020) Simulation and performance analysis of reverse osmosis water desalination system operated by a high concentrated photovoltaic system. Desalination and Water Treatment 177:29–39. https://doi.org/10.5004/dwt.2020.24895
24. Sobana S, Panda RC (2013) Development of a transient model for the desalination of sea/brackish water through reverse osmosis. Desalination and Water Treatment 51(13–15):2755–2767. https://doi.org/10.1080/19443994.2012.749376
25. Joseph A, Damodaran V (2019) Dynamic simulation of the reverse osmosis process for seawater using LabVIEW and an analysis of the process performance. Comput Chem Eng 121:294–305. https://doi.org/10.1016/j.compchemeng.2018.11.001
26. Kim JS, Chen J, Garcia HE (2016) Modeling, control, and dynamic performance analysis of a reverse osmosis desalination plant integrated within hybrid energy systems. Energy 112:52–66. https://doi.org/10.1016/j.energy.2016.05.050
27. Joseph A, Vasanthi D (2019) Performance analysis of PID control loops in desalination process using LabVIEW. In: 2019 Innovations in power and advanced computing technologies, i-PACT 2019, pp 1–9
28. Bartman AR, Zhu A, Christofides PD, Cohen Y (2010) Minimizing energy consumption in reverse osmosis membrane desalination using optimization-based control. Journal of Process Control 20(10):1261–1269. https://doi.org/10.1016/j.jprocont.2010.09.004
29. Singh VP, Rathore NS (2019) Whale optimisation algorithm-based controller design for reverse osmosis desalination plants. Int J Intell Eng Inform 7(1):77. https://doi.org/10.1504/ijiei.2019.10018732
30. Choi Y, Lee Y, Shin K, Park Y, Lee S (2020) Analysis of long-term performance of full-scale reverse osmosis desalination plant using artificial neural network and tree model. Environ Eng Res 25(5):763–770. https://doi.org/10.4491/eer.2019.324
31. Porrazzo R, Cipollina A, Galluzzo M, Micale G (2013) A neural network-based optimizing control system for a seawater-desalination solar-powered membrane distillation unit. Comput Chem Eng 54:79–96. https://doi.org/10.1016/j.compchemeng.2013.03.015
32. Aish AM, Zaqoot HA, Abdeljawad SM (2015) Artificial neural network approach for predicting reverse osmosis desalination plants performance in the Gaza Strip. Desalination 367:240–247. https://doi.org/10.1016/j.desal.2015.04.008
33. Gaudio MT, Coppola G, Zangari L, Curcio S, Greco S, Chakraborty S (2021) Artificial Intelligence-Based Optimization of Industrial Membrane Processes. Earth Syst Environ 5(2):385–398. https://doi.org/10.1007/s41748-021-00220-x
34. Barello M, Manca D, Patel R, Mujtaba IMM (2014) Neural network based correlation for estimating water permeability constant in RO desalination process under fouling. Desalination 345:101–111. https://doi.org/10.1016/j.desal.2014.04.016
35. Cabrera P, Carta JA, González J, Melián G (2017) Artificial neural networks applied to manage the variable operation of a simple seawater reverse osmosis plant. Desalination 416(October 2016):140–156. https://doi.org/10.1016/j.desal.2017.04.032
36. Karimanzira D, Rauschenbach T (2020) Deep Learning Based Model Predictive Control for a Reverse Osmosis Desalination Plant. J Appl Math Phys 08(12):2713–2731. https://doi.org/10.4236/jamp.2020.812201
37. Hafner R, Riedmiller M (2011) Challenges and benchmarks from technical process control, pp 137–169. https://doi.org/10.1007/s10994-011-5235-x
38. Sutton RS, Barto AG (2018) Reinforcement Leaning An Introduction, pp 481. A Bradford Book; 2ND edn
39. Yoo H, Kim B, Kim JW, Lee JH (2021) Reinforcement learning based optimal control of batch processes using Monte-Carlo deep deterministic policy gradient with phase segmentation. Comput Chem Eng 144. https://doi.org/10.1016/j.compchemeng.2020.107133

40. Bonny T, Kashkash M, Ahmed F (2022) An efficient deep reinforcement machine learning-based control reverse osmosis system for water desalination. Desalination 522(October 2021):115443. https://doi.org/10.1016/j.desal.2021.115443
41. Krishnan S, Boroujerdian B, Fu W, Faust A, Reddi VJ (2021) Air Learning: a deep reinforcement learning gym for autonomous aerial robot visual navigation. Mach Learn 110(9):2501–2540. https://doi.org/10.1007/s10994-021-06006-6
42. Dulac-Arnold G, Levine N, Mankowitz DJ, Li J, Paduraru C, Gowal S, Hester T (2021) Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. Mach Learn 110(9):2419–2468. https://doi.org/10.1007/s10994-021-05961-4
43. Sassi KM, Mujtaba IM (2012) Effective design of reverse osmosis based desalination process considering wide range of salinity and seawater temperature. Desalination 306:8–16. https://doi.org/10.1016/j.desal.2012.08.007
44. Palacin LG, Tadeo F, De Prada C, Johanna S (2011) Operation of desalination plants using renewable energies and hybrid control. Desalination and Water Treatment 25(1–3):119–126. https://doi.org/10.5004/dwt.2011.1433
45. El-Dessouky HT, Ettouney HM (2002) Reverse Osmosis. In: Fundamentals of salt water desalination, pp 409–437. Elsevier. https://doi.org/10.1016/B978-044450810-2/50009-9
46. Gambier A, Krasnik A, Badreddin E (2007) Dynamic modeling of a simple reverse osmosis desalination plant for advanced control purposes. Proc Am Control Conf 26:4854–4859. https://doi.org/10.1109/ACC.2007.4283019
47. Schock G, Miquel A (1987) Mass transfer and pressure loss in spiral wound modules. Desalination 64:339–352. https://doi.org/10.1016/0011-9164(87)90107-X
48. Zhu M, El-Halwagi MM, Al-Ahmad M (1997) Optimal design and scheduling of flexible reverse osmosis networks. J Membr Sci 129(2):161–174. https://doi.org/10.1016/S0376-7388(96)00310-9
49. Wilf M, Klinko K (1994) Performance of commercial seawater membranes. Desalination 96(1–3):465–478. https://doi.org/10.1016/0011-9164(94)85196-4
50. Syafiie S, Tadeo F, Palacin L, Prada CD (2008) Membrane modeling for simulation and control of reverse osmosis in desalination plants
51. Zhou SL, McMahon TA, Walton A, Lewis J (2002) Forecasting operational demand for an urban water supply zone. J Hydrol 259(1–4):189–202. https://doi.org/10.1016/S0022-1694(01)00582-0
52. Donkor EA, Mazzuchi TA, Soyer R, Alan Roberson J (2014) Urban Water Demand Forecasting: Review of Methods and Models. J Water Resour Plan Manag 140(2):146–159. https://doi.org/10.1061/(asce)wr.1943-5452.0000314
53. Lillicrap, T.P, Hunt, J.J, Pritzel, A, Heess, N, Erez, T, Tassa, Y, Silver, D, Wierstra, D.: Continuous control with deep reinforcement learning. In: 4th International conference on learning representations, ICLR 2016 - conference track proceedings (September) (2016). arXiv:1509.02971
54. Mnih, V, Kavukcuoglu, K, Silver, D, Graves, A, Antonoglou, I, Wierstra, D, Riedmiller, M.: Playing Atari with Deep Reinforcement Learning, pp 1–9 (2013). arXiv:1312.5602
55. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533. https://doi.org/10.1038/nature14236
56. Saavedra A, Valdés H, Mahn A, Acosta O (2021) Comparative analysis of conventional and emerging technologies for seawater desalination: Northern chile as a case study. Membranes 11(3). https://doi.org/10.3390/membranes11030180

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.