Applied Energy 268 (2020) 114943

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Operation scheduling in a solar thermal system: A reinforcement learning-based framework

Camila Correa-Jullian[a,*], Enrique López Droguett[a,b], José Miguel Cardemil[a]

[a] Mechanical Engineering Department, Universidad de Chile, Santiago, Chile
[b] Center for Risk and Reliability, University of Maryland, College Park, USA

HIGHLIGHTS

- Condition-based decision-making framework with Reinforcement Learning.
- Framework used for scheduling the operation of a solar thermal system.
- Synthetic data generation from TRNSYS simulation.
- Sensitivity analysis based on energy-related KPI for selected actions.
- Beneficial alternative schedules found for July (low solar radiation) setting.

ARTICLE INFO

Keywords:
Solar hot water systems
Reinforcement learning
Intelligent control systems
Condition-based decision making
Q-learning
Machine learning

ABSTRACT

Reinforcement learning (RL) provides an alternative method for designing condition-based decision making in engineering systems. In this study, a simple and flexible RL tabular Q-learning framework is employed to identify the optimal operation schedules for a solar hot water system according to action–reward feedback. The system is simulated in TRNSYS software. Three energy sources must supply a building's hot-water demand: low-cost heat from solar thermal collectors and a heat-recovery chiller, coupled to a conventional heat pump. Key performance indicators are used as rewards for balancing the system's performance with regard to energy efficiency, heat-load delivery, and operational costs. A sensitivity analysis is performed for different reward functions and meteorological conditions. Optimal schedules are obtained for selected scenarios in January, April, July, and October, according to the dynamic conditions of the system. The results indicate that when solar radiation is widely available (October through April), the nominal operation schedule frequently yields the highest performance. However, the obtained schedule differs when the solar radiation is reduced, for instance, in July. On average, with prioritization of the efficient use of both low-cost energy sources, the performance in July can be on average 21% higher than under nominal schedule-based operation.

1. Introduction

It is estimated that the energy consumption for domestic hot water (DHW) accounts for up to 10% of the total end energy use [1]. In this context, solar hot water (SHW) systems are a sustainable alternative to conventional fossil fuel- or electricity-driven devices for delivering low-grade heat in residential and commercial buildings, as well as industrial applications [2]. As solar resource is variable, thermal storage and auxiliary thermal sources are frequently integrated to increase the availability of the system [3]. In addition to the main thermal and hydraulic components in an SHW system, the integration of a comprehensive control system is critical to ensure the high performance of each component and for the thermal energy management of the system [4]. Coordination mechanisms must be considered between the different elements of the system (hot-water production, demand profiles) to reduce the operational costs of the system [5]. For instance, the optimal flow control and energy-efficiency strategies for SWH systems with forced circulation have been studied [6,7] to increase the thermal energy output and reduce the pump energy consumption. Furthermore, time delays caused by thermal inertia can potentially have negative consequences for the control-loop performance, owing to the introduction of unstable behavior [8]. In this context, the development of intelligent and flexible control systems is essential for improving the energy-management strategies, to enhance not only the energy savings

* Corresponding author.
E-mail address: camila.correa.j@ug.uchile.cl (C. Correa-Jullian).

https://doi.org/10.1016/j.apenergy.2020.114943
Received 12 November 2019; Received in revised form 26 March 2020; Accepted 1 April 2020
Available online 24 April 2020

Acronyms

RL	Reinforcement Learning
SHW	Solar Hot Water System
TRNSYS	Transient System Simulation Tool
KPI	Key Performance Indicators
DHW	Domestic Hot Water Systems
HVAC	Heating, Ventilation and Air Conditioning Systems
SHIP	Solar Heat Industrial Processes
DL	Deep Learning
EVs,	Electric Vehicles
PV/T	Thermo-Photovoltaic Systems
RBC	Rule-Based Control
MDP	Markov Decision Process
FPC	Flat Plate Solar Collector
ETC	Evacuated Tube Solar Collector
IAM	Incidence Angle Modifier
GHI	Global Horizontal Irradiance
DFI	Diffuse Horizontal Irradiance
DNI	Direct Normal Irradiance

Subscripts

i	fluid inlet
a	ambient
sp	speed
dir	direction
o	optical property
T	transverse
L	longitudinal
t	time
G	global KPI
v	tempering valve outlet

Nomenclature

s_t	state of the system at time t
S	state space
a_t	action taken by the agent at time t
$A(s_t)$	state-dependent action Space
$p(s_{t+1} \mid s_t, a_t)$	action-dependent transition probabilities between states
$R(s_t, a_t)$	action-state dependent reward function
γ_r	discount factor of future rewards
$Q(s, a)$	Q-function or action-value function
r_t	reward perceived at time t
π	policy guiding the permissible actions taken
α_r	learning rate
R_k	total reward at the end of episode
τ_o	transmittance
α_o	absorptance
θ	incidence angle of solar radiation
$K_{\tau\alpha}$	optical efficiency factor
F_R	heat removal factor
U_L	overall thermal loss coefficient
T	temperature
$(\tau\alpha)_n$	transmittance-absorptance product at normal incidence angle
η_0	optical efficiency of ETC solar collector
a_{1a}, a_{2a}	first and second order loss coefficients of ETC solar collector
I	solar irradiance
M	meteorological conditions
W	wind
P	atmospheric pressure
O	operational condition
E	energy gain from solar field
CH	electrical power input to the chiller
HP	energy rate delivered by heat pumps
α, β, γ	weights for sensitivity analysis

and efficiency but also the user satisfaction.

With the increased availability of monitoring data and the development of versatile data-driven applications such as machine learning, various algorithms have been developed and applied to perform decision-making tasks in various research areas involving engineering systems. These applications have mainly focused on design, manufacture, operation, and maintenance [9–12]. Reinforcement learning (RL) offers an alternative approach to decision making based on the system's experience and conditions. In addition to the physical, cost, and comfort restrictions of the system, time-deferred effects of actions and policies are taken into account and identified as beneficial or damaging to its overall goals. In a typical RL setting, the objective is to extract optimal decision policies under a sequence of states, actions, and rewards with which the decision-maker (called the agent) learns to optimize a cost function by interacting with an environment. Diverse RL techniques have been developed for specific applications. An RL agent can be trained using historical operation and sensor data, as well as stochastic ambient effects, thus shifting the focus from model- and risk-based analysis to data-driven condition-based decisions.

Relevant application areas in different engineering systems are listed as follows: the design and optimization of production and maintenance scheduling policies [13,14], power-grid management [15], demand response models to market electricity pricing [16], and autonomous driving [17]. Recently, RL algorithms have been applied to adaptive controls in energy systems. This has allowed the integration of renewable energy sources into electrical grids and real-time decision making for microgrid energy management [18,19]. A review of these applications was presented in [20], in which the studied systems included heating, ventilation, and air conditioning (HVAC) systems [21], DHW systems [22–24], smart appliances, hybrid and electric vehicles (EVs) [25], and distributed generation coupled to energy storage [26]. The main objective of the RL agent in these studies is to manage the energy costs, user satisfaction and comfort; manage the peak-demand performance; and reduce the fuel consumption. For instance, in [27], an adaptive and occupant-centered controller for lighting in commercial buildings was successfully implemented, balancing the occupant comfort and energy consumption, in contrast to schedule- and occupancy-based control scenarios.

In thermal systems, similar adaptive controls for DHW systems have been developed to integrate renewable energy sources and reduce the overall energy consumption. The mass flow rate in combined thermo-photovoltaic (PV/T) systems and geothermal heat pumps in buildings were optimized according to the heat demand, net output power, and optimal operational temperatures via numerical simulations [28]. Methods such as Tabular Q-learning and Batch Q-learning with Memory Replay were compared with standard rule-based control (RBC). All the tested solar PV/T RL control methods outperformed the RBC by > 10% after the third year of simulation. Additionally, the more general case of a heterogeneous cluster of thermal demand response electric water heaters was investigated using Batch RL algorithms [29]. By using a detailed stratified thermal model of storage tanks, the RL was compared with the traditional hysteresis controller. It outperformed the traditional controller by reducing the electricity cost while maintaining the user satisfaction under time-varying electricity-price scenarios.

Experimental validation of the demand response of DHW buffers was performed, with the aim of optimizing the heating cycles for maximizing the use of local PV production. Compared with standard thermostat controllers, the use of a RL-based algorithm combined with a data-driven weather forecast increased the amount of PV energy delivered to the DHW system by > 20% [30].

This study analyzes the operation of an SHW subsystem that interacts with other heat sources to deliver hot water to a university building. Three heat sources deliver energy to the system sequentially: (1) low-temperature solar thermal collectors; (2) mid-temperature excess heat from a heat-recovery chiller; and (3) high-temperature conventional air–water heat pumps. Given the design of the system, the heat driven by the solar collectors and the heat-recovery chiller is considered low-cost; it is attributed to the capture of free solar radiation and waste heat (a byproduct of the operation of the water-cooled chiller), respectively. An RL agent is given control over the operation of these two low-cost sources, seeking to find the optimal operation schedule of the circulation pumps that manage these heat sources, while prioritizing the participation of the solar field. To examine this balance among the energy consumption, renewable participation, and hot-water supply at the design temperatures, different key performance indicators (KPIs) are proposed to guide the rewards that the agent perceives according to the energy efficiency, renewable capacity, and economic considerations. The daily operation of the system under different scenarios is investigated. As is demonstrated, the RL agent can extract the optimal policies from the permissible actions in the system by incorporating its thermal behavior, reflecting the importance given to each KPI. Therefore, the main contribution of this work is the development of a framework for condition-based operation scheduling for SHW systems, emphasizing the flexibility of production policies via RL with the Q-learning algorithm. This approach considers both the comfort-demand profiles and the overall production goals in the operation scheduling. Additionally, with the incorporation of the availability of solar radiation as a key driver of the system's goals, the overlooked complexity of the thermal inertia behavior of the solar thermal system plays a pivotal role.

The remainder of this paper is organized as follows. Sections 2 and 3 provide the necessary background for RL and SHW systems, respectively. Section 4 describes the proposed framework, including the simulation approach and the construction of the RL state–action–reward setting. Section 5 presents the results and a detailed sensitivity analysis of the KPI. Finally, Section 6 presents concluding remarks.

2. Reinforcement learning

The basic concept of an RL algorithm is the interaction between an agent and its environment through feedback loops based on actions and rewards, leading to a specific goal. Through this feedback loop, the agent can derive an optimal policy given the restrictions established by the environment (e.g., a physical asset or system). The agent–environment interaction sequence is mathematically described by a Markov decision process (MDP). An MDP formally describes a stochastic dynamic system defined by a tuple $\langle S, A, T, R\rangle$. The elements that define this system are as follows:

(1) A discrete and finite state space that, at a certain time t, is represented by $s_t \in S$.
(2) A finite action space that depends on a given state $a_t \in A(s_t)$. The effect on an action on the state s_t will cause the transition to the next state s_{t+1}.
(3) Transition probabilities between states give the previous state–action tuple $p(s_{t+1}| s_t, a_t)$, which is defined as T: $S \times A \times S \rightarrow [0, 1]$. These transitions are Markovian if the future state resulting from an action is only dependent on the present system's state.
(4) A reward function $R(s_t, a_t)$ defined as R: $S \times A \times S \rightarrow \mathbb{R}$, which represents the consequence of taking an action a_t at a certain state s_t.
(5) A discount factor $\gamma_r \in (0, 1)$ that allows the maximum future reward sequence that a certain action a_t can possibly yield to be quantified. Thus, the feedback loop that the agent perceives is based on future discounted rewards to encompass the time delay between the action and the resulting reward.

Various RL algorithms have been developed and widely used in a variety of settings, including traditional algorithms such as Q-learning [31] and SARSA [32] and more complex deep learning (DL) algorithms, such as the Deep Q-Networks (DQN) for feature enhancement [33] and human-level decision making [34]. In practice, the difference between these algorithms lies in how the worth of an action is translated into perceptible rewards through explicit functions (as in Q-learning and SARSA) or represented by the trained weights of artificial neural networks in DL settings. Among these approaches, model-free RL algorithms are a class of data-driven approaches in which the internal logic of the system is represented as a black box from the agent's perspective. Model-free RL algorithms do not explicitly have the transition-probability matrix of the MDP, but they obtain an approximation of the action-values by exploring and interacting with the environment or system [35]. This allows the exploration of large action–state spaces with a reduced computational cost compared with dynamic-programming approaches [36].

The framework used in this study, i.e., tabular Q-learning, is based on a modified version of the value function method, where the utility or worth of each corresponding action–state is solely expressed by a quantified value (Q-value), which is given by an action–value function [35]. An action–value function $Q(s, a)$, which is also called a Q-function, is defined as the maximum expected return r_t of a specific action a_t over a state s_t given a policy π, as shown:

$$Q(s, a) = \max_{a} E(r_t|s_t = s, a_t = a, \pi). \tag{1}$$

The future optimal action a^* and the optimal policy π^* are then expressed through the Q-function as follows:

$$Q^*(s, a) = \max_{a^*} Q(s, a^*) \tag{2}$$

$$\pi^*(s) = \operatorname*{argmax}_{a} Q^*(s, a) \tag{3}$$

where $Q^*(s, a^*)$ represents the maximum future Q-value from the possible action–state (s, a) tuples. The optimal policy π^* is extracted as the set of actions that maximize the Q-value. The Q-learning algorithm is a value iteration method adaptation of the Bellman Equation used to estimate $Q^*(s, a)$. The linear approximation of the Q-function updates the corresponding Q-values using the maximum Q-value available for the present state [20]:

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha[r_t + \gamma_r Q^*(s', a') - Q_t(s, a)] \tag{4}$$

where $Q_{t+1}(s, a)$ represents the updated Q-values for the action–state tuple calculated from the present Q-value $Q_t(s, a)$ and the present reward function r_t, as well as the discounted maximum future Q-value given the possible (s, a). Here, α represents the learning rate of the algorithm, and γ_r represents the discount factor. Both are in the range of [0,1].

The Q-learning algorithm is an off-policy learning method, which uses a simulator of the environment as an inexpensive method for generating and sampling a large number of training examples that map out the real-life problem under analysis [37]. In comparison, an on-policy setting requires continuous interactions between the agent and the environment to determine future states of the system to simultaneously derive the optimal policy while mapping out the state space. Such an on-policy is used in SARSA. Additionally, when interacting with a deterministic setting, the Q-learning method allows the most promising set of actions given a specific environment state to be identified [38]. A typical issue when deriving the optimal control policy is

exploration versus exploitation of the environment. Exploration can be introduced as random mapping of the Q-values before the systematic selection of the best output [35]. As the Q-learning method seeks the highest available future reward to select an action, in long decision-making sequences, the agent may become biased according to the first experienced state of the system. To counter this, the Q-learning method requires a detailed configuration of the action–state space to avoid biased estimation of the Q-values. In this case, as the action–state space is reduced and the reward function's values are directly calculated from the state of the system, a greedy policy is used to maximize the exploitation through the training and evaluation phase; i.e., the taken action always yields the greatest reward given the present state.

In this study, a tabular approach is used for the Q-learning method, given that the action–state space is reduced. This, in contrast to larger action–state spaces, which benefit from approximate solutions to estimate the Q-values, such as the use of neural networks [34]. This tabular approach has a simple formulation and straightforward implementation, as described in the flowchart presented in Fig. 1. Depending on the agent–environment interaction setting, the agent retrieves information on the state of the system via continuous or periodical observations. Then, according to the observed Q-values, the agent selects the available action with the highest value, retrieving the reward or penalty for it, updating the Q-value, and causing the transition to the next state. Both parameters α and γ_r focus on the learning rate and the monitored effects of actions in the system. The parameter α defines the rate of new knowledge rewriting already stored information. Low values of γ_r favor short-term effects, and the agent is focused on long-term developments in the system with values close to 1 [35].

Based on the cited literature, in the context of maintenance and operation scheduling, the use of RL can be applied for downtime and maintenance costs minimization or the maximization of production goals, according to quantifiable rewards extracted from the system's monitoring data. Additionally, this process considers the delay between an action and the corresponding reward, which is known as a temporal credit-assignment problem and is particularly important in systems with long response times, such as thermal systems, owing to their inherent thermal inertia.

3. Solar hot water systems

Both domestic and industrial applications of SHW systems exhibit potential for reducing the greenhouse-gas emissions, and currently one of the most widely used water heating systems worldwide [39,40]. The installed capacity of SHW systems for buildings increased by approximately 250% over the past decade, and by the end of 2017, it exceeded 470 GW_{th}, surpassing the 402-GW_{el} installed capacity of solar photovoltaic technologies. During the same year, the operation of SHW systems achieved reductions of 41.7 million tons of oil and 134.7 million tons of CO_2 emissions [41]. Additionally, in recent years, there has been increasing interest in large-scale solar thermal systems (> 350 kW_{th}; 500 m^2) integrated into the building design, district heating networks, and solar heat industrial process (SHIP) applications [42]. Despite their growth and outstanding benefits, the global installed capacity of solar thermal systems covered only 2.1% of the total demand for space and water heating in 2018 [43].

Most commercial and residential SHW applications are based on non-concentrating solar thermal technologies, which are mainly classified into flat plate (FPC) and evacuated tube (ETC). In the latter category, the heat-pipe ETC variation combines advantages of both standard types of solar collectors. These collectors incur relatively low maintenance costs and are less affected by unfavorable weather conditions than the FPC; moreover, they have good anti-freezing properties and a high thermal conductivity to prevent internal overheating, which is a frequent issue in the ETC [44]. Under standard EN 12975-2 testing conditions, ETC efficiency values obtained are estimated within 50–60% [45]. Ayompe et al. [46] performed year-round energy performance monitoring for different SHW configurations, reporting an annual solar fraction, collector efficiency, and system efficiency of 40.2%, 60.7%, and 50.3%, respectively, for a heat-pipe ETC.

3.1. SHW case study

The SHW system analyzed in this study is presented in Fig. 2, in which three heating stages operating in series can be identified. The red and blue lines represent the flow of hot and cold streams in the system, respectively. The goal of this heating system is to deliver a load of 24,000 L at 45°C with a daily operating schedule from 7AM to 9PM. This setting is based on the installation located at a building at the Physical and Mathematical Sciences Faculty of the Universidad de Chile, in Santiago, Chile, which was previously studied in [47] (the simulation's representability of the system was assessed). The first heating stage corresponds to the renewable section, consisting of the solar field of a heat-pipe ETC, a preheating storage tank, and the mains water inlet to the system. The second section integrates a hot-water flow from a heat-recovery chiller, which receives a water flow previously heated by the solar loop. The tank is designed to store water between 35 and 40°C. Finally, in the heating section, electricity-driven heat pumps (with a temperature set point of 50°C) and an additional mains water inlet are used to regulate the temperature of the water delivered to the load through a tempering valve, which is dispatched at 45°C. Heat exchangers and constant-speed centrifugal pumps connect the different closed loops. The monitored variables mainly consist of the temperature and operational status of the aforementioned equipment. The solar field is composed of a total absorption area of 105.6 m^2 of the ETC, which is tilted at 15° and is north-oriented. Tables 1 and 2 present the optical and thermal efficiencies of the installed solar collector (Hitek Solar NSC 58-30 model[1]).

Table 1 presents the IAM values obtained under the test conditions for the selected ETC model, expressing how the optical properties, such as the transmittance τ_o and absorptance α_o, of the solar collector's components, vary depending on the incidence angle θ of solar radiation on the collector's surface. This is expressed as the $K_{\tau\alpha}$ factor in Eq. (5), affecting the solar collector's overall efficiency, as well as F_R, representing the heat-removal factor; U_L, the overall thermal-loss coefficient; T_i, the fluid inlet temperature; T_a, the ambient temperature; and $(\tau\alpha)_n$, the transmittance-absorptance product at the normal incidence angle [48]. In the case of the ETC, these models present non-symmetrical cover optical properties; thus, the effects of both the transverse θ_T and longitudinal θ_L incidence angles are taken into account, as indicated by Eq. (6) [49]. The heat-recovery chiller used is a Thermocold CWC Prozone 1320 Z C model of 254 kW nominal capacity, with a set point of 10 °C. All the storage tanks have a constant volume of 4 m^3.

$$\eta_i = F_R[G_T K_{\tau\alpha}(\tau\alpha)_n - U_L(T_i - T_a)] \quad (5)$$

$$K_{\tau\alpha}(\theta) = K_{\tau\alpha}(\theta_T)\cdot K_{\tau\alpha}(\theta_L) \quad (6)$$

3.2. Simulation approach

A physical representation of the SHW system was developed in the Transient System Simulation Program (TRNSYS) [50] to generate synthetic data that represent the different scenarios in which the agent operates. The different scenarios are a result of altering the control system and operating hours of the solar and heat-recovery chiller circulation pumps.

TRNSYS operates through types (or component blocks) which

[1] For reference, the datasheet reports that a difference of 10 °C between ambient and mean solar collector temperature results in a power output per collector unit of 1686 W, with an approximated total of 74 kW for the complete solar field.

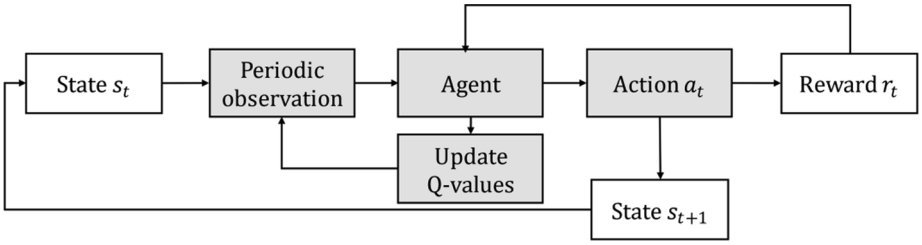

Fig. 1. Information flowchart for a Q-learning agent with periodic observations.

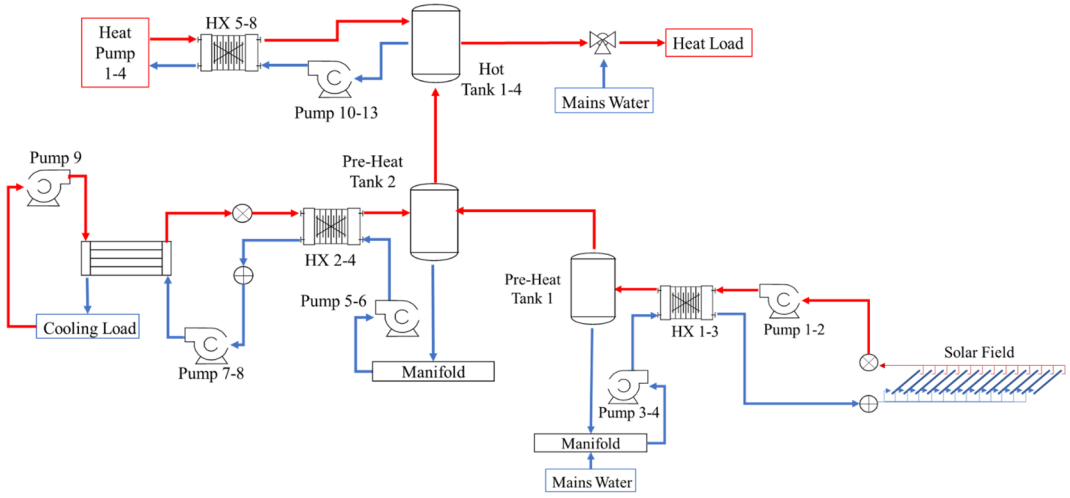

Fig. 2. Schematic of the SWH system, showing the preheating and heating sections.

Table 1
Incidence angle modifier (IAM) values for the Hitek Solar NSC Heat Pipe ETC solar collector.

	10°	20°	30°	40°	50°	60°	70°
$K_\theta(\theta_T)$	1.010	1.019	1.056	1.151	1.452	1.462	1.261
$K_\theta(\theta_L)$	0.999	0.994	1.018	0.974	0.952	0.913	0.833

Table 2
Thermal capacities of the Hitek Solar NSC Heat Pipe ETC solar collector (parameters related to the aperture area).

Parameter	Reference Value
η_0	0.618
a_{1a} [$\mathbf{W/(m^2K)}$]	1.377
a_{2a} [$\mathbf{W/(m^2K^2)}$]	0.018
Effective thermal capacity [$\mathbf{kJ/(m^2K)}$]	5.684

integrate experimentally validated or theoretical equations to express the operation of the thermal, hydraulic, and control components present in the system. TRNSYS has been applied in several applications related to solar energy systems and validated experimentally. Different technologies, such as flat plate and heat pipe evacuated tube collectors, have been validated in diverse applications [51], such as pool heating [52], large heating networks [53], and coupling with HVAC systems [54], as well as representing the stratification of hot-water storage systems [55]. While stochastic effects can be introduced in a TRNSYS setting, simulations of thermal behavior are deterministic in nature.

For the SHW system studied, the nominal design conditions are introduced in the corresponding operational types representing the solar collectors, heat-recovery chiller, heat pumps, heat exchangers, single-speed centrifugal pumps, and control system. Actual meteorological data, e.g., the solar radiation, ambient temperature, wind speed, and wind directions, are used to describe the environment in which the SHW system operates. Additionally, the following simplifications are made.

- The temperature of the water entering from the mains system is calculated according to the numerical correlations presented in [56] and adapted to the local weather.
- Heat pumps are represented by an auxiliary water heater configured with nominal heating capacities corresponding to the design conditions.
- The system's control has automatic responses, as follows:
  - o The flow from the preheating tanks to the heat tanks is equivalent to the flow required by the users. Replacement water from the mains system is introduced into the solar preheating tank (Pre-Heat Tank 1 in Fig. 2). This allows the mass of water stored in the system to be constant.
  - o For safety purposes, if the outlet temperature dispatched from the heat tanks is > 45°C, mixing valves are activated, and the mains water is introduced until the temperature is below this threshold. The extra water is then reintroduced into the hot-water tank circulation flows.
- The operational statuses of both the solar and heat-recovery chiller circulation pumps (Pumps 1, 2 and 7, 8 in Fig. 2, respectively) are altered to simulate the different scenarios, reflecting the actions that the RL agent can control during the daily operation schedule.

In the design of the SHW system, it is considered that the heat load is delivered at a constant rate. However, on the basis of operational experience of both the users and the building's management team, an estimate of the daily demand profile is presented in Fig. 3, where the greatest demand occurs between 12 PM and 2 PM, coinciding with the highest availability of solar radiation during the year. This profile is introduced in the TRNSYS deck, along with the component's design characteristics. For the agent–environment interaction, samples are

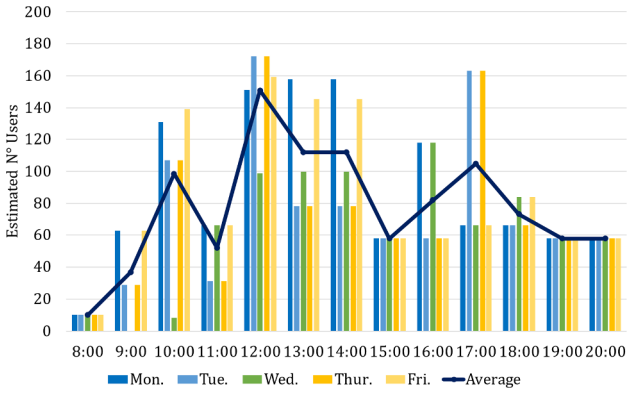

Fig. 3. Estimated hot-water demand weekly profile for the SWH system (in number of users).

extracted for four different meteorological conditions throughout the year. As it is of interest to extend the use of the solar field and increase the thermal efficiency of the system, the RL agent is presented with four evenly spaced time windows with remarkably different solar-radiation daily profiles. These daily profiles are simulated between the 1st and 10th of January, April, July, and October to examine the sensitivity and adaptability of the agent's decisions to solar-radiation availability and thus the performance of the solar field. Sample solar-radiation profiles for the Global Horizontal (GHI), Diffuse Horizontal (DFI), and Direct Normal (DNI) components are presented in Fig. 4.

The performance of the ETC mainly depends on the GHI solar radiation, as its geometry allows for passive sun-tracking throughout the day. While the radiation profile for July is the lowest, it also exhibits a smooth behavior, similar to the January profile. The radiation profiles for April and October exhibit more irregular behavior. Finally, the selection of the variables and the analyzed timeframes depends on the thermal behavior of the system and is thus based on the previously presented solar-radiation profiles and the thermal inertia of the system.

4. Proposed solar thermal system operation scheduling framework

In this section, the condition-based operation scheduling framework is presented, combining the data-driven decision-making algorithm with performance indicators. This allows the exploration of the optimal operating configuration for an SHW system under different meteorological conditions and overall policy goals. Here, a balance among reducing the energy consumption, increasing the participation of the solar field, and reaching the design outlet temperatures is desired.

In previous studies [57–60], RL applications have focused on maintenance scheduling, given the ability to properly define states, actions, and rewards. The nature of the environment, as well as the agent–environment interactions, depends on the nature of the TRNSYS simulation. By building detailed simulation models, a significant amount of synthetic data can be generated for training the RL model. To simplify the experimental setting, all possible scenarios are simulated beforehand. Thus, the interaction between the agent and the environment is limited. Future states are dependent on the previous actions taken; thus, there are inaccessible action–state combinations during the interaction episode. Furthermore, the RL agent interacts with a deterministic environment, with finite-length episodes that cover the daily operation of the SHW system. The transition probabilities for each state are stationary. Observations, states, and actions performed on the environment are defined as discrete interactions with the agent. As the action–state space is well defined and limited, the values of the Q-function are approximated via a tabular approach. A weighted sum of KPIs is used as the reward function for balancing the performance of the

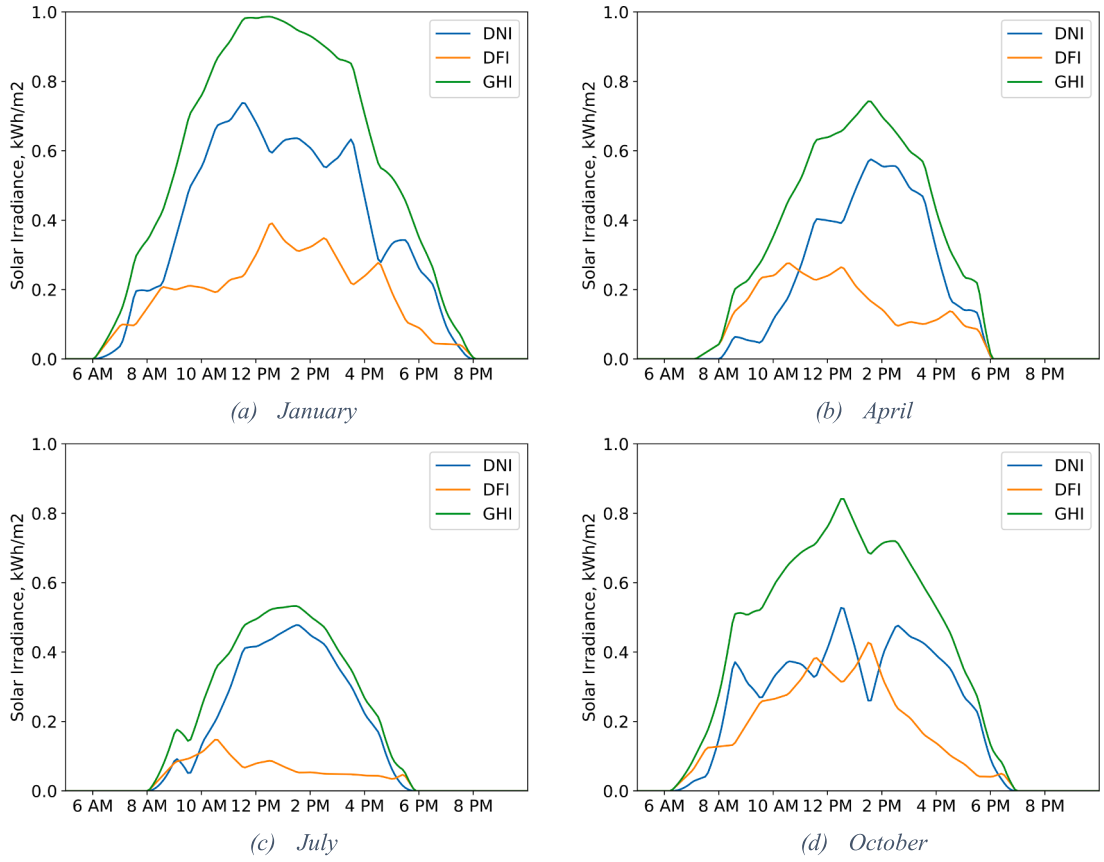

Fig. 4. Solar-radiation daily profiles for selected days in January, April, July, and October.

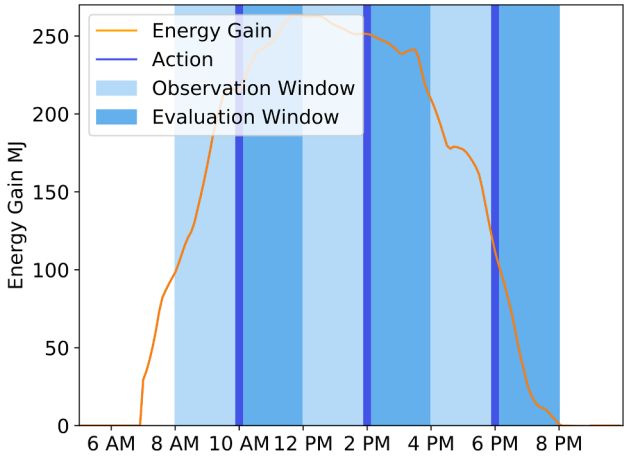

Fig. 5. Daily interaction instances and assessment time windows.

system with regard to the energy efficiency, heat-load delivery, and operational costs.

The proposed approach considers (a) the short-term effect of actions on the system in daily operation scheduling for the main thermal components; (b) simple decisions regarding the control system of side components (i.e., on/off signals); (c) no downtime as a consequence of these actions, other than the inherent thermal inertia; and (d) a reward function that allows the prioritization of different aspects of the SHW's operation, which are represented by different KPIs.

4.1. State space and variables

To define the state space, the temperature, meteorological conditions, and energy flow variables are extracted from the TRNSYS simulation. The state at each timestep is then defined by a function of the monitored variables given the combination of actions and observations of the system:

$$s_t = f\left(T{:}\{T_{1,\dots,N},\ T_{amb}\},\ I{:}\{GHI,\ DFI,\ DNI\},\ M{:}\{RH,\ W_s,\ W_d,\ P\},\ O{:}\{O_{1\dots N}\}\right) \quad (7)$$

Here, the state s_t is constructed as a function of the following variables: T and O represent the temperature and operational status monitored for N components, respectively; I represents the solar-radiation measurements; and M represents other meteorological conditions. The solar irradiance measurements include the DNI, GHI, and DFI components, and other ambient measurements include the relative humidity (RH), wind speed and direction (W_{sp}, W_{dir}), and atmospheric pressure (P), representing the meteorological conditions. Finally, a binary (0,1) operational status is used as a substitute for flow measurements, as single-speed centrifugal pumps are used.

Special consideration is given to the use of radiation values, as they provide valuable information about the expected performance of the solar field in deterministic steady-state simulations. Additionally, information regarding the temperatures and heat flows is obtained for each component. The components selected to describe the state of the system are the solar field, heat-recovery chiller, heat pumps, hot-water storage tanks, mains water inlet, and tempering valve outlet. These main components affect the overall efficiency of the system, which consists of the three heat sources and three temperature points relevant to the delivery of the heat load. This selection results in a total of 43 variables, which are arranged in a multi-dimensional array [observations, time window, variables] for each daily profile. A sampling frequency of 6 min is considered to account for the natural thermal inertia of the system and the variability of the solar radiation, as minute-based measurements may suffer from uncertainties caused by atmospheric phenomena [47].

4.2. Agent–environment interaction

For simplifying the interaction between the agent and the environment and considering the thermal inertia of the SHW system, the agent acts sequentially on the environment at three instances daily, which are denoted as "interaction windows." Given the size of the system, it is estimated that at least 1 h is needed to observe significant thermal changes in the monitored points [47]. During these three daily observation-action time windows, the agent assesses the current situation and selects a future action according to the obtained KPI. As the possible future states depend on the present state, the decision making process of the control system is condition-based, i.e., sensible to both the meteorological conditions and the thermal performance of the system. The number of interaction windows is selected to observe the development of the system's state during the insolation hours of the design operation schedule (7 AM to 9 PM) in time windows that allow the assessment of the selected action's effect considering the inertia of the system.

To cover the hours in which solar radiation is mostly available during the user consumption profile shown in Fig. 3, the action events are distributed at 10 AM, 2 PM, and 6 PM. Thus, a daily operation is divided into three time windows: 8 AM to 12 PM, 12 PM to 4 PM, and 4 PM to 8 PM. Additionally, these time-windows are divided into two phases: an observation window in which the agent can assess the present state of the SHW system and select an adequate action and an assessment window in which the rewards are calculated. In Fig. 5, this temporal arrangement is superimposed on the energy-gain profile of the solar collector for January.

4.3. Permissible actions

Permissible actions in the system are focused on the operation of the solar field and heat-recovery chiller circulation pumps. The objective is to investigate operation schedules that may not be intuitive while also considering the effects of the thermal inertia of the SHW system, as well as different meteorological conditions, economic factors, and heat loads. Among the three heat sources in the SHW system, only two (the

Table 3
Admissible paths for solar and chiller by-pass pump control.

Path	10 AM	2 PM	6 PM	Path	10 AM	2 PM	6 PM	Path	10 AM	2 PM	6 PM
1	{1,1}	{1,1}	{1,1}	9	{1,0}	{1,1}	{1,1}	17	{0,1}	{1,1}	{1,1}
2	{1,1}	{1,1}	{1,0}	10	{1,0}	{1,1}	{1,0}	18	{0,1}	{1,1}	{1,0}
3	{1,1}	{1,1}	{0,1}	11	{1,0}	{1,1}	{0,1}	–	{0,1}	{1,1}	{0,1}
4	{1,1}	{1,0}	{1,1}	12	{1,0}	{1,0}	{1,1}	19	{0,1}	{1,0}	{1,1}
5	{1,1}	{1,0}	{1,0}	13	{1,0}	{1,0}	{1,0}	20	{0,1}	{1,0}	{1,0}
6	{1,1}	{1,0}	{0,1}	14	{1,0}	{1,0}	{0,1}	–	{0,1}	{1,0}	{0,1}
7	{1,1}	{0,1}	{1,1}	15	{1,0}	{0,1}	{1,1}	–	{0,1}	{0,1}	{1,1}
8	{1,1}	{0,1}	{1,0}	16	{1,0}	{0,1}	{1,0}	–	{0,1}	{0,1}	{1,0}
–	{1,1}	{0,1}	{0,1}	–	{1,0}	{0,1}	{0,1}	–	{0,1}	{0,1}	{0,1}

solar field and the chiller bypass pumps) are treated as independent variables in the TRNSYS simulation. The heat pumps are controlled by an integrated thermostat with a set point of 50°C. To focus on the guiding principles underlying the control logic (whether to (a) prioritize the renewable capacity factor, (b) maximize the ratio between free and priced energy, or (c) minimize the energy consumption), no further assumptions are made about the operation logic of the SHW (other than the operation schedule).

A state of the system is defined by the operational condition of the single-speed centrifugal pumps that recirculate flows from the solar field and the heat-recovery chiller as a tuple $\{x, y\}$. The first digit represents the state of the solar circulation pump, and the second digit represents the state of the heat-recovery circulation pump. At each of the three agent–environment interactions, there are four possible states considering both components and on/off or 1/0 pump control signals: {1,1}, {1,0}, {0,1}, and {0,0}. For practical reasons, the latter state in which both heat sources are shut-off is ignored. As the participation of the solar field is of interest, daily sequences involving two periods in which it is shut-off from operation are also ignored. Table 3 presents the 20 possible sets of actions, paths, or scenarios per day, which are represented by a trio of tuples of control signals for the solar and chiller circulation pumps. Here, paths are defined by tuples (solar pump control, chiller bypass pump control) where 1/0 represents the on/off conditions, respectively.

4.4. Condition-based rewards

KPIs are used to describe the effects of selected actions on the state of the system. These KPIs are related to the energy consumption and overall efficiency. From the TRNSYS simulation, the following main variables and components are used to analyze the agent's decisions: (a) the total energy gain of the solar field, (b) the required electrical power of the heat-recovery chiller, (c) the energy delivered to the stream by the heat pumps, and (d) the outlet temperature of the tempering valve. These variables are selected because they can be corroborated with data collected in a real system through temperature and control signal measurements. While the first four variables describe the energy gains and consumptions of the system, the latter is crucial for determining whether the hot-water demand is satisfied under the design conditions (i.e., 45°C).

The proposed KPIs are presented in Eqs. (8)–(12), where three main ideas are expressed: the renewable capacity factor, amount of hot water supplied, and energy consumption. KPI_1 and KPI_2 both represent a renewable capacity factor. The first covers only components in the preheating section, and the latter also includes the heating section (heat pumps). The energy consumption and cost are represented by KPI_3, as the ratio of the low-cost energy to the total energy entering the system. In this case, the heat-recovery chiller is considered a low-cost energy source, as the main function of this equipment is to deliver cold water to a separate section of the HVAC system. Thus, the water used for internal refrigeration of the chiller requires no additional energy consumption to reach temperatures beneficial to the hot-water load. A higher KPI_3 represents delivery of a larger amount of energy from the preheating section, reducing the need for the traditional heat pumps and reducing the overall energy consumption in the system. KPI_4 quantifies the use of the traditional heat sources, i.e., the heat-recovery chiller and heat pumps. Finally, KPI_5 represents the number of time periods in which the design temperature is satisfied by the system. It is the sum of a binary function based on the outlet temperature of the tempering valve. The designed outlet temperature is 45°C; however, fluctuations of up to 5°C may naturally occur within the system and do not require further action from the heat sources. Yet, the penalization must be asymmetrical, as hot water at 50°C is not desired, for safety reasons. Thus, the safety threshold is set in the range of 40–45°C.

$$KPI_1 = \frac{E}{CH + E} \tag{8}$$

$$KPI_2 = \frac{E}{CH + E + HP} \tag{9}$$

$$KPI_3 = \frac{E + CH}{E + CH + HP} \tag{10}$$

$$KPI_4 = CH + HP \tag{11}$$

$$KPI_5 = \frac{\sum f_t}{\sum t};\quad f_t = \begin{cases} 1 & if \quad T_v \in [40\ ^\circ\text{C},\ 45\ ^\circ\text{C}]\ during\ time\ t \\ 0 & if \quad not \end{cases} \tag{12}$$

In Eqs. (8)–(12) E represents the energy gain of the solar field, CH represents the electrical power required by the chiller, HP represents the rate at which energy is delivered to the stream by the heat pumps (in kJ/h), and T_V represents the outlet temperature of the tempering valve (in °C).

The KPIs are calculated during each observation and assessment window after a certain action has been taken, yielding a matrix of [timesteps, interactions, KPI values] = [20,3,2] for all possible state–action combinations. Then, the mean difference between the values $(t + 1, t)$is used to represent the reward of the action on the state of the system. As the overall performance of the system depends on the energy consumption, the participation of the solar field, and the outlet temperatures, these KPIs are combined to form a global KPI_G. However, depending on the authority in charge of the system, the priorities differ. To evaluate the influence of each of the original KPIs, four different combinations are studied, reflecting the three main ideas in each term (the renewable capacity factor, amount of hot water supplied, and energy consumption). These are defined in Eqs. (13)–(16) according to the values of the weights α, β, γ and the KPI, subject to the restrictions of Eq. (17):

$$\text{Case1}\quad KPI_{G,1} = \alpha \cdot KPI_1 + \beta \cdot KPI_5 + \gamma \cdot KPI_3 \tag{13}$$

$$\text{Case2}\quad KPI_{G,2} = \alpha \cdot KPI_2 + \beta \cdot KPI_5 + \gamma \cdot KPI_3 \tag{14}$$

$$\text{Case3}\quad KPI_{G,3} = \alpha \cdot KPI_1 + \beta \cdot (KPI_5 - KPI_4) + \gamma \cdot KPI_3 \tag{15}$$

$$\text{Case4}\quad KPI_{G,4} = \alpha \cdot KPI_2 + \beta \cdot (KPI_5 - KPI_4) + \gamma \cdot KPI_3 \tag{16}$$

$$\begin{aligned} &\alpha + \beta + \gamma = 1 \\ &\alpha, \beta \in [0, 1] \\ &\gamma = 1 - \alpha - \beta \\ &\gamma > 0 \end{aligned} \tag{17}$$

The difference between KPI_1 and KPI_2 is that the latter also considers the effect of the heat pumps. Thus, $KPI_{G,1}$ and $KPI_{G,2}$ produce similar results, but the latter is representative of the whole system (not only the preheating section). Similarly, KPI_4 and KPI_5 are both related to the energy consumption. Combining these two KPIs adds a penalty term specific to the use of traditional sources when hot water is supplied at the desired temperatures, which gives priority to the solar field's participation. Thus, the similarity in the results provided by $KPI_{G,3}$ and $KPI_{G,4}$ is expected. According to the foregoing definition, a higher score of KPI_G is obtained with higher values of KPI_1, KPI_2, KPI_3, and KPI_5, while the cost ofKPI_4 is minimized. For each admissible (action, state) pair, the instantaneous reward function is expressed as follows:

$$r_t(s, a) = KPI_{G,t}(s, a) \tag{18}$$

Thus, the total reward for each daily scenario k,which is composed of the instantaneous observed rewards r_i^{obs} at each interaction period i obtained from the combination of visited states s^{vis} and selected actions a^{sel}, is defined as

$$R_k = \sum_i^3 r_i^{obs}(s^{vis}, a^{sel}). \tag{19}$$

4.5. Q-table with action-dependent states

A Q-table approach is used to map the corresponding combinations of available actions and viable states of the system. After the training process, each Q-value in the table represents the worth of pursuing an action in a certain state of the system. As the permissible states are action-dependent, there are Q-values corresponding to inadmissible (state, action) pairs that are not updated with the reward function in Eq. (18).

An asymmetric [12 × 32] Q-table is used to present the combination of possible states and actions. The rows, as well as the first 12 columns, represent the possible intermediate states. The remaining 20 columns represent the final absorbing states or daily outcomes. Each state is defined by a tuple that contains the control signal for the solar and chiller heat rejection recirculation pumps: {1,1}, {1,0}, and {0,1} (as defined in previous sections). A schematic of the daily decision process is presented in Fig. 6. Here, the dependency of the actions regarding the states can be observed, as the actions branch out to the future permissible action–states, as described in Section 4.3. For instance, if the first action selected corresponds to {1}, the second action can only be selected from options {4, 5, 6}. This is a representation of the possible state sequence combinations presented in Table 3.

4.6. Q-learning algorithm implementation

The implemented Q-learning algorithm comprises the following steps, as shown in Fig. 7.

(1) Initialization of the Q-Table. As the reward function r_t can obtain zero values for certain (state, action) combinations, the table is initialized with –1.
(2) Selection of the daily initial state. This corresponds to {1,1}, i.e., a state where the solar field and the chiller circulation pumps are activated at 7:00 AM.
(3) At each interaction opportunity between the agent and the environment, depending on the present state and time, the possible actions and admissible future states must be defined.
(4) The agent selects the action with the highest Q-value. If all the Q-values are equal, as for the first daily decision, the selection is performed randomly. The optimal policy is to select the action yielding the highest return for each state, which corresponds to the maximum available Q-value.
(5) After selecting an action, the new current state must be defined, and the corresponding reward must be calculated, representing the change between the present and previous states of the system.
(6) Finally, the corresponding Q-value of the selected (state, action) tuple must be updated with the reward function and the approximation of the Q-function presented in Eq. (4).
(7) Steps 3–6 are repeated for the selected number of training iterations. In the proposed setting, 1000 iterations are conducted.

A sensitivity analysis is performed for the weights α, β, and γ related to each KPI. Analysis of the sensitivity of the agent's decisions to the weights' values reveals that the selection of the optimal policy depends on the system's management priorities regarding energy efficiency and user comfort. With a higher value ofα, the agent's decision is biased toward KPI_1; thus, the renewable capacity factor in the preheating section impacts the final r_k score more significantly than the other KPIs. This allows an additional analysis based on the preference of the authority in charge of the system's management, e.g., energy efficiency, reduced economic costs, or user satisfaction. The detailed process in Fig. 7 is repeated for all the combinations of the policy weights α, β, and γ defined by Eq. (17), resulting in 36 different scenarios. In the next section, the results are discussed, and the paths selected by the agent under different meteorological conditions and reward functions, as well as the resulting scores, are analyzed.

For clarity, an overall flow diagram of the framework's implementation is presented in Fig. 8, summarizing the details of this section. This figure shows how the TRNSYS simulation data are incorporated into the defined states, actions, and rewards of the Q-learning algorithm, as well as the calculation of the rewards for the different scenarios presented.

5. Results and discussion

This section presents the results of the agent–environment interaction simulation and reward calculation based on different meteorological scenarios and overall cost functions. Importantly, the objective of this analysis was to highlight the simplicity and flexibility that RL introduces for a multi-objective optimization problem such as scheduling the operation of components based entirely on the operational conditions of the complete system.

The overall scores R_k were analyzed for the 36 possible combinations per meteorological scenario. For this, a sensitivity analysis of α, β, and γ was performed in determining the agent's decisions throughout the simulated interactions. The weight of each factor influenced the

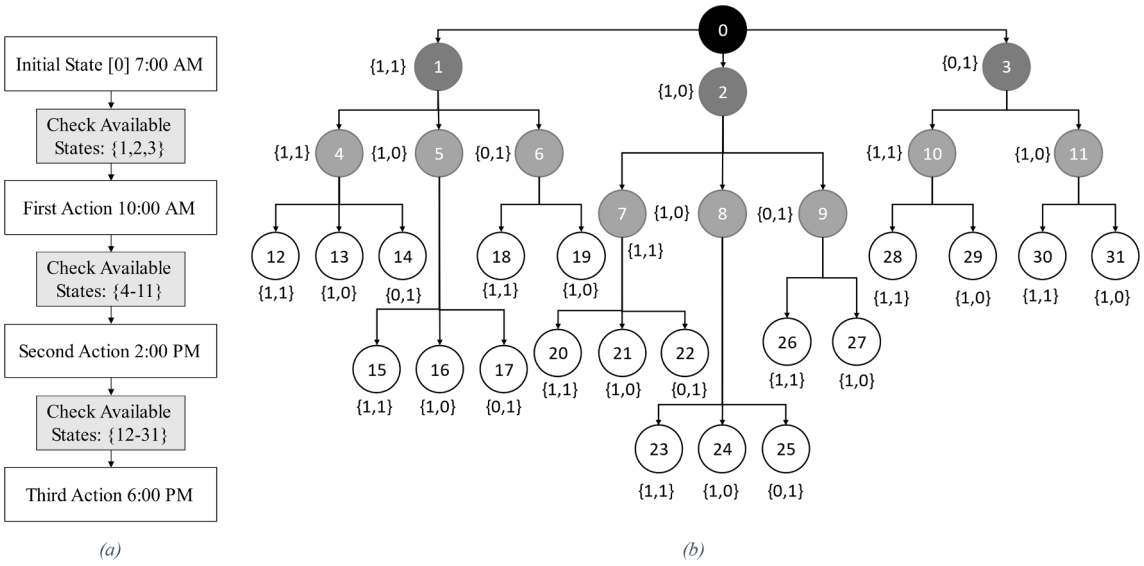

Fig. 6. (a) Agent–environment interaction and (b) action-dependent state space.

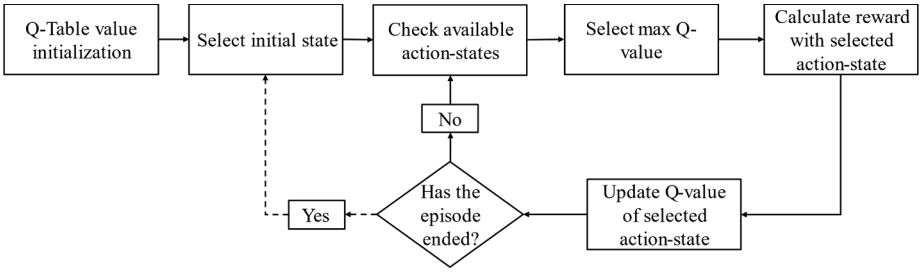

Fig. 7. Flow diagram of the implemented Q-learning algorithm.

participation of the renewable energy source (the solar thermal system), the heat-load supply, and the operational costs of the system. Thus, α was denoted the "renewable factor," β was denoted as the "supply factor," and γ was denoted as the "cost factor." An analysis is presented for the cases studied for the months of January and July, comparing opposite meteorological conditions and how these affected the agent's decisions. As shown in Table 3, each path sequence describes a combination of three actions per day. A weight sensitivity analysis is presented for the maximum R_k scores and selected paths and the frequency of selected actions. Furthermore, the maximum scores per path and the optimal weight values for all the studied cases are presented in Tables 4 and 5. The maximum scores and corresponding paths for each case and month are highlighted in bold font in both tables.

Table 4
Most frequently selected path for different cases and months.

Cases	Month	Alpha	Beta	Gamma	Max. Score	Av. Score	Path
Case 1	January	0.1	0.3	0.6	1.78	1.62	9
	April	0.1	0.1	0.8	1.75	1.56	9
	July	0.1	0.2	0.7	**1.93**	**1.66**	**11**
	October	0.1	0.2	0.7	1.82	1.64	9
Case 2	January	0.1	0.1	0.8	1.98	1.73	1
	April	0.1	0.1	0.8	1.74	1.56	9
	July	0.1	0.1	0.8	**2.04**	**1.79**	**11**
	October	0.1	0.8	0.1	1.51	1.40	12
Case 3	January	0.7	0.1	0.2	**1.62**	**1.17**	**2**
	April	0.6	0.3	0.1	0.81	0.32	14
	July	0.7	0.2	0.1	0.95	0.28	16
	October	0.7	0.2	0.1	1.04	0.45	6
Case 4	January	0.1	0.1	0.8	**1.90**	**1.52**	**1**
	April	0.1	0.1	0.8	1.60	1.39	1
	July	0.2	0.3	0.5	0.87	0.35	13
	October	0.2	0.3	0.5	0.85	0.53	12

5.1. Weight sensitivity score analysis

First, the effects of the four combinations of the weight parameters on the overall scores R_k were examined. Surface graphs are presented to show the permissible and non-permissible action–states determined by the relationships in Eq. (17). As shown in Fig. 9, for January, the global R_k scores obtained with different combinations of α, β, and γ were not significantly sensitive to the different cases, given the higher availability of solar radiation; thus, the performance of the solar field was enhanced. Cases 1 and 2 exhibited a more independent response to the β factor, and Cases 3 and 4 did not exhibit a high sensitivity to changes in α. Higher scores were obtained by prioritizing the γ factor (obtaining low α and β values), expressing the highest ratio of free versus total energy flux in the system. The lowest scores were caused by low participation of the solar field (low α) and a large heat-load supply (high β), which implied high energy consumption.

The difference between the cost-function scenarios was observed more clearly for July, as shown in Fig. 10. These differences were more significant when the available solar radiation was minimized and therefore the performance of the SHW system was reduced. The scores obtained reflect more efficient use of energy resources compared with other months, as high overall R_k scores were obtained (lower maximum scores were obtained for April and October, as shown in Table 5).

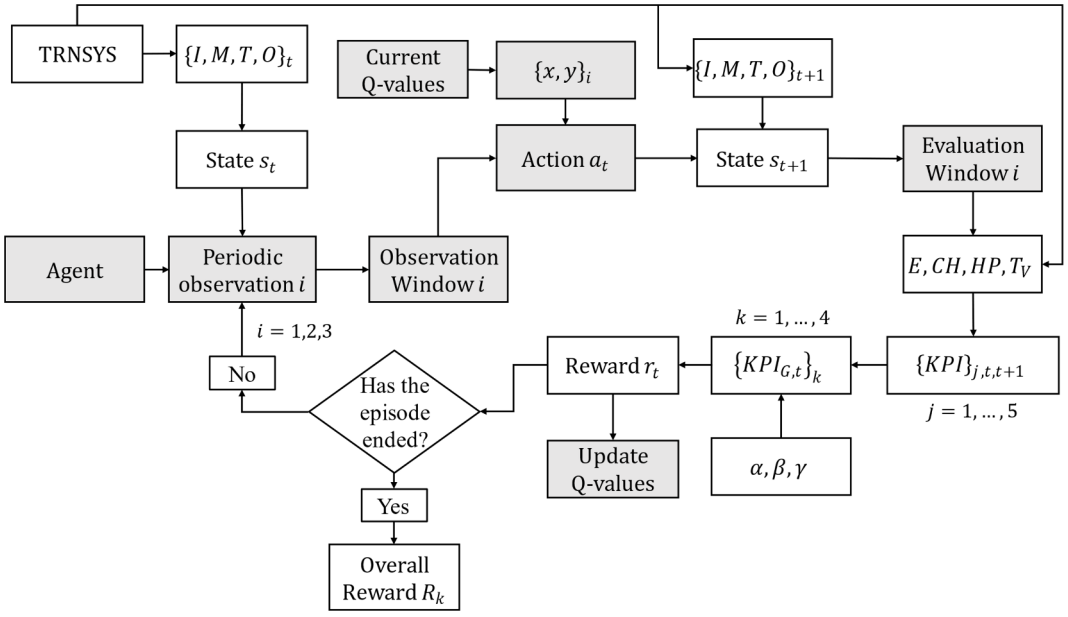

Fig. 8. Flow diagram of the overall framework.

Table 5
Maximum scores for different weight values, selected paths, and cases.

Month	Case	Alpha	Beta	Gamma	Max Score	Path
January	1	0.1	0.1	0.8	1.967	1
	2	0.1	0.1	0.8	**1.984**	**1**
	3	0.1	0.1	0.8	1.878	1
	4	0.1	0.1	0.8	1.895	1
April	1	0.1	0.1	0.8	**1.746**	**9**
	2	0.1	0.1	0.8	1.742	9
	3	0.1	0.1	0.8	1.586	1
	4	0.1	0.1	0.8	1.601	1
July	1	0.1	0.1	0.8	1.928	4
	2	0.1	0.1	0.8	**2.035**	**11**
	3	0.1	0.1	0.8	1.775	3
	4	0.1	0.1	0.8	1.793	3
October	1	0.1	0.2	0.7	1.820	9
	2	0.1	0.1	0.8	**1.822**	**1**
	3	0.1	0.1	0.8	1.690	1
	4	0.1	0.1	0.8	1.705	1

As indicated by Eqs. (14) and (15), cases 3 and 4 were significantly penalized with a high β value, which responded to the effect of KPI_4 (representing increased energy consumption), balancing out a higher KPI_5 (increased heat delivery under design conditions). In principle, these settings reduce unnecessary use of the heat-recovery chiller while penalizing the use of the heat pumps, pushing to deliver the load at 40°C. In contrast, cases 1 and 3 exhibited a significant decrease in the scores in the proximity of $\alpha = 0.5$ at $\beta = 0.1$, while this effect was negligible in cases 2 and 4. This was caused by the difference between KPI_1 and KPI_2; the latter was less sensible to changes in the operation of the heat-recovery chiller, counteracting the effect of the heat pumps. This effect is explained by the operation of the heat pumps, independent of the RL agent. With a reduced participation of the heat-recovery chiller, water temperatures prior to the heating section are frequently lower than desired. This is of relevance under low availability of solar radiation, which results in a larger heat input from conventional sources, compensating for the low heat input from the field to deliver the heat load at the desired temperature (thus increasing the penalty for higher β values).

Owing to the higher availability of solar radiation during January, the system did not struggle to satisfy the demanded heat load. Thus, the value of KPI_5 did not significantly affect the overall score for different values of β in cases 1 and 2. However, in cases 3 and 4, the additional penalization for the use of the heat-recovery chiller and the automatic heat pumps reduced the scores as β increased. This effect was more clearly observed during July, when the increase in β led to lower global KPI scores, as the system struggled to balance meeting the heating load and reducing the operational costs with the limited participation of the solar field. These surface graphs also suggest that the effect of the chiller and heat pumps was greater than that of the solar field, which is consistent with the system's design described in Section 3.1. Additionally, this allows us to quantify the maximum importance assigned to the solar field, which does not affect the system's operation in a significantly negative way. At approximately $\alpha = 0.3$, the overall score reduction is $< 10\%$.

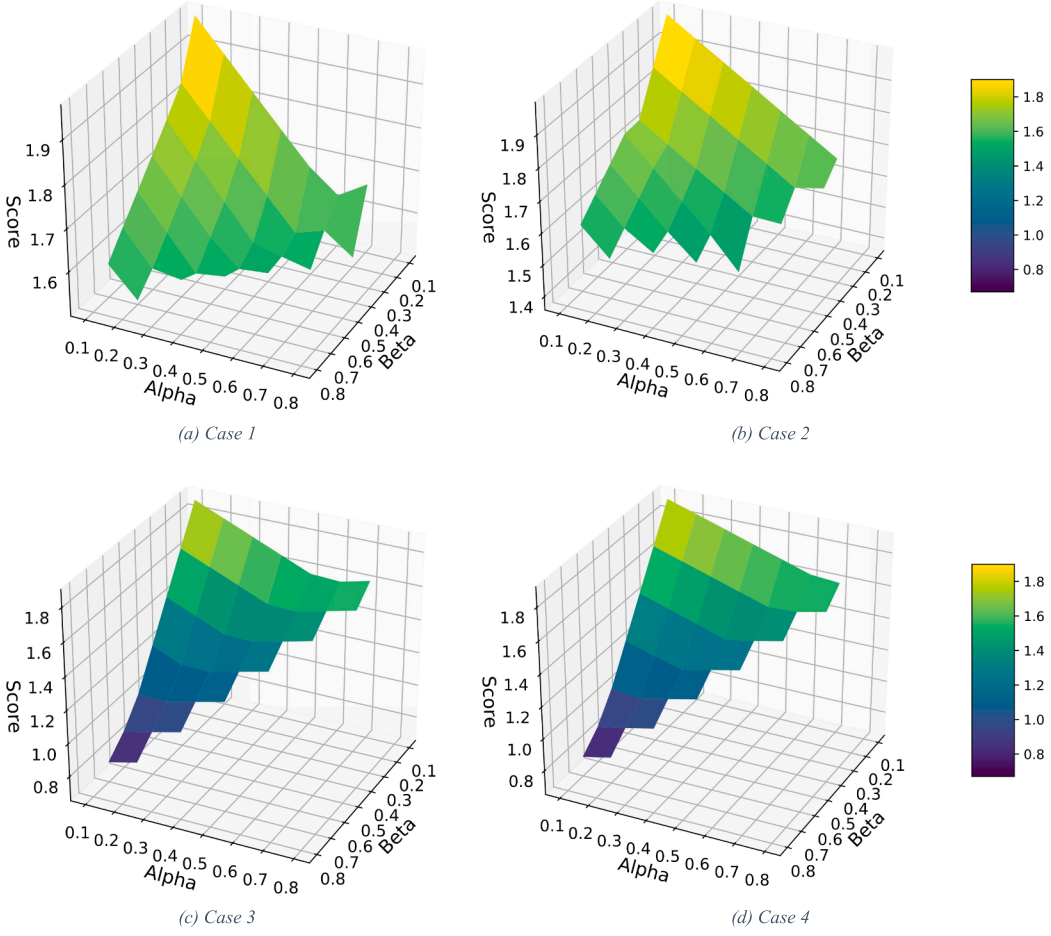

Fig. 9. Scores for January (Cases 1–4).

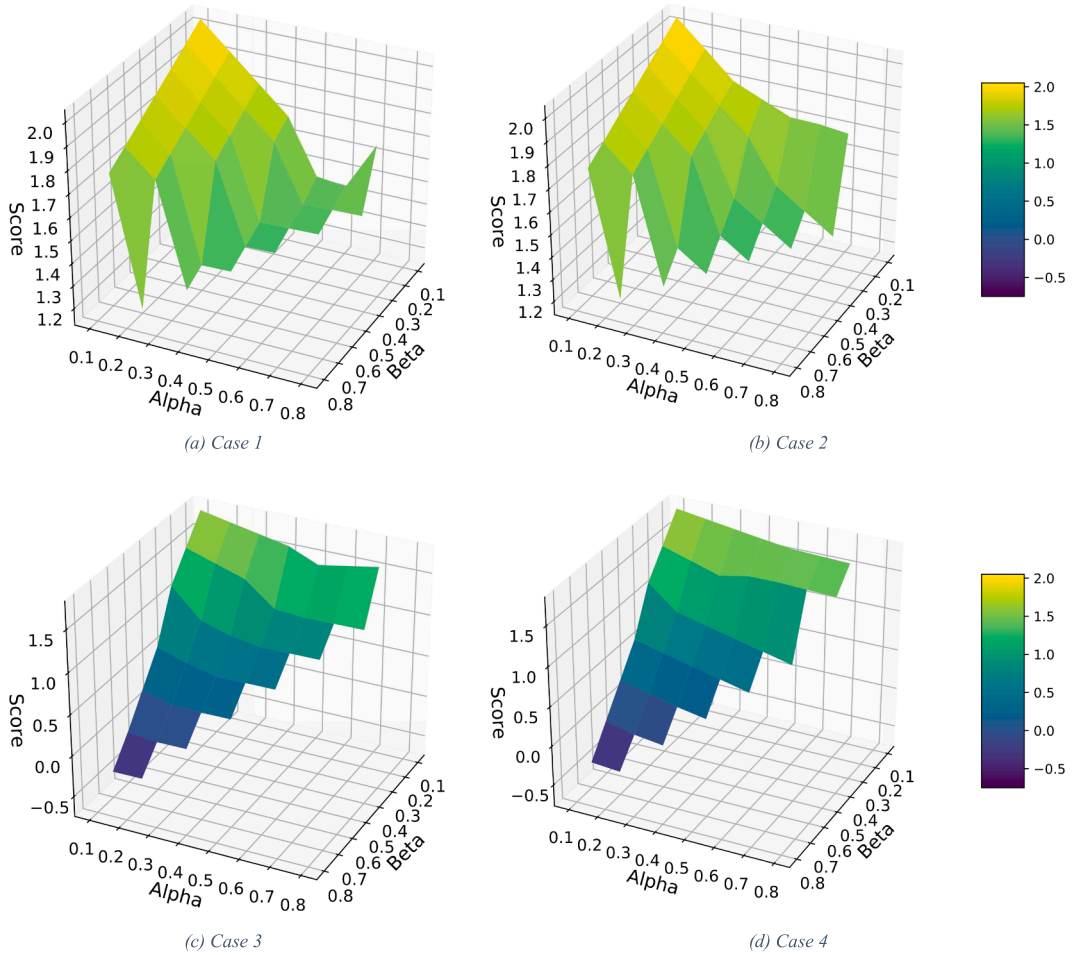

Fig. 10. Scores for July (Cases 1–4).

5.2. Selected path sensitivity to weight value

When analyzing the variability of the agent's decision (i.e., a particular selected path) as a result of the different configurations of the reward function, the relationship between the weights and selected actions can be discussed. As in the previous results, there was an area of permissible and non-permissible paths based on the weight values. This analysis allows to estimate the regions defined by the importance assigned by the values of α, β, and γ. It also revealed the prevalence of certain schedules based primarily on the meteorological conditions. For January, path 1 (nominal conditions, {1,1}-{1,1}-{1,1}) and path 2 (chiller is shut-off during the last interaction window, {1,1}-{1,1}-{1,0}) were present in all the r_t cases. Path 1 was more often selected at low β and low-to-mid αvalues in all the studied cases, while paths 1 and 2 were both dominant in cases 3 and 4, as shown in Fig. 11.

As mentioned previously, cases 3 and 4 reduced the use of the chiller in favor of increasing the use of the solar field. In contrast, for cases 1 and 2, paths 1 ({1,1}-{1,1}-{1,1}) and 9 (chiller was shut-off during the first interaction, {1,0}-{1,1}-{1,1}) were dominant. In these cases, at low α and high β values, the agent was willing to reduce the participation of the chiller during the first period of the day. This effect was greater in case 1 than in case 2, given the replacement of KPI_1 with KPI_2 in the latter, which balanced the heat entering both phases of the system. Furthermore, for similar values of α andβ, in the center part of the figures and towards higher α values, the agent also tends to reduce the participation of the chiller, choosing paths 5, 10, 12, and 14 (the chiller was shut-off during 2/3 of the day in these paths).

For July, different path choices and behavior patterns were observed, as shown in Fig. 12. There was a strong predilection for path 11 ({1,0}-{1,1}-{0,1}) in all the cases. Other paths, such as 3 ({1,1}-{1,1}-{0,1}) and 14 ({1,0}-{1,0}-{0,1}), were present in three of the four cases, limiting the participation of the solar field when the available solar radiation was insufficient to have a beneficial effect on the system. This was also a logical approach of the RL's agent based on the meteorological conditions.

Path 11 was dominant in cases 1 and 2, balancing the participation of the solar field and the chiller at the minimum cost for every value of β. In cases 1 and 2, other investigated alternatives were characterized for the reduction of the chiller's participation, particularly for similar values of $\alpha$and β. Cases 3 and 4 exhibited similar path selections, although case 4 particularly penalized the operation of the chiller, given the influence of KPI_2 (paths 13 {1,0}-{1,0}-{1,0} and 14 {1,0}-{1,0}-{0,1}).

5.3. Maximum scores by most frequent selected policies

An overall view of the paths selected by the agent is presented in this section. From the viewpoint of the operation schedule manager, it is of interest to determine which paths or combinations of actions yield higher scores, independent of which KPI is prioritized. To analyze the effect of the selected paths on the total R_k scores, these maximum scores were selected from all the combinations of weight factors. The most frequently selected paths did not always generate high scores. However, this approach allowed to examine how the different r values affected the agent's decision making.

The results are presented in Table 4. For instance, the path most frequently selected for case 1 was the most insensitive to the meteorological conditions, as among the months, only July exhibited a

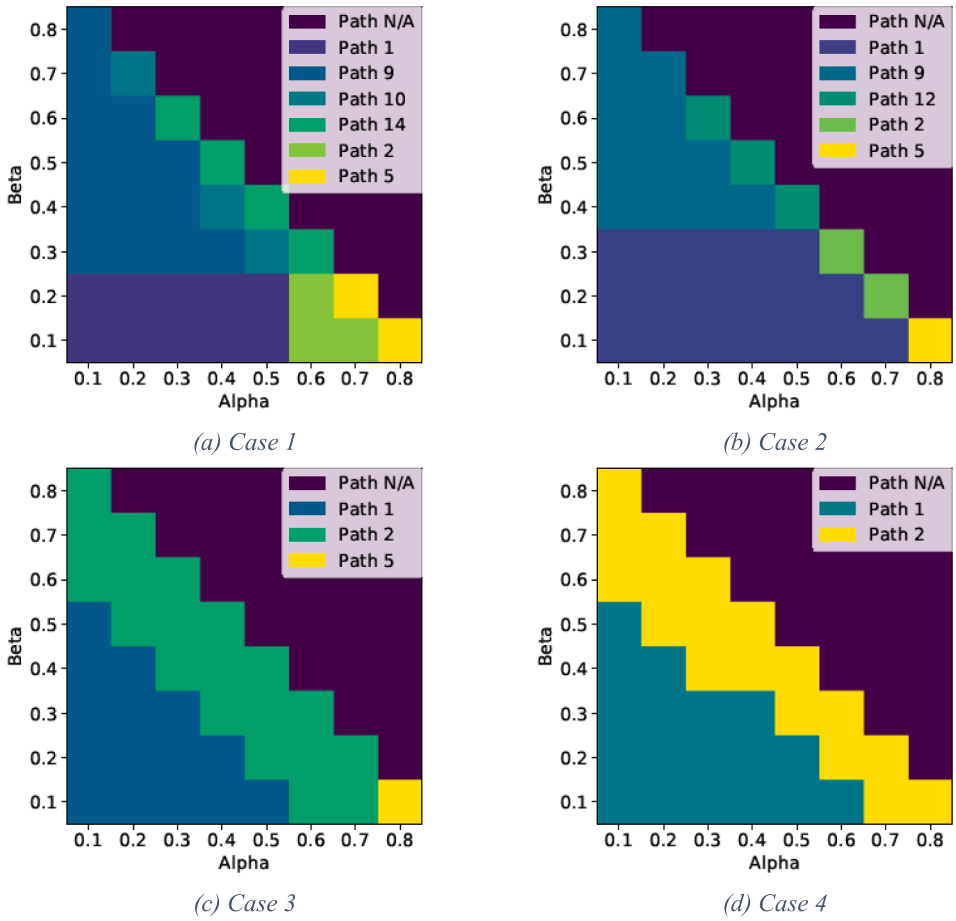

Fig. 11. Selected paths for January (Cases 1–4).

different path. While the agent tended to reduce the participation of the chiller during the first period of the day (path 9 {1,0}-{1,1}-{1,1}), the highest score was unexpectedly achieved in July (path 11 {1,0}-{1,1}-{0,1}), in which both pumps connected to the chiller (first interaction, more solar radiation available) and the solar field (last interaction, less solar radiation available) were shut-off at one point during the day. These high scores obtained in July can be counterintuitive, as lower participation of the solar field reduces the renewable capacity factor and the amount of low-cost energy in the system (reflected by KPI_1 andKPI_3, respectively). However, the ability of the system to reach desired temperatures (KPI_5) balanced out these lower values. Related to this behavior, note that $\alpha = 0.1$ was maintained, and the other factors exhibited minor differences. These results can be interpreted as follows: the energy delivered by the heat-recovery chiller was higher than that delivered by the solar field, and considering the thermal inertia in the system, it could alone satisfy the heat load with little input from the heat pumps. As the KPI_3 values increased (caused by a reduction in the activity of the heat pump), and γ increased, higher global R_k scores were obtained. Case 2 yielded different results for January and October (more solar radiation available); however, the highest score was obtained in July, with path 11 ({1,0}-{1,1}-{0,1}). The weight values selected in case 2 were similar to the combinations that yielded high scores in case 1, as shown in Figs. 9 and 10. The selected paths for April and July exhibited the same behavior that was observed in case 1 (paths 9 {1,0}-{1,1}-{1,1} and 11 {1,0}-{1,1}-{0,1}). However, in comparison to case 1, the path selection in case 2 was more significantly influenced by the operational costs of the chiller, reducing its participation also in October (path 12 {1,0}-{1,0}-{1,1}), even when there was a higher availability of solar radiation.

Case 3 exhibited different results. The highest scores were obtained in January. Additionally, the use of the chiller was highly penalized, even when it was shut off during two periods per day in April and July (paths 14 {1,0}-{1,0}-{0,1} and 16 {1,0}-{0,1}-{1,0}, respectively) and even when little solar radiation was available. This bias can be interpreted as a more “renewable efficient” policy, reflecting the effect of KPI_4. However, compared with the other cases, it produced lower scores overall. Additionally, case 3 was the only scenario where high α values were selected. As case 4 was similar to case 3, January exhibited the highest scores under nominal conditions ({1,1}-{1,1}-{1,1}). For July and October, the aforementioned trend of penalizing the use of the chiller was present, but the point of full shut-down of the chiller operation was reached (path 13 {1,0}-{1,0}-{1,0}). The low scores obtained despite the large γ factor could be due to the high usage of the heat pumps to reach the desired output temperatures in the system to counter the lower temperatures reached with the solar field. In the present case, this is reflected by both KPI_2 and KPI_4.

5.4. Maximum scores for different selected paths

When analyzing the paths that generated the highest scores for the different cases, the number of viable paths was reduced to five. Table 5 presents the selected paths that generated the highest scores for the different cases and months. The weights were mostly constant at $\alpha = 0.1$, $\beta = 0.1$, $\gamma = 0.8$ throughout the different scenarios, except for case 1 in October.

Regarding the selected paths, those for January was highly regular. Here, the agent chose to operate under nominal conditions throughout the day. April also exhibited simple behavior. Even when the solar radiation was limited, cases 1 and 2 penalized the use of the heat-recovery chiller during the first action window (10 AM) rather than

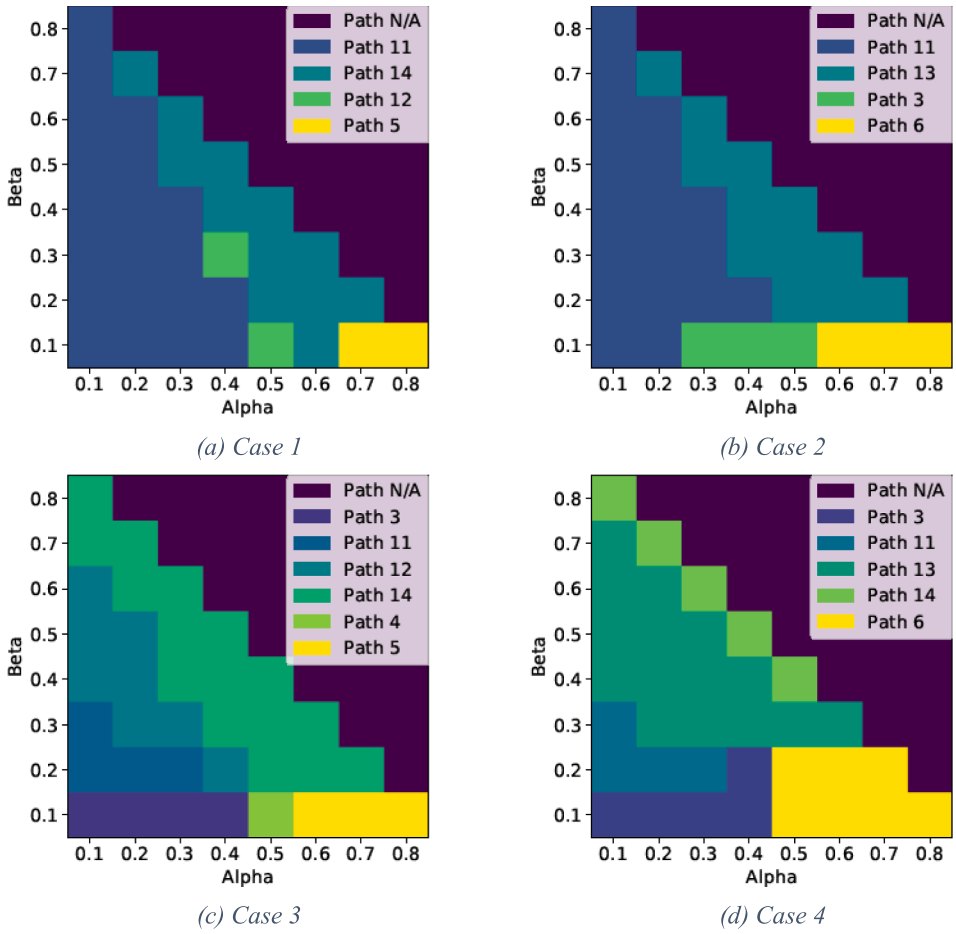

Fig. 12. Selected paths for July (Cases 1–4).

participation of the solar field (path 9 {1,0}-{1,1}-{1,1}). Similarly, cases 3 and 4 returned to nominal conditions rather than limiting the participation of the solar field. July exhibited irregular behavior. Cases 3 and 4 reduced the participation of the solar field during low-radiation hours (after 6 PM, path 3 {1,1}-{1,1}-{0,1}), and case 1 penalized the unnecessary use of the heat-recovery chiller (2 PM, path 4 {1,1}-{1,0}-{1,1}). For case 2, path 11 ({1,0}-{1,1}-{0,1}) had the highest score, and it was also the most frequent choice for July in cases 1 and 2 (see Table 4). Finally, October exhibited both the nominal path 1 ({1,1}-{1,1}-{1,1}) and path number 9 (1,0}-{1,1}-{1,1}). Nevertheless, case 1 for October was the only scenario where the optimal weight configuration resulted in $\alpha = 0.1$, $\beta = 0.2$, and$\gamma = 0.7$.

The maximum scores for the cases were mostly caused by the agent's decision to maintain nominal operation. This is reasonable, considering the overall performance of the solar field, except during July, when the solar radiation is at its minimum level. However, the maximum scores were observed for January and July, representing the most economic path choices given the available resources and heat-load goals. The maximum score for July was 2.035, which represents increases of 3%, 17%, and 12% compared with the maximum scores for January, April, and October, respectively. On average, July's performance exhibits a 14% increase compared to the nominal operation schedules during the other tested months.

The RL agent is able to derive operation schedules which exhibit better performance than under nominal conditions in April and July. For the different values of α, β and γ, maximum scores under nominal conditions (path 1 {1,0}-{1,0}-{1,0}) are presented in Table 6. For January and October these scores and corresponding operation schedules were confirmed by the RL agent (see Table 5). However, global KPI scores increase a 11% and 21% on average for April and July, respectively, as detailed in Table 7. This increase in July is caused by schedules which either shut-off the heat-recovery chiller or the solar field at different daily instances, in favor of maximizing heat gain and minimizing costs under unfavorable meteorological conditions.

The sensitivity analysis revealed that this method can be applied regardless of the meteorological conditions, yielding consistent configurations of weight values (α, β, γ) for the maximum scores. Regardless of the selected weights, the highest performance of the system was achieved in January and July, corresponding to the maximum and minimum availability of solar radiation, respectively (albeit the selected paths varied significantly). The April and October scenarios were less predictable owing to the unstable weather conditions; the agent was forced to compensate for the effects of conditions changing daily and the thermal inertia of the system. The global KPI choice had the principal effect on the choice of daily action paths. This highlights the flexibility of the method when it is applied to the SHW system.

In this agent–environment interaction setting, the RL agent extracted logical and physically reasonable scheduling operations for the SHW system, improving its ability to adapt to meteorological conditions

Table 6
Maximum Scores obtained under nominal operation.

Max. Scores	Month			
Case	January	April	July	October
1	1.967	1.477	1.648	1.815
2	1.984	1.669	1.700	1.822
3	1.878	1.586	1.564	1.690
4	1.895	1.601	1.326	1.705

Table 7
Increase of maximum scores compared to nominal operation.

Increase %	Month			
Case	January	April	July	October
1	–	18.2%	17.0%	0.3%
2	–	4.4%	19.7%	–
3	–	–	13.5%	–
4	–	–	35.2%	–

and different operating principles. By taking into account the thermal inertia of the system, this method has an advantage over regular thermostat-based controls, as it reduces the frequency of actions and thus enhances the stability of the system. In most cases where solar radiation was widely available at the selected location (October through April), the nominal operational condition frequently yielded the highest performance. However, the operation schedule differed significantly during July, when the solar radiation was at its minimum. Both extreme scenarios—January and July—exhibited the best-performing schedules. A deeper study should be performed to investigate different agent–environment interaction frequencies during the studied time windows.

Methods such as RL allow decision-making situations to be investigated via different approaches. Depending on the level of knowledge and complexity of a system, the environment, states, actions and rewards can be defined in ways specific to the system, to evaluate the time-deferred effects of these actions on the system. Thus, the design and approach of the RL agent is entirely dependent on the studied system, rendering it highly flexible. The present case study revealed how this general methodology can be applied to a fairly complex system that deals with demand profiles, thermal inertia, and varying sources of energy to determine operational schedules according to the energy efficiency, comfort levels, and participation of renewable energy sources. The tabular Q-learning approach is used owing to its simplicity; however, more complex methodologies should be investigated for similar settings. Additionally, the proposed method allows the analysis of a multi-objective optimization problem based entirely on the operational conditions of the complete system. With this data-driven approach, a specific physics model for solving the optimization problem is not required, providing an alternative technique for complex system analysis.

6. Conclusions

A decision-making environment representing a condition-based operational scheduling is proposed. On the basis of the use of an RL agent, optimal operational schedules are derived according to the integration scheme and performance of a hot-water system driven by a renewable energy source. The system is controlled by a heat-recovery chiller and a single-speed centrifugal pump, which is used to circulate water through an evacuated tube solar collector field. Both the pump and the chiller are subjected to the agent's decisions to maximize a reward function based on the energy efficiency and user satisfaction. Through operating-profile samples generated in TRNSYS and subsequent development of the agent–environment interaction space, decisions regarding the operation of the circulation pumps were assessed for three daily instances. Four global KPI were proposed and used as reward functions under different meteorological conditions. Finally, a sensitivity analysis was performed on the agent's priorities: maximize the renewable energy source participation, reduce the operational costs, and increase the supply rate under the established design.

This setting of condition-based scheduling has several advantages over programmed controlled systems. First, as long time windows are considered, there is a lower probability of causing and reacting to unstable control signals, which can potentially damage the circulation pumps and affect the delivery of hot water under the design conditions. As in many dynamic systems, different operation scheduling plans have long-term effects on the future states of the system and thus require an approach that can encompass the delayed thermal responses throughout the system. This time difference between the action and the reward is a crucial component in the RL-based framework. Second, by replacing the temperature-based controls and taking into account the system's current performance, efficient schedules can be identified according to the established global KPI configuration. Third, this setting allows the exploration of different policies depending on the prioritized criteria represented by the global KPI reward function. In this case, while the nominal operational conditions exhibited the highest performance when solar radiation was available, new beneficial alternative schedules were identified for July.

Although the proposed RL-based solar thermal system operation scheduling framework is a promising approach, further development and experimental validation are needed to assess its advantages. Furthermore, given the available data, the agent–environment setting, which is coupled to the condition-based KPI representing the reward function, is a straightforward tool. This approach does not require a large amount of computational resources to simulate action–reward sequences and can reveal or confirm beneficial scheduling alternatives for different systems.

CRediT authorship contribution statement

Camila Correa-Jullian: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. **Enrique López Droguett:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. **José Miguel Cardemil:** Conceptualization, Funding acquisition, Resources, Methodology, Software, Supervision, Writing - review & editing.

Declaration of Competing 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.

Acknowledgments

The authors appreciate the support from ANID/FONDAP 15110019 "Solar Energy Research Center" SERC-Chile. The authors acknowledge the partial financial support of the Chilean National Fund for Scientific and Technological Development (FONDECYT) under Grant No. 1190720.

References

[1] Aguilar C, White D, Ryan D. Domestic water heating and water heater energy consumption in Canada. Canadian building energy end-use data and analysis centre. 2005https://sites.ualberta.ca/~cbeedac/publications/documents/domwater_000.pdf.
[2] Sharma AK, Sharma C, Mullick SC, Kandpal TC. Solar industrial process heating: A review. Renew Sustain Energy Rev 2017;78:124–37. https://doi.org/10.1016/j.rser.2017.04.079.
[3] Hossain MS, Saidur R, Fayaz H, Rahim NA, Islam MR, Ahamed JU, et al. Review on solar water heater collector and thermal energy performance of circulating pipe. Renew Sustain Energy Rev 2011;15:3801–12. https://doi.org/10.1016/j.rser.2011.06.008.
[4] Halvgaard R, Bacher P, Perers B, Andersen E, Furbo S, Jørgensen JB, et al. Model predictive control for a smart solar tank based on weather and consumption forecasts. Energy Procedia 2012;30:270–8. https://doi.org/10.1016/j.egypro.2012.11.032.
[5] Herrera E, Bourdais R, Guéguen H. A hybrid predictive control approach for the management of an energy production-consumption system applied to a TRNSYS solar absorption cooling system for thermal comfort in buildings. Energy Build 2015;104:47–56. https://doi.org/10.1016/j.enbuild.2015.06.076.
[6] Nhut LM, Park YC. A study on automatic optimal operation of a pump for solar domestic hot water system. Sol Energy 2013;98:448–57. https://doi.org/10.1016/j.

solener.2013.08.040.
[7] Badescu V. Optimal control of flow in solar collector systems with fully mixed water storage tanks. Energy Convers Manage 2008;49:169–84. https://doi.org/10.1016/j.enconman.2007.06.022.
[8] Garcia-Gabin W, Zambrano D, Camacho EF. Sliding mode predictive control of a solar air conditioning plant. Control Eng Pract 2009;17:652–63. https://doi.org/10.1016/j.conengprac.2008.10.015.
[9] Zhang Y, Yang R, Zhang J, Weng Y, Hodge BM. Predictive analytics for comprehensive energy systems state estimation. Big data application in power systems Elsevier Inc; 2017. p. 343–76. https://doi.org/10.1016/B978-0-12-811968-6.00016-4.
[10] Wang J, Ma Y, Zhang L, Gao RX, Wu D. Deep learning for smart manufacturing: Methods and applications. J Manuf Syst 2018;48:144–56. https://doi.org/10.1016/j.jmsy.2018.01.003.
[11] Sharma A, Kakkar A. Forecasting daily global solar irradiance generation using machine learning. Renew Sustain Energy Rev 2018;82:2254–69. https://doi.org/10.1016/j.rser.2017.08.066.
[12] Khan S, Yairi T. A review on the application of deep learning in system health management. Mech Syst Signal Process 2018;107:241–65. https://doi.org/10.1016/j.ymssp.2017.11.024.
[13] Waschneck B, Reichstaller A, Belzner L, Altenmüller T, Bauernhansl T, Knapp A, et al. Optimization of global production scheduling with deep reinforcement learning. Procedia CIRP, vol. 72, Elsevier; 2018, p. 1264–9. Doi: 10.1016/j.procir.2018.03.212.
[14] Xanthopoulos AS, Kiatipis A, Koulouriotis DE, Stieger S. Reinforcement learning-based and parametric production-maintenance control policies for a deteriorating manufacturing system. IEEE Access 2017;6:576–88. https://doi.org/10.1109/ACCESS.2017.2771827.
[15] Rocchetta R, Bellani L, Compare M, Zio E, Patelli E. A reinforcement learning framework for optimal operation and maintenance of power grids. Appl Energy 2019;241:291–301. https://doi.org/10.1016/j.apenergy.2019.03.027.
[16] Lu R, Hong SH, Zhang X. A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. Appl Energy 2018;220:220–30. https://doi.org/10.1016/j.apenergy.2018.03.072.
[17] Wang P, Chan C-Y. Formulation of deep reinforcement learning architecture toward autonomous driving for on-ramp merge. 2017 IEEE 20th conference on intelligent transportation systems, proceedings ITSC, IEEE; 2017. p. 1–6. https://doi.org/10.1109/ITSC.2017.8317735.
[18] Kuznetsova E, Li YF, Ruiz C, Zio E, Ault G, Bell K. Reinforcement learning for microgrid energy management. Energy 2013;59:133–46. https://doi.org/10.1016/j.energy.2013.05.060.
[19] Levent T, Preux P, Henri G. Energy Management for Microgrids: a Reinforcement Learning Approach. IEEE PES Innovative Smart Grid Technologies Europe, ISGT-Europe 2019;2019:1–5.
[20] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl Energy 2019;235:1072–89. https://doi.org/10.1016/j.apenergy.2018.11.002.
[21] Chen Y, Norford LK, Samuelson HW, Malkawi A. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. Energy Build 2018;169:195–205. https://doi.org/10.1016/J.ENBUILD.2018.03.051.
[22] Ruelens F, Claessens BJ, Quaiyum S, De Schutter B, Babuška R, Belmans R. Reinforcement learning applied to an electric water heater: from theory to practice. IEEE Trans Smart Grid 2018;9:3792–800. https://doi.org/10.1109/TSG.2016.2640184.
[23] Al-Jabery K, Xu Z, Yu W, Wunsch DC, Xiong J, Shi Y. Demand-side management of domestic electric water heaters using approximate dynamic programming. IEEE Trans Comput Des Integr Circuits Syst 2017;36:775–88. https://doi.org/10.1109/TCAD.2016.2598563.
[24] Kazmi H, Mehmood F, Lodeweyckx S, Driesen J. Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems. Energy 2018;144:159–68. https://doi.org/10.1016/j.energy.2017.12.019.
[25] Qi X, Luo Y, Wu G, Boriboonsomsin K, Barth M. Deep reinforcement learning enabled self-learning control for energy efficient driving. Transp Res Part C Emerg Technol 2019;99:67–81. https://doi.org/10.1016/J.TRC.2018.12.018.
[26] Lu R, Hong SH. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. Appl Energy 2019;236:937–49. https://doi.org/10.1016/j.apenergy.2018.12.061.
[27] Park JY, Dougherty T, Fritz H, Nagy Z. LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. Build Environ 2019;147:397–414. https://doi.org/10.1016/j.buildenv.2018.10.028.
[28] Yang L, Nagy Z, Goffin P, Schlueter A. Reinforcement learning for optimal control of low exergy buildings. Appl Energy 2015;156:577–86. https://doi.org/10.1016/j.apenergy.2015.07.050.
[29] Ruelens F, Claessens BJ, Vandael S, Iacovella S, Vingerhoets P, Belmans R. Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning. Proceedings – 2014 power systems computation conference, PSCC 2014 2014. https://doi.org/10.1109/PSCC.2014.7038106.
[30] De Somer O, Soares A, Vanthournout K, Spiessens F, Kuijpers T, Vossen K. Using reinforcement learning for demand response of domestic hot water buffers: a real-life demonstration. 2017 EEE PES innovative smart grid technologies conference Europe, ISGT-Europe 2017 – Proceedings vol. 2018. Janua, IEEE; 2017. p. 1–7. https://doi.org/10.1109/ISGTEurope.2017.8260152.
[31] Watkins CJCH. Learning from delayed rewards. PhD Thesis, Univ Cambrigde, England 1989; 15: 233–5.
[32] Singh SP, Sutton RS. Reinforcement learning with replacing eligibility traces. Mach Learn 1996;22:123–58. https://doi.org/10.1007/BF00114726.
[33] Claessens BJ, Vrancx P, Ruelens F. Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. IEEE Trans Smart Grid 2018;9:3259–69. https://doi.org/10.1109/TSG.2016.2629450.
[34] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. Nature 2013;518:529–33. https://doi.org/10.1038/nature14236.
[35] Sutton RS, Barto AG. Reinforcement learning: an introduction. IEEE Trans Neural Networks 1998;9. https://doi.org/10.1109/TNN.1998.712192.
[36] Barde SRA, Yacout S, Shin H. Optimal preventive maintenance policy based on reinforcement learning of a fleet of military trucks. J Intell Manuf 2019;30:147–61. https://doi.org/10.1007/s10845-016-1237-7.
[37] Wiering M, Martijn van O. Reinforcement Learning vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. https://doi.org/10.1007/978-3-642-27645-3.
[38] Panait L, Luke S. Cooperative multi-agent learning: the state of the art. Auton Agent Multi Agent Syst 2005;11:387–434. https://doi.org/10.1007/s10458-005-2631-2.
[39] Ge TS, Wang RZ, Xu ZY, Pan QW, Du S, Chen XM, et al. Solar heating and cooling: Present and future development. Renew Energy 2018;126:1126–40. https://doi.org/10.1016/j.renene.2017.06.081.
[40] Ntsaluba S, Zhu B, Xia X. Optimal flow control of a forced circulation solar water heating system with energy storage units and connecting pipes. Renew Energy 2016;89:108–24. https://doi.org/10.1016/j.renene.2015.11.047.
[41] Weiss W, Spörk-Dür M. Solar Heat Worldwide 2018. Global Market Development and Trends in 2017. Detailed Market Figures. IEA Sol Heat Cool Program 2016;2018:94https://www.iea-shc.org/Data/Sites/1/publications/Solar-Heat-Worldwide-2018.pdf.
[42] Baljit SSS, Chan HY, Sopian K. Review of building integrated applications of photovoltaic and solar thermal systems. J Clean Prod 2016;137:677–89. https://doi.org/10.1016/j.jclepro.2016.07.150.
[43] U.S. Energy Information Administration (EIA). Assumptions to the Annual Energy Outlook 2015 - Oil and Gas Supply Module. Washington DC; 2015.
[44] Shafieian A, Khiadani M, Nosrati A. A review of latest developments, progress, and applications of heat pipe solar collectors. Renew Sustain Energy Rev 2018;95:273–304. https://doi.org/10.1016/j.rser.2018.07.014.
[45] Zambolin E, Del Col D. Experimental analysis of thermal performance of flat plate and evacuated tube solar collectors in stationary standard and daily conditions. Sol Energy 2010;84:1382–96. https://doi.org/10.1016/j.solener.2010.04.020.
[46] Ayompe LM, Duffy A, Mc Keever M, Conlon M, McCormack SJ. Comparative field performance study of flat plate and heat pipe evacuated tube collectors (ETCs) for domestic water heating systems in a temperate climate. Energy 2011;36:3370–8. https://doi.org/10.1016/j.energy.2011.03.034.
[47] Correa-Jullian C, Cardemil JM, Droguett EL, Behzad M. Assessment of Deep Learning techniques for Prognosis of solar thermal systems. Renew Energy 2020;145:2178–91. https://doi.org/10.1016/j.renene.2019.07.100.
[48] Duffie JA, Beckman WA, McGowan J. Solar engineering of thermal processes. Am J Phys 1985;53(4). https://doi.org/10.1119/1.14178. 382–382.
[49] Kalogirou SA. Solar energy engineering processes and systems. Second. Elsevier; 2009. https://doi.org/10.1016/B978-0-12-374501-9.00014-5.
[50] Klein SA. TRNSYS: A transient systems simulation program v.18.00.0019 2018.
[51] Ayompe LM, Duffy A, McCormack SJ, Conlon M. Validated TRNSYS model for forced circulation solar water heating systems with flat plate and heat pipe evacuated tube collectors. Appl Therm Eng 2011;31:1536–42. https://doi.org/10.1016/j.applthermaleng.2011.01.046.
[52] Ruiz E, Martínez JP. Analysis of an open-air swimming pool solar heating system by using an experimentally validated TRNSYS model. Sol Energy 2010;84:116–23. https://doi.org/10.1016/j.solener.2009.10.015.
[53] Bava F, Furbo S. Development and validation of a detailed TRNSYS-Matlab model for large solar collector fields for district heating applications. Energy 2017;135:698–708. https://doi.org/10.1016/j.energy.2017.06.146.
[54] Chen JFF, Dai YJJ, Wang RZZ. Experimental and analytical study on an air-cooled single effect LiBr-H2O absorption chiller driven by evacuated glass tube solar collector for cooling application in residential buildings. Sol Energy 2017;151:110–8. https://doi.org/10.1016/J.SOLENER.2017.05.029.
[55] Kalogirou SA, Agathokleous R, Barone G, Buonomano A, Forzano C, Palombo A. Development and validation of a new TRNSYS Type for thermosiphon flat-plate solar thermal collectors: energy and economic optimization for hot water production in different climates. Renew Energy 2019;136:632–44. https://doi.org/10.1016/j.renene.2018.12.086.
[56] Burch J, Christensen C. Towards development of an algorithm for mains water temperature. InterSolar 2007 Conf 2007: 5–10.
[57] Aissani N, Beldjilali B, Trentesaux D. Dynamic scheduling of maintenance tasks in the petroleum industry: A reinforcement approach. Eng Appl Artif Intell 2009;22:1089–103. https://doi.org/10.1016/j.engappai.2009.01.014.
[58] Kuhnle A, Jakubik J, Lanza G. Reinforcement learning for opportunistic maintenance optimization. Prod Eng 2019;13:33–41. https://doi.org/10.1007/s11740-018-0855-7.
[59] Compare M, Bellani L, Cobelli E, Zio E. Reinforcement learning-based flow management of gas turbine parts under stochastic failures. Int J Adv Manuf Technol 2018;99:2981–92. https://doi.org/10.1007/s00170-018-2690-6.
[60] Wang X, Wang H, Qi C. Multi-agent reinforcement learning based maintenance policy for a resource constrained flow line system. J Intell Manuf 2016;27:325–33. https://doi.org/10.1007/s10845-013-0864-5.