Applied Energy 318 (2022) 119206

Contents lists available at ScienceDirect

Applied Energy

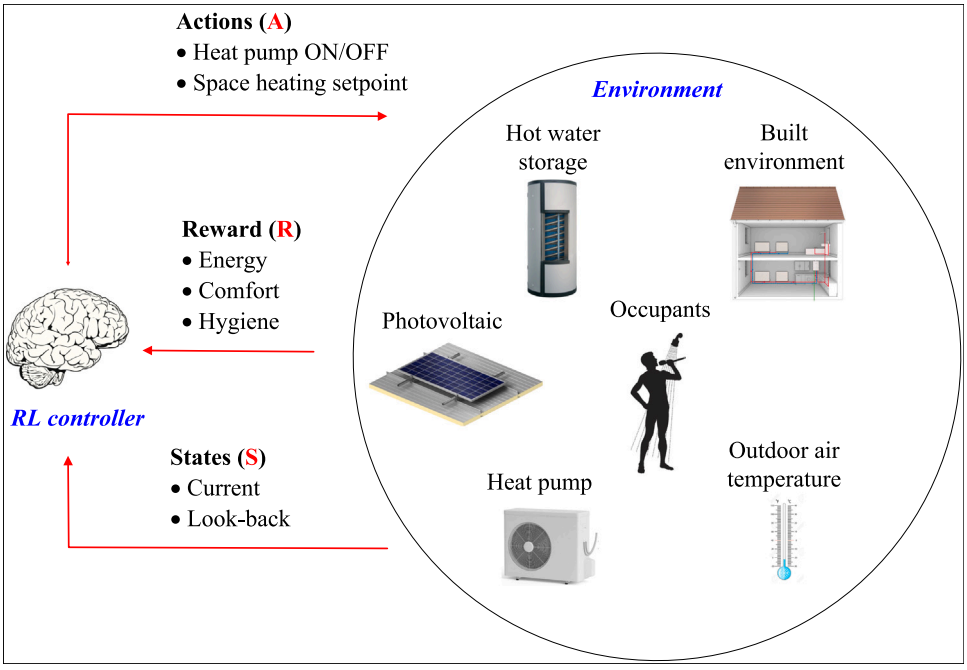
journal homepage: www.elsevier.com/locate/apenergy

Reinforcement Learning for proactive operation of residential energy systems by learning stochastic occupant behavior and fluctuating solar energy: Balancing comfort, hygiene and energy use

Amirreza Heidari [a,b,*], François Maréchal [b], Dolaana Khovalyg [a]

[a] Ecole Polytechnique Fédérale de Lausanne (EPFL), Laboratory of Integrated Comfort Engineering (ICE), CH-1700 Fribourg, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Laboratory of Industrial Process and Energy Systems Engineering (IPESE), CH-1951 Sion, Switzerland

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Reinforcement Learning
Space heating
Solar
Building
Control
Occupant behavior

ABSTRACT

When it comes to residential buildings, there are several stochastic parameters, such as renewable energy production, outdoor air conditions, and occupants' behavior, that are hard to model and predict accurately, with some being unique in each specific building. This increases the complexity of developing a generalizable optimal control method that can be transferred to different buildings. Rather than hard-programming human knowledge into the controller (in terms of rules or models), a learning ability can be provided to the controller such that over the time it can learn by itself how to maintain an optimal operation in each specific building. This research proposes a model-free control framework based on Reinforcement Learning

* Correspondence to: Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
E-mail addresses: amirreza.heidari@epfl.ch (A. Heidari), francois.marechal@epfl.ch (F. Maréchal), dolaana.khovalyg@epfl.ch (D. Khovalyg).

https://doi.org/10.1016/j.apenergy.2022.119206
Received 11 February 2022; Received in revised form 23 March 2022; Accepted 23 April 2022
Available online 10 May 2022

that takes into account the stochastic hot water use behavior of occupants, solar power generation, and weather conditions, and learns how to make a balance between the energy use, occupant comfort and water hygiene in a solar-assisted space heating and hot water production system. A stochastic-based offline training procedure is proposed to give a prior experience to the agent in a safe simulation environment, and further ensure occupants comfort and health when the algorithm starts online learning on the real house. To make a realistic assessment without interrupting the occupants, weather conditions and hot water use behavior are experimentally monitored in three case studies in different regions of Switzerland, and the collected data are used in simulations to evaluate the proposed control framework against two rule-based methods. Results indicate that the proposed framework could achieve an energy saving from 7% to 60%, mainly by adapting to solar power generation, without violating comfort or compromising the health of occupants.

1. Introduction

Occupant behavior is a major driver of energy use in buildings [1]. Occupants influence the building energy use by their presence, activation, or dis-activation of energy devices and adjustment of desired setpoints [2]. The role of occupant behavior is specifically important for indoor conditioning, and hot water production systems [3]. Occupant behavior is considered as a major source of uncertainty for optimal operation of building energy systems [4]. Modeling the occupant behavior may, therefore, help to better understand and integrate it into the control of energy systems in buildings [5,6]. However, occupant behavior can be affected by many different parameters, including environment-related, time-related, and random factors, which makes it extremely stochastic and complex [7,8]. Even when an advanced modeling method is developed to predict occupant behavior, it is challenging to quickly apply that model to a similar, but distinct building [4]. Consequently, it is challenging to develop a holistic and transferable model of occupant behavior to be used in the design phase of buildings without any prior data of that specific occupants. Current control of building energy systems are, therefore, detached from occupant behavior and follow a conservative and energy-intensive approach.

Besides occupant behavior, integration of renewable energy sources to the buildings forms another source of uncertainty for their optimal operation. The share of renewable energy sources in the building sector is projected to be doubled by 2030 [9]. While this increasing share would reduce CO2 emissions, the fluctuating and stochastic nature of renewable energy sources increases the complexity of optimal control. Due to the intermittent nature of renewable energy sources, injecting the surplus power into the grid also complicates the grid operation and can pose problems (e.g. voltage fluctuation) [10]. One way to cope with the fluctuating supply is to make the local electricity demand flexible and responsive to the supply, aiming to maximize the self-consumption and to ensure the balance in the grid [11]. Demand flexibility can be provided through several methods, such as flexible thermal generators, electrical or thermal energy storage, demand-side measures, or even grid-connected electric vehicles [12]. Among these options, storing the surplus energy as heat (power-to-heat) is considered to be particularly promising because both the cost of generating heat from electricity and the cost of heat storage are relatively low [13]. Air-to-water heat pumps emerge as a favorable power-to-heat option that can provide a great opportunity for solar energy integration in the building sector. This is because, first of all, the number of heat pumps as an energy-efficient technology is steadily increasing in the building sector. For example, the number of installed heat pumps in Germany has almost doubled over the last 6 years [14]. Secondly, hot water storage of a heat pump is cost-effective energy storage that can provide the same level of self-consumption of electric storage, but at half of the levelized electricity cost [15]. Furthermore, the thermal mass of the building itself can serve as an additional heat storage for heat pumps, making it possible to further increase the flexibility without additional investments [11]. Buildings, therefore, can be seen as free batteries for the grid. To incorporate the energy flexibility of residential heat pumps, their operation should be responsive to the stochastic occupant behavior, climate conditions that affect the heat pump efficiency, and solar power production. The most conventional heat pump controllers today are *rule-based controllers*, which follow a set of rules defined at the design stage. These methods are computationally inexpensive and can be easily programmed on a cheap hardware. However, rule-based controllers totally neglect the stochasticity of the environment and follow a static operational strategy which is usually far from optimal strategy. A more advanced control method is Model Predictive Control (MPC), which uses a model of the system to make predictions about the future outputs. It solves an optimization problem at each time step to determine the next actions that drive the predicted output as close as possible to the desired reference. MPC has shown a promising performance when applied to complex air conditioning systems [16–19]. However, there are several limitations to the application of MPC in practice. First of all, the performance of MPC and other model-based control methods is highly dependent on the accuracy of the developed model and prediction of the stochastic parameters. However, developing an accurate model of the system is extremely time-consuming and, therefore, not practical in most cases [20]. Moreover, even if an accurate model is developed, it can become fairly inaccurate over the time due to, for example, aging or modification of the system. Being dependent on an accurate model also makes the MPC building-specific, limiting the transferability to the other buildings and widespread adoption in the building sector [21]. To optimize the developed model at each time-step, MPC requires a considerable computational power which further limits its implementation in practice [22].

An alternative to hard-programming the expert knowledge as rule-based or model-based control methods is to give the learning ability to the controller, so it can learn by itself how to optimally control the energy system when it is applied to each new building. With recent advances in the Internet of Things (IoT) technology on the one hand, and vast progress in Machine Learning methods, on the other hand, the development of controllers which can learn by themselves is ever more realistic [21]. Among Machine Learning methods, Reinforcement Learning (RL) has recently gained popularity as a model-free control method [23]. In RL, the learning controller, known as agent, interacts with its environment and uses feedback from the environment to select the best possible action given the current state [24]. RL is gaining increasing attention for the built environment applications due to its three main advantages. First of all, it can be model-free, which therefore does not require a complicated and costly model of the system. It is a big advantage specifically when the system is complex [4]. Secondly, it is computationally efficient (after training), even when the state-space has a high dimension [22]. Finally, an RL agent can continuously adapt to the changes in the environment to maintain an optimal control policy. It makes RL an ideal method for integrating time-varying parameters such as solar energy potential, environmental conditions, or even occupant behavior into the controller. The RL agent treats occupant behavior as an unknown factor and learns and adapts to it over the time [4].

In recent years, RL has been investigated for a diverse set of applications in buildings. Park et al. [25] proposed a device called Lightlearn, which uses RL for occupant-centric control of lights in offices. The device was installed in five different offices for eight weeks. The performance of the proposed solution was compared with conventional occupancy-based and schedule-based methods in case of energy use and comfort of the people. Results showed that the occupant-centric

control based on RL successfully made a balance between occupant comfort and energy use and provided energy saving compared to both conventional methods. RL is also studied for other applications such as thermal storage inventory [26], natural ventilation [27] or integrated lighting and blind control [28]. However, regarding the big share of thermal conditioning energy use in buildings, most of the studies on RL have been focused on air conditioning systems. Zou et al. [29] developed an RL model for optimal control of air handling units to minimize the energy use, while preserving the comfort of occupants. The operational results indicate that the agent has learned how to adapt to the occupancy schedule to save energy, for example, by pre-cooling the spaces before the start of occupied hours. Schreiber et al. [22] proposed the application of RL for load shifting of a cooling network under the dynamic pricing. The cooling network included a chiller that supplied cooling to 3 different sites. The RL agent in this system was supposed to regulate the cooling supply to each site, to shift the power consumption to periods with lower electricity prices or lower outdoor air temperature while keeping the indoor air temperature violations in an acceptable range. Brandi et al. [23] implemented double deep Q-learning to control the operation of a water-based space heating system in an office building. In this study, the static deployment, where the agent is no longer trained over the deployment phase, is compared to the dynamic deployment where the agent continues training even over the deployment phase. It was shown that the RL agent with carefully designed state-space is capable of providing the necessary adaptability even in case of static deployment. Comparison with the rule-based method showed that the RL-based controller could provide 5% to 12% energy saving with an enhanced comfort. Valladares et al. [30] evaluated the potential of deep Q-learning for controlling the indoor air temperature and air quality (CO2 concentration) while reducing energy use. Two different case studies were evaluated, a laboratory room having around 2–10 occupants and a classroom with up to 60 students. The trained agent was tested in experimental setup using IoT sensors and actuators. The proposed method was then compared to the conventional rule-based control. Results show that the proposed framework could provide a better comfort (measured by Predicted Mean Vote (PMV) index) and 10% lower CO2 levels than the current control system while using about 4%–5% less energy.

There are only a few studies that have taken hot water production into account, while it accounts for a big share of buildings' energy use, and is usually integrated into the space heating systems,. Kazmi et al. [31] proposed a model-based RL control framework to balance comfort and energy use in heat pump water heating systems. In particular, they used model-based heuristics that incorporate the state of hot water tank and occupant behavior into the optimal control problem. The models for heat pump, storage tank, and occupant behavior prediction were probabilistic, data-driven models that learned from historical data. Thirty two net-zero buildings in the Netherlands using heat pumps and storage tanks were studied. It was shown that the proposed RL control approach reduces energy use for hot water production by roughly 20% with no loss of occupant comfort. Heidari et al. [32] proposed an RL-based control framework to learn and adapt to the occupants' hot water use behavior, and make a balance between energy use, comfort and water hygiene. The proposed framework was tested over data collected in a Swiss residential house. While the monitoring campaign was during COVID-19 pandemic with an abnormal occupant behavior, the proposed framework could quickly learn the occupant behavior and provide 24% of energy saving over the conventional rule-based method.

Regarding the increasing interest in integrating solar energy into buildings, a number of studies have also focused on solar-assisted space heating and hot water production. Correa-Jullian et al. [33] proposed a condition-based control approach based on tabular Q-learning for the optimal control of a solar-assisted water heating system. The Reinforcement Learning agent in this system was supposed to determine the operational schedules of the solar field and heat recovery chiller according to the energy efficiency, comfort levels, and participation of renewable energy sources. The results showed that the Reinforcement Learning-based operation performed better than the nominal operation schedule when solar radiation was low. On the other hand, nominal operation yielded a higher performance when the solar radiation was highly available. Ali and Kazmi [34] proposed an RL-based control framework for Photovoltaic-assisted (PV-assisted) domestic hot water production systems. The control approach tried to maximize the self-consumption of PV production by shifting the consumption into the periods of PV power production. However, temperatures above 50 °C were awarded equally so preventing the over-consumption of PV power for overheating the water. Comparison of the RL-based control with the rule-based control over 6 different case studies showed that the RL-based control successfully increased the self-consumption of PV production. Lissa et al. [35] proposed a framework for optimal control of PV-assisted space heating and hot water system. The proposed framework aimed to reduce energy use by optimizing the operation of the heat pump and maximizing the PV self-consumption while keeping the comfort of occupants. To monitor the comfort aspect, higher and lower temperature limits were considered for indoor air and hot water temperatures. The limits of indoor air temperature were based on the hourly average temperatures recorded in the case study building, and the limits for hot water temperature are 40 °C and 55 °C. It was indicated that as indoor heating is a slow process, the agent can better follow the comfort limits, but the water heating is a faster process and, therefore, there is a higher probability of surpassing the comfort limits. The evolution of reward term showed that after the first month of training, the agent learned to keep the occupant comfort and the occupants no longer experienced high deviations from comfort limits. The proposed framework could provide 8% to 16% energy saving compared to the rule-based controller.

1.1. Objectives and contributions

This paper proposes an RL-based control framework for PV-assisted space heating and hot water production, which can learn and adapt to the stochastic parameters, namely hot water use behavior of occupants, PV power production, and outdoor air temperature, and accordingly make a balance between energy use, comfort, and water hygiene. Very few studies have investigated RL for the entire system of solar energy, space heating, and hot water production. This study intends to further broaden the current knowledge by investigating the following aspects:

- **Model-free:** This framework does not use any model, such as a data-driven or thermodynamic model of the system, and rather learns the required knowledge from scratch. Experts usually assume that by providing their prior knowledge to the agent, it will learn easier and perform better. However, it is not always the case. An example is AlphaZero, an RL model developed to play the Go game, which by learning from zero and playing by itself significantly performed better than its prior model AlphaGo [36]. Most importantly, being model free will facilitate the transferability of the control framework to the other residential buildings with different system specifications;
- **Integration of water hygiene:** Legionella is a waterborne bacteria that grows in hot water between 25 °C and 57 °C and pose health risks to the occupants. According to the literature review, the hygiene aspect of water is never investigated in previous studies on RL. This is while the hygiene aspect, mainly Legionella, is the main barrier for reducing water temperature to save energy [37]. This study integrates water hygiene into the control framework by taking into account the Legionella growth model. This will help the agent to properly adjust hot water temperature for reducing energy use without endangering the health of occupants;

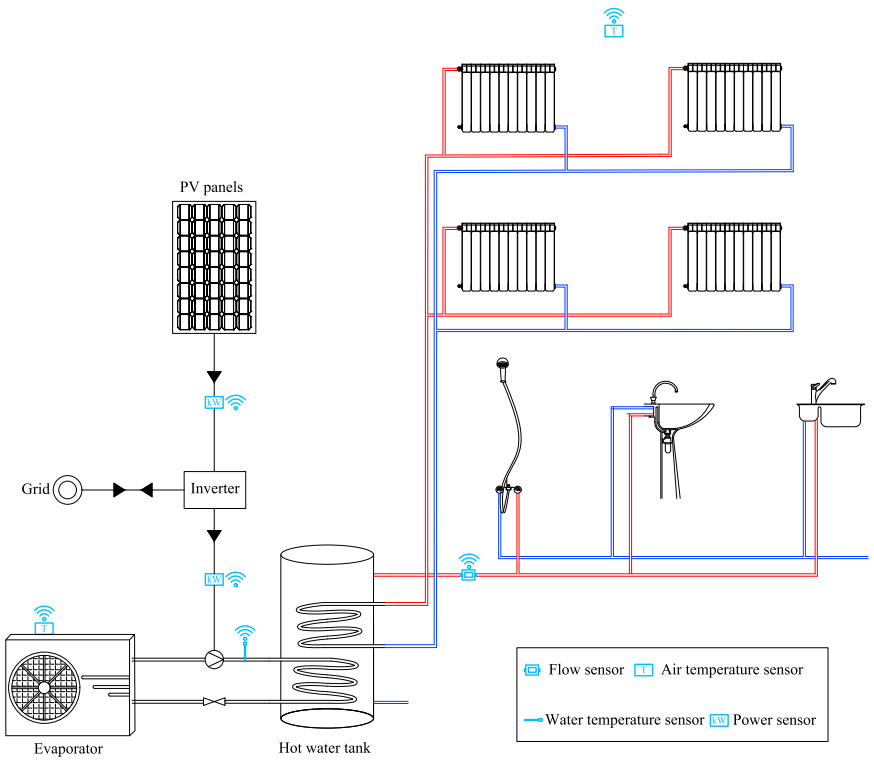


Fig. 1. Configuration of system to be controlled by RL.

- **Stochastic-based offline training:** To speed-up the convergence and to minimize the risk of violating comfort or hygiene aspects on the target house, a stochastic-based offline training phase is designed to provide enough experience to the agent in the safe simulation environment before being implemented on the target house. The offline training phase integrates a stochastic hot water use model and trains the agent over a variety of system sizes, geographical locations, and hot water use behaviors to ensure the agent has obtained a generalized experience and can quickly adapt to different houses. Offline training is done in simulation, which is a safe environment where the agent can learn from scratch and even try random actions without any consequences on the real occupants;
- **Investigating the adaptation potential to different hot water use behaviors:** Hot water usage is the most stochastic parameter that the agent needs to adapt to. To evaluate the adaptation potential, real-world hot water use behavior is monitored in 3 Swiss residential houses. As the monitoring campaign was performed over the COVID-19 pandemic, it will allow investigating the adaptation potential from the normal behavior observed in the offline training into the abnormal behavior during the COVID-19 pandemic. Also, the behavior of 3 cases was found to be very different, which allows to further investigate the adaptation potential of the agent to different occupant behaviors;
- **Investigating the generalization potential of the knowledge gained in offline training:** A well-designed training procedure should provide a generalizable knowledge to the agent. If the knowledge gained in offline training is generalizable, it can minimize or at the best case eliminate the need for an online training on the real house. Since the online training of the agent on the cloud can be challenging and costly, in practice, it would be much easier if an agent could be only trained on simulations and directly deployed on the target environment. This paper investigates two scenarios. The first scenario is the direct deployment of the agent, where the agent is directly deployed on the target house after offline training, without any online training on that specific house. The second scenario is long-time deployment, where after a short-time online training, the agent is deployed for a long time to see if there is a need for sequential trainings or one initial training is enough. These scenarios can provide insight for elimination or reduction of training phase on the target house by a generalizable offline training, which will facilitate the practical implementation of RL in residential buildings;

The remaining of this paper is organized into four sections. The first section presents the methodology of the research. The second section gives a brief overview of the case study houses and the monitoring campaign. The results of the study are outlined in the third section. Finally, the fourth section concludes the paper.

2. Methodology

The methodology section presents the energy system to be controlled, monitoring campaign in the case study houses, developed model used for estimation of Legionella concentration, the proposed RL control framework as well as baseline control scenarios.

2.1. Case study description

2.1.1. System configuration

The proposed framework is focused on a residential energy system including space heating, hot water production and PV power generation. There are many alternative configurations for this system, such as integrated or separated thermal storage for space heating and hot water production. However, as the aim of this study is to prove the potential of RL for optimal operation of these systems, one common configuration of the system is examined as an example. The proposed framework

Table 1
Area and number of occupants in case study houses.

	Heated area (m^2)	Adults number	Children number
House 1	160	2	3
House 2	120	2	2
House 3	150	2	1

can be easily adjusted to other configurations. The configuration used in this study is shown in Fig. 1.

The heating system is air-to-water heat pump, a favorable power-to-heat option with increasing number in building sector, that can provide a great opportunity for solar energy integration. Heat pump has a variable Coefficient of Performance (COP) depending on outdoor air and hot water temperature. This dependency makes it more challenging for the RL agent to schedule heating cycles. Secondly, hot water tank is considered as an energy storage, because it is more cost-effective than electric storage [15], provides both functionalities of energy storage and hot water provision, and is available in many buildings. While the space heating can be integrated or detached from energy storage, in this configuration it is considered to be integrated to storage to provide further energy flexibility. In this case, the surplus solar energy can be either stored in the tank and be used for space heating later on. PV panels are considered to be grid-connected, so the surplus power can be also injected to the grid.

2.1.2. Monitoring campaign

Hot water demand is less predictable than space heating demand, can be very different between similar buildings [38], can impose a fast change in the hot water tank temperature which causes the violation of user comfort [39]. Thus, for the proposed framework, the most challenging task for the agent is to learn the hot water use behavior of occupants in each building. This study intents to evaluate the performance of framework over the actual hot water usage measurements. In this research, a cost-effective, low-power and water-proof monitoring system is implemented to monitor all the assets, and then the flow rate of all end-uses are summed to obtain the main flow rate. Monitoring all the end-uses is not needed for this framework, and a single sensor on the tank outlet can provide the demand data. The detailed monitoring in this research was to provide a high-resolution dataset for future research.

Three residential houses in Switzerland are monitored for 20 weeks. Geographical location of buildings is indicated in Fig. 7. Monitoring period of houses 1 and 2 was entirely during cold season (28 August 2020 to 15 January 2021), while for house 3 it also includes the hot season (23 March 2021 to 10 August 2021). The third house is to analyze how the agent will adapt to a period where PV power production is high but energy demand is low (as there is no space heating demand in this period). The heated area and number of adults and children in each building are shown in Table 1. As shown in this table, the case study buildings are selected to include a variety of family compositions, which allows to further evaluate the adaptation potential of the agent to different houses.

The case studies were equipped with heat pump. But since they were occupied residential buildings, in this phase of early evaluation it was not desired to test the proposed framework directly on the actual systems as it could result to discomfort and dissatisfaction of tenants who were volunteer in this study. Rather, the real-life collected data can provide a realistic evaluation in simulation environment, without violating the comfort of occupants.

2.2. Legionella concentration model

Legionella is a water-born bacteria that grows in water between 25 °C and 47 °C and can be transferred to humans by breathing in the contaminated water droplets. Infection with this bacteria results in a respiratory illness, known as Legionnaires' disease (LD) [37]. Hot water systems are responsible for the most number of infection cases, as they can provide the desirable temperature regime for the growth of Legionella [40].

While there are several disinfection methods, such as chemical methods, one of the most conventional methods is thermal disinfection [40]. With a temperature of 60 °C Legionella cells die in only 2 min [41]. Therefore, as a common practice the hot water tank temperature is constantly kept above 60 °C to ensure Legionella cannot grow in the tank. The high temperature of hot water tank will reduce the heat pump COP, increase the heat loss, and also increase the risk of scalding at the point-of-use. This conservative operational approach is because controller does not have any sense about the real-time risk of Legionella in the tank. This framework aims to quantify the risk of Legionella for the agent in real-time, so it can overheat and disinfect the tank only when it is needed. Legionella growth is a complicated process that depends on many different factors such as temperature, PH, and existence of nutrients [42]. It is therefore complicated to develop a model for accurate calculation of Legionella concentration. Few mathematical models are developed to estimate the Legionella concentration only based on water temperature variations [37,43,44]. Assuming that the hot water tank has not been initially contaminated with Legionella and biofilm, and also the network water is properly treated, these models can be used to provide the real-time estimation of Legionella concentration only based on temperature. Controlling the hot water tank temperature by considering Legionella concentration can make a shift from energy-intensive conservative control approaches into energy-efficient while safe methods. However, little attention has been given to the integration of Legionella risk assessment into the control systems. Kenhove et al. [45] integrated a model of Legionella concentration into the rule-based controller, where the controller heats the tank when the estimated concentration passes a threshold. Based on the literature review, there is no study on the integration of Legionella growth into RL-based control frameworks. Different from the rule-based control which only overheats the tank when a threshold is passed, an RL agent can learn how to proactively plan overheating cycles while minimizing energy use, for example, by overheating the tank when there is a surplus of PV power, or heat pump COP is higher, or a demand is expected to happen in near future.

Estimation of Legionella concentration in this study is based on the model proposed by Amerongen et al. [44]. In this model, for the temperature range of 25 °C and 47 °C, the doubling time (the number of hours required for Legionella concentration to get doubled) is calculated as:

$$DO = 0.5702 \times T_{tank}^2 - 43.3 \times T_{tank} + 829 \tag{1}$$

where T_{tank} is the hot water tank temperature (°C) and DO is doubling time (hours). Using the doubling time, and integrating the effect of inlet and outlet water streams, the following equation is used to calculate the concentration of Legionella in this temperature range:

$$C = \frac{(C_{initial} + \frac{C_{initial}}{DO}) \times V_{tank} + C_{network} \times Demand - C_{initial} \times Demand}{V_{tank}} \tag{2}$$

where $C_{initial}$ is the concentration of Legionella at the beginning of timestep (CFU/L), the $C_{network}$ is the concentration of Legionella in network water (CFU/L), $Demand$ is the hot water demand (L), and V_{tank} is the tank volume (L), and C is the concentration of Legionella at the end of that timestep. Regarding that in the hot water tanks the same amount of consumed hot water is replaced by the cold network water, the term $C_{network} \times Demand$ is the amount of Legionella entering the tank from network water, and $C_{initial} \times Demand$ is the amount of Legionella exiting the tank. For the temperature above 60 °C, the reduction in concentration is calculated as:

$$C = \frac{(C_{initial} - 0.999 \times C_{initial}) \times V_{tank} + C_{network} \times Demand - C_{initial} \times Demand}{V_{tank}} \tag{3}$$

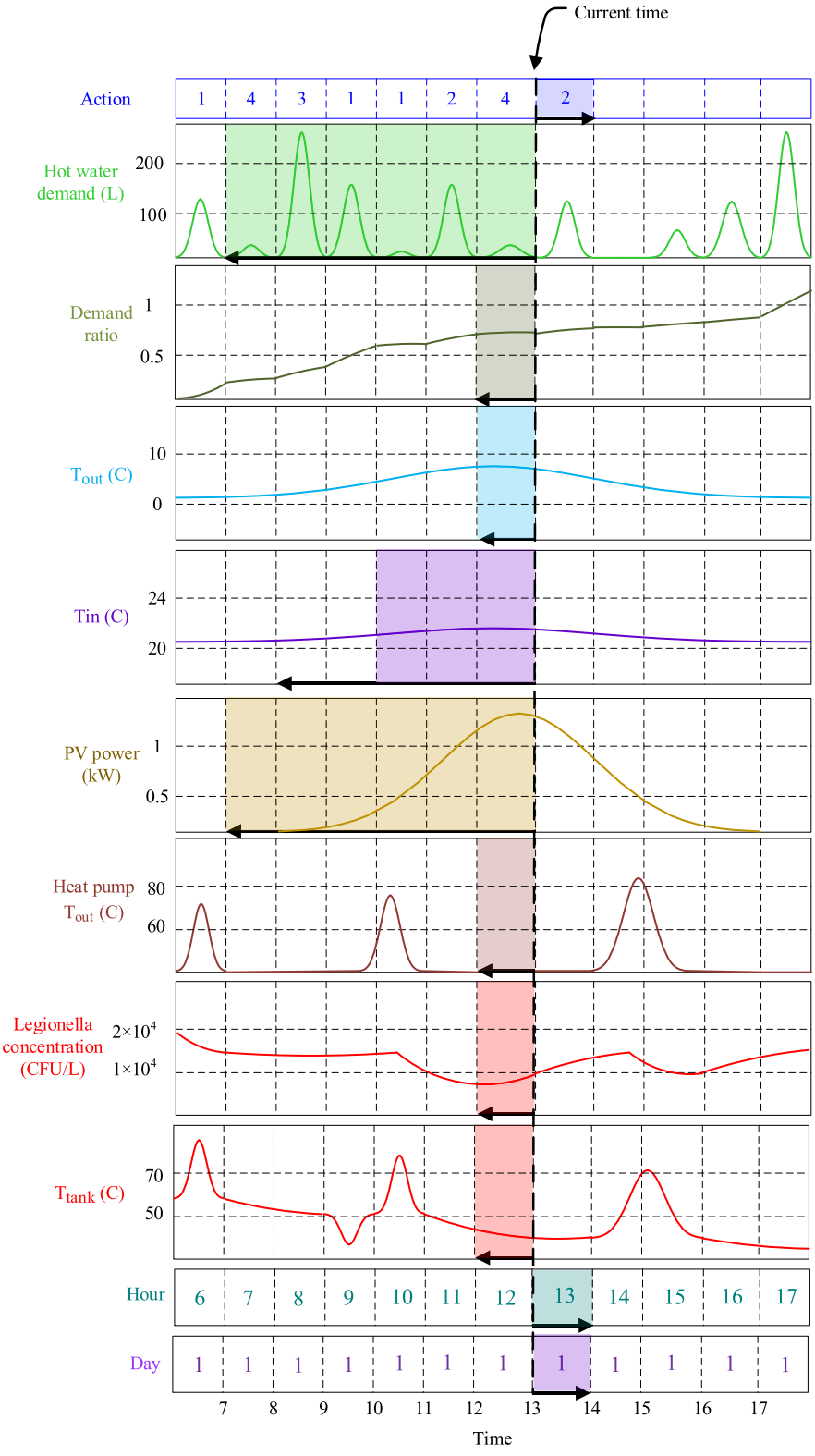


Fig. 2. Visual representation of states and actions.

For the temperatures below 25 °C or between 47 °C and 60 °C, the concentration of Legionella is assumed to be constant. It is a conservative assumption to further ensure the health of occupants, because for a temperature above 50 °C the disinfection still happens but with a lower rate [41].

2.3. Reinforcement learning control framework

A variety of RL algorithms have been developed so far. These algorithms can be divided into two main categories of policy-based and value based methods. Policy-based methods are suitable for problems with a continuous action space (such as robotic applications), while

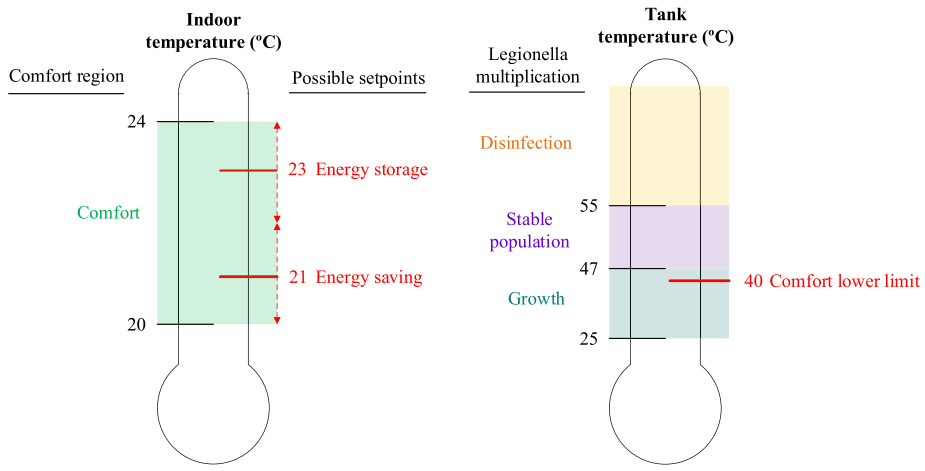


Fig. 3. Temperature ranges for comfort limits of indoor air and Legionella multiplication and comfort limit for hot water tank.

value-based methods are suitable for environments with a discrete action space, where the agent implicitly finds a policy by learning the optimal value function [46]. It is shown that value-based methods learn faster, as they include a limited number of possible actions, and are less sensitive to hyper-parameter tuning [47]. One of the most widely used value-based RL algorithms is deep Q-learning. Deep Q-learning tries to estimate the value of each action, known as Q values, and select the action with the highest estimated value. These values are calculated based on the following formula:

$$Q^{new}(s_t, a_t) = Q(s_t, a_t) + \alpha.(r_t + \gamma.max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (4)$$

where $Q(s_t, a_t)$ is the old estimated value, $\alpha[0, 1]$ is the learning rate, r_t is the immediate reward, γ is the discount factor, and $max_a Q(s_{t+1}, a)$ is the estimated future reward. As the original deep Q-learning use the same network for the estimation of future values, it can lead to the overestimation of value for some specific actions and therefore a non-optimal action can be selected. To solve this issue, a modified technique called double deep Q-learning is recently developed. The main characteristic of this technique is the presence of two networks to counteract the overestimation of the Q-values. The second network is an exact copy of the first one, but is only updated every τ steps, and is used to calculate the target Q-values for expectation [48]. Therefore, this research use double deep Q-learning algorithm to develop the control framework. Tensorforce library [49] is used to program this framework in Python. In spite of other RL libraries that are mainly developed for computer games, Tensorforce is developed with a Modular design, making it easy to adjust the agent and the environment for other domains.

2.3.1. State, actions and reward design

The RL agent observes the state of the environment, then selects an action based on the observed states, and tries to maximize a reward. The proper setup of states, actions and rewards is an important aspect to design a robust RL framework. State parameters should provide all necessary information for the agent to predict future immediate reward, and also should be possible to be collected by sensors in practice [23]. The following parameters are included in the state vector:

- **History of hot water demand:** As one of the most important aspects of this framework, the agent is supposed to learn and predict future hot water use behavior of occupants. Studies have shown that there are some routines in hot water use behavior of occupants in residential buildings, and therefore future hot water use is correlated with the historical demand [32,50–52]. Therefore, a look-back vector of previous hot water demands is included in the state vector to enable the agent to forecast future demands. The length of this vector (the number of previous hours to be included) for this parameter and also other parameters of state will be determined based on the sensitivity analysis.
- **Demand ratio:** It would be useful to the agent to estimate how much hot water would be used in total up to the end of day. As the number of people and their daily behavior is almost persistent, the total hot water demand of today can be close to yesterday. The remaining hot water demand at each hour up to the end of day, can be close to the remaining demand of the same hour up to the end of the day for the previous day. The following percentage quantifies the ratio of consumption up to the current time of today, over the total consumption of previous day.

$$DR = \frac{\sum_{h=0}^{H} Demand_{Day=D}}{\sum_{h=0}^{24} Demand_{Day=D-1}} \quad (5)$$

where H is the current time of day, D is the day number, $Demand$ is the volumetric demand (L), and DR is Demand Ratio.
- **Outdoor air temperature:** The outdoor air temperature affects the space heating demand and also heat pump COP. A look-back vector of outdoor air temperature lets the agent to learn the future variations of outdoor air temperature and heat pump COP.
- **Indoor air temperature:** Indoor air temperature is important for the agent from two aspects. First of all, it affects the occupants comfort and should be carefully adjusted. Secondly, as the building thermal mass is also a potential energy storage, it is indicating the current level of stored energy in the building thermal mass.
- **PV power production:** Another important functionality of this framework is to learn and predict PV power production and optimally schedule the future actions. The look-back vector of PV power production enables the agent to learn the future PV power production.
- **Heat pump outlet water temperature:** The heat pump outlet water temperature determines the maximum possible temperature in tank, and also affects the rate of energy delivery to the tank and subsequently indoor air.
- **Legionella concentration:** For optimal adjustment of the hot water tank temperature, the agent should know the current estimated concentration of Legionella in hot water tank (CFU/L). This lets the agent to prevent unnecessary thermal disinfection of tank, and only overheat the tank when it is needed or when surplus of PV power production needs to be stored in the tank.
- **Hot water tank temperature:** The hot water tank temperature indicates the level of stored energy in the tank, which also should be kept above the comfort temperature.
- **Hour of the day:** Many of the stochastic parameters, such as occupants hot water use behavior, solar energy and outdoor air

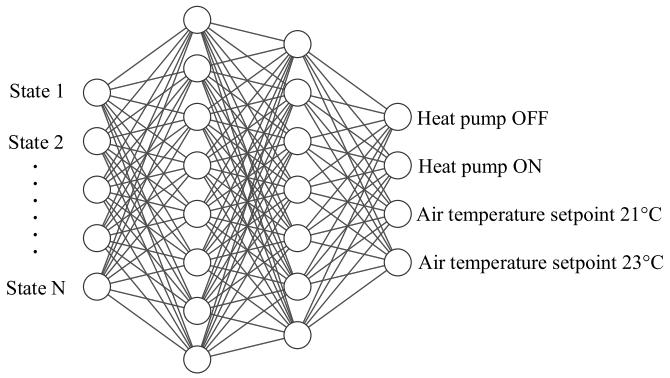


Fig. 4. Possible actions for the agent.

temperature are strongly correlated with the hour of the day. To further assist the agent to learn and predict these parameters, hour of the day for the upcoming hour is also provided to the agent. Different from other parameters, this is not a look-back vector but is associated to the upcoming timestep.

- **Day of the week:** There is a significant difference between the hot water use profile of working days and weekends. Also, the hot water use profile of each day is found to be highly correlated with the profile of the same day over the last week [32,50–52]. Accordingly, to learn and predict the future hot water demands it would be helpful for the agent to know what is the current day of week. The day number for the upcoming timestep is therefore provided to the agent.

A visual representation of state parameters at each time step is shown in 2. The length of look-back vector indicated for each parameter is symbolic in this figure as it will be determined over the sensitivity analysis.

Possible actions should also provide enough flexibility for the agent to maintain an optimal operation. The possible actions in this study are selected according to the comfort limits and hygiene aspects. As shown in Fig. 3 the comfort limits for indoor air temperature in winter are between 20 °C and 24 °C based on *ISO7730* [53]. Regarding that this range is quite narrow, to ensure the comfort of occupants while providing enough flexibility for the agent, it is assumed that the agent only selects a setpoint to overwrite the existing thermostat. The possible setpoints are 21 °C and 23 °C, with a dead-band of 2 °C, which therefore covers all the comfort region. The option of 21 °C is an energy saving choice, while the option of 23 °C provides the opportunity of storing surplus PV power in building thermal mass. In case of hot water tank, the multiplication of Legionella at each temperature range, as well as the comfort limit for hot water are shown in Fig. 3. While the required temperature of mixed water at each point-of-use is different, 40 °C is assumed as the minimum required supply temperature [38]. In this research, 40 °C is considered as the minimum comfort level for the average hot water tank temperature. This will further ensure the comfort of occupants as the hot water is supplied from the top of the tank which has a higher temperature due to the stratification of tank. Since the range of possible temperatures for hot water is quite wide, and discretization of setpoints would result in many different actions, possible actions for hot water temperature adjustment is considered as turning ON and OFF the heat pump. This would give the possibility to the agent to adjust any temperature with only two actions. On the other hand, the agent should properly learn the relationship between the hot water tank temperature and all the affecting factors, such as future hot water demands, and schedule the ON/OFF actions properly to avoid any comfort violations.

The possible actions are presented in Fig. 4. Actions related to the hot water tank are separated from the ones related to the space heating, meaning that the agent cannot simultaneously change the indoor air setpoint and heat pump status, and should prioritize between them. While it is possible to combine the tank and space heating actions, for example one action representing turning ON the heat pump and also indoor setpoint of 21 °C, primary tests indicated that such combined actions make it more complicated for the agent to learn the relationship between performing each action and the associated impact on the environment.

The reward function should be well designed to clearly reflect the aims and priorities as simple as possible. This control framework intends to minimize the energy usage of heat pump, and maximize the self-consumption of PV power, while maintaining the occupants comfort and water hygiene. The reward function is composed of four different terms as follow:

- **Energy term:** The energy term penalizes the agent for (1) any energy usage of heat pump and (2) the surplus of PV power not used by the heat pump. This term is defined as

$$R_{energy} = -a\times \mid HP_{power} - PV_{power} \mid \tag{6}$$

where HP_{power} and PV_{power} are the power usage of heat pump (kW) and power production of PV panels (kW), accordingly. a is the weighting factor, used to make the same scale between different terms and also put more emphasize on higher priority terms.

- **Hot water comfort term:** This term penalizes the agent if the temperature of hot water tank falls below the comfort level of 40 °C.

$$if\ T_{tank} \geq 40,\ R_{DHWcomfort} = 0\ else\ -b \tag{7}$$

where T_{tank} is the hot water tank temperature and b is the weighting factor. It is also possible to penalize the agent proportional to the temperature deviation, but here the comfort and hygiene penalizations are done with constant numbers to speed up the learning process.

- **indoor air temperature comfort term:** This term penalizes the agent if the indoor air temperature is out of comfort limits. While the possible setpoints are inside of comfort region, still a comfort violation can happen if the hot water tank temperature is not high enough to provide required heat for radiators. This term is defined as

$$if\ 20 \leq T_{indoor} \leq 24,\ R_{Indoorcomfort} = 0\ else\ -c \tag{8}$$

where T_{indoor} is the indoor air temperature and c is the weighting factor.

- **Hygiene term:** This term penalizes the agent if the estimated concentration of Legionella in hot water tank exceeds the maximum acceptable level. This term is defined as

$$if\ C \leq C_{max},\ R_{Hygiene} = 0\ else\ -d \tag{9}$$

if $Concentration$ is the current concentration of Legionella (CFU/L), and $Concentration_{max}$ is the maximum acceptable concentration (CFU/L).

The total reward, which is going to be maximized by the agent, is then the summation of these rewards as

$$R_{total} = R_{energy} + R_{DHWcomfort} + R_{Indoorcomfort} + R_{Hygiene} \tag{10}$$

2.3.2. Training procedure

To train the RL agent, it is required to establish an interaction between the agent and environment, which lets the agent to perform actions on the environment, receive back the next state and calculate

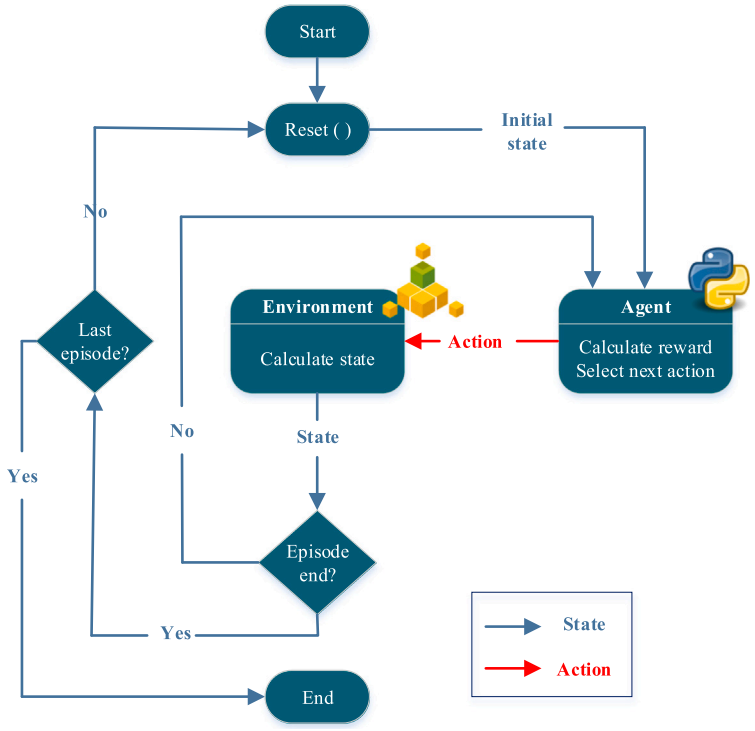


Fig. 5. Procedure of interactions between the agent developed in Python and system model developed in TRNSYS.

the subsequent reward. The interactive procedure in this research is established by coupling the agent developed in Python with the dynamic model of system developed in TRNSYS. Fig. 5 presents how the agent and environment interact with each other. At each timestep, the agent writes the selected action to the input file of TRNSYS, runs the TRNSYS model for one hour, and then reads the subsequent parameters of state from TRNSYS output file. If the episode is not ended, the state is again used by the agent to select the next action. And if it is the last timestep of episode, the state is reset and sent to the agent. As the agent tries to maximize the reward over the period of each episode, the reset is to set the timestep counter as zero and inform the agent that the episode is ended. The length of episode is considered as one week.

This study propose a multi-step training procedure in which the agent is first trained offline on a safe virtual environment, then trained online on the target house and finally is deployed on that house. The overall procedure is presented in Fig. 6. During the offline training phase, the agent is interacting with the virtual model of the system for 10 years. An important consideration in the offline training phase is to provide a generalizable knowledge to the agent, so it can be transferred to different houses with different system sizes, located in different weather conditions of Switzerland, and with different occupant behavior. To provide a generalizable knowledge of the occupants hot water use behavior, an stochastic hot water use model driven by actual data from other buildings [54] is used to simulate the hourly demand data. Actual weather data from multiple weather stations in Switzerland are also collected, and for each year of the offline training phase, the solar and weather data of a different city is used as indicated on Fig. 7. In addition, a different set of system sizes (e.g. heat pump capacity, hot water tank volume, radiators and PV panels area, etc.) are used in each year. Including these variations in the long-time offline training phase also reduces the possibility of overfitting to a specific case. The pre-trained agent is then saved to be used for online training on each of the target houses. It should be noted that the simulation model is only used to provide an initial experience for the agent, so it is not part of the framework and does not need to be an exact model of the target system. Therefore, the proposed framework fits in the category of model-free RL.

On the online training phase, the pre-trained agent is again trained with the actual hot water demand and weather data of the target house. While the offline training phase might be enough for the agent, the online training on the target house can help the agent to further adapt the agent to the specific house. In this phase, the actual hot water demand data that are measured experimentally in each house is used to represent the real behavior of occupants. Detailed description of monitoring campaign is provided in the next sections. Once the agent is trained for several weeks on the target house, it is then deployed on this house. It means that the agent is no longer learning but only controlling the system. This phase is computationally efficient and can be done offline on a low-price hardware such as a Raspberry Pi.

While TRNSYS models are used in all phases, it should be noted that the model in offline training phase is a virtual model to be used in a laboratory, while the model used in online train and deployment phases is to represent an actual building.

2.3.3. *Different training scenarios*

To get the full potential of RL it should be continuously trained, enabling it to adapt to all changes during the life-time of the system, which is however costly and computationally expensive. This research aims to gain a good level of adaptation, without being continuously trained online, by incorporating the intensive stochastic-based offline training procedure. Online training of the agent is challenging, costly, and less robust as it would depend on the continuous interaction between sensors and agent on the server. It would be therefore easier in practice if the duration of online training phase would be reduced as much as possible, or it would be even totally eliminated. One of the aims of this study is to assess if the stochastic-based intensive offline training can (1) eliminate the need for continuous or sequential online training phases on the target house, and (2) totally eliminate the need for online training on the target house. To this aim, three different scenarios for training phase are evaluated. These scenarios include:

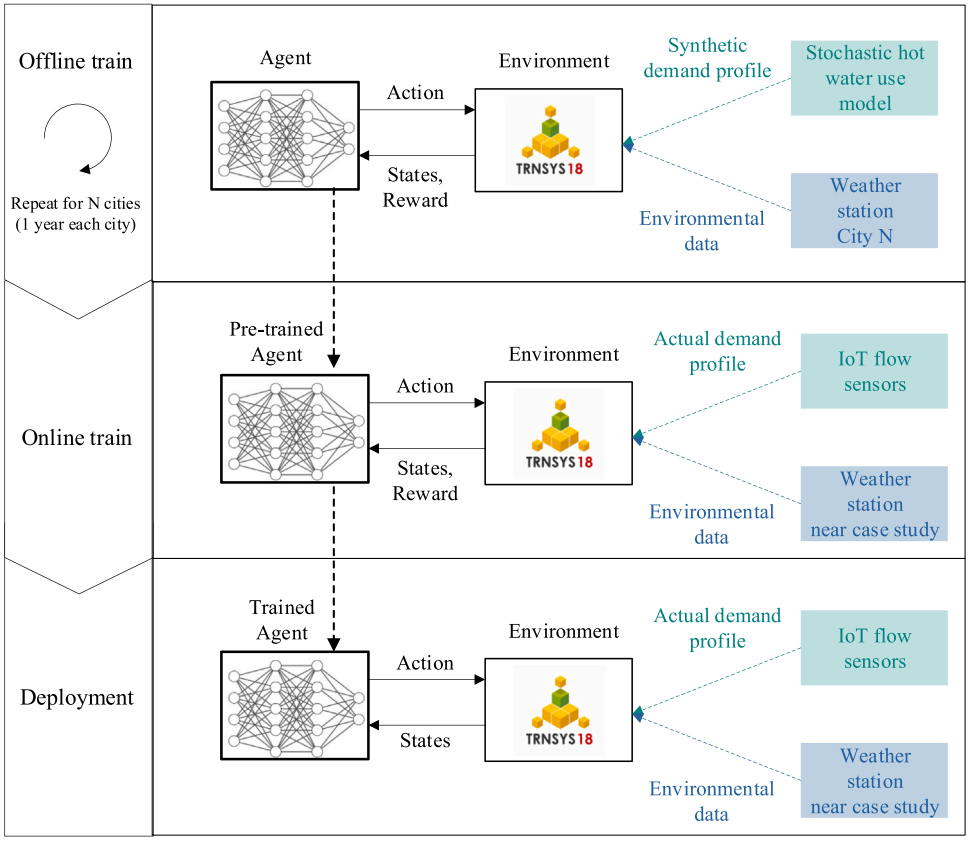


Fig. 6. Training procedure.

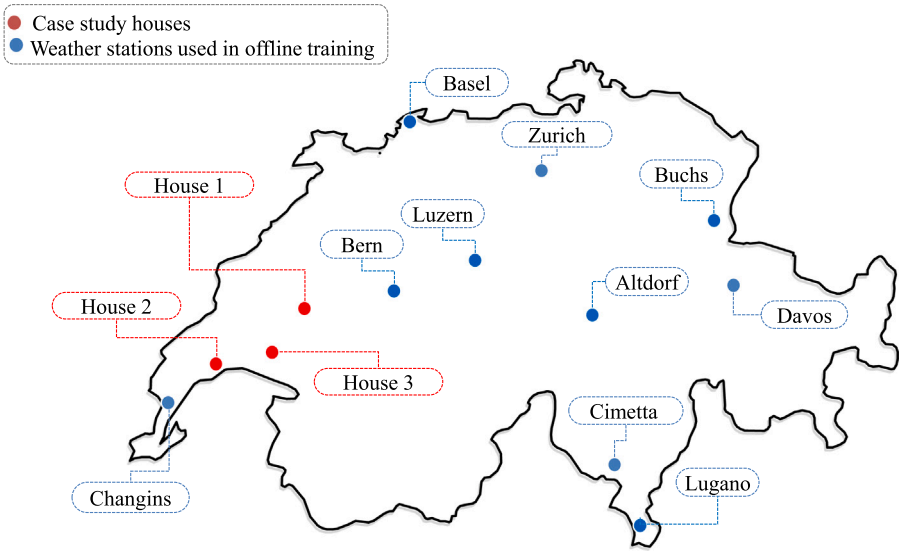


Fig. 7. Location of cities used in offline training phase as well as case study houses on the Swiss map.

- **Online training and Short-time Deployment (RL-OSD):** After offline training, the agent is trained online on the target house, and then deployed for a short period of 1 month;
- **Online training and Long-time Deployment (RL-OLD):** After offline training, the agent is trained online on the target house, and then deployed for a long period of 8 month;

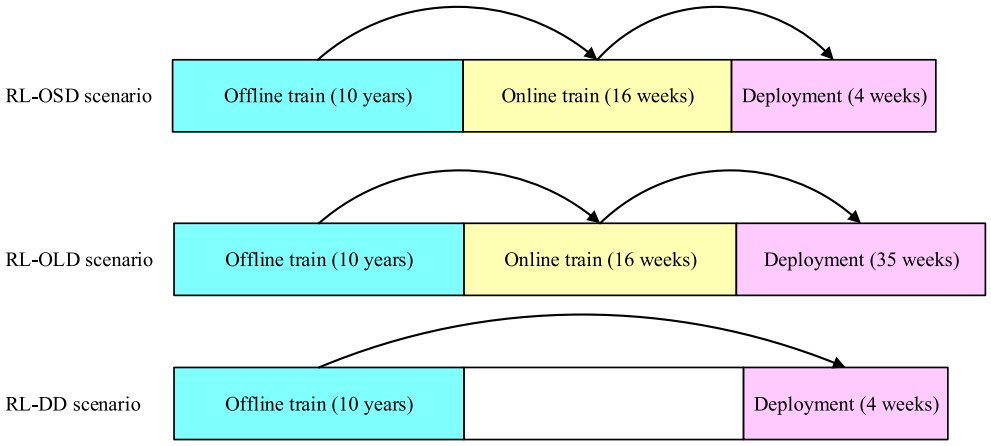


Fig. 8. Visual presentation of different training and deployment scenarios.

- **Direct Deployment (RL-DD):** After offline training, without any online training, the agent is directly deployed on the target house for a short period of 1 month;

For a better understanding, these three scenarios are visually presented in Fig. 8.

2.4. Baseline control methods

In order to better highlight the advantage of a learning controller, it can be compared to the conventional rule-based controllers that only follow static programmed rules while ignoring the variations of occupant behavior, solar energy or weather conditions. Two following rule-based controllers are also modeled in this study:

- **Rule-based controller with Conventional setpoints (RC):** A rule-based method which uses the setpoints of common practice. In this method, setpoint air temperature is considered as 21 °C with a deadband of 2 K, which is a recommended setpoint for healthy and comfortable air temperature, [55,56], and 60 °C with a deadband of 10 K for hot water tank, which is a commonly used setpoint to follow hygiene requirements in storing hot water [57,58];
- **Rule-based controller with Energy saving setpoints (RE):** A rule-based method with similar setpoint air temperature to the RC method, but with the setpoint tank temperature of 50 °C for energy saving;

While due to the hygiene aspects, the RE scenario is not common in practice, in this study it is considered to illustrate that the energy saving of proposed control framework is not only achieved by lowering the setpoint temperatures, but rather by learning how to optimally schedule the heating cycles. There are many other alternative control methods, such as using a heat curve, that are today applied in the buildings. These methods are similar in the sense that they follow static rules, which are detached from occupant behavior or renewable energy. Similar results are expected if a comparison is made between the learning agent and other rule-based controllers.

2.5. System sizes

Table 2 shows the specifications of modeled systems used in the offline training and target houses. The agent is supposed to be able to adapt to a new building, with different area and different system sizes than what it has observed during the offline training phase. The heated area and heat pump capacity in case study buildings are bigger (House 1), smaller (House 2), and almost similar (House 3) to the offline training phase. Area of PV panels is equal to the available area for tilted roofs calculated based on [59]. Heat pump rated heating is also proportional to the heated area, and is sized based on the capacity per area of a real-word similar installation presented in detail in [15]. The same tank size is considered in all houses for simplicity.

3. Results

In summary, the results of this study are presented in 5 sections as below:

- **Dataset overview:** Provides an overview of collected datasets during monitoring campaign;
- **Hyper-parameters:** Describes the hyper-parameters selected for the proposed framework;
- **Reward evolution:** Evaluates the convergence of proposed framework;
- **Visual assessment:** Some operational parameters (e.g. air temperature, water temperature, hygiene, etc.) are visualized to provide a detailed and hourly presentation of the agent performance;
- **Quantified assessment:** Quantification metrics are used to summarize and compare the agent performance (such as total energy use) with respect to the conventional methods;

3.1. Overview of datasets of different houses

Fig. 9 shows the hourly variations of hot water demand, PV power production and outdoor air temperature in three different case studies. It can be seen that there is a good diversity in hot water use behavior of case study houses, as the Houses 1, 2 and 3 can be categorized as high volume (up to 250 L/h), low volume (mostly below 50 L/h), and middle volume (up to 150 L/min) consumers. There is a good variation also in the trend of PV power production in case study houses. Hourly variations of PV power on the first and second case studies show a decreasing trend, with higher values during the training phase compared to the deployment phase. On the third house, the date of monitoring campaign has been different from the first and second case studies, with the training phase starting from cold weeks and the deployment phase during the warmer weeks. Therefore, the trend of PV power production is increasing in this house, with higher hourly production during the deployment phase compared to the training phase. Variations of hourly outdoor air temperature also show a similar trend, with a decreasing trend on the first and second case studies and an increasing trend on the third case studies. The deployment phase of the first and second case studies is during the cold weeks, when both space heating and hot water production is required, while the deployment phase of the third house is during the warm weeks,

Table 2
System sizes used in offline training and different case studies.

	Offline training	House 1	House 2	House 3
Total heated area (m^2)	140	160	120	150
Heat pump rated heating (kW)	6	7	5	6
Heat pump compressor power (kW)	0.95	1.1	0.8	0.95
Tank size (L)	500	500	500	500
PV panels type	Monocrystalline	Monocrystalline	Monocrystalline	Monocrystalline
PV panels total area (m^2)	10	11	8	12
Panels slope	45	45	45	45

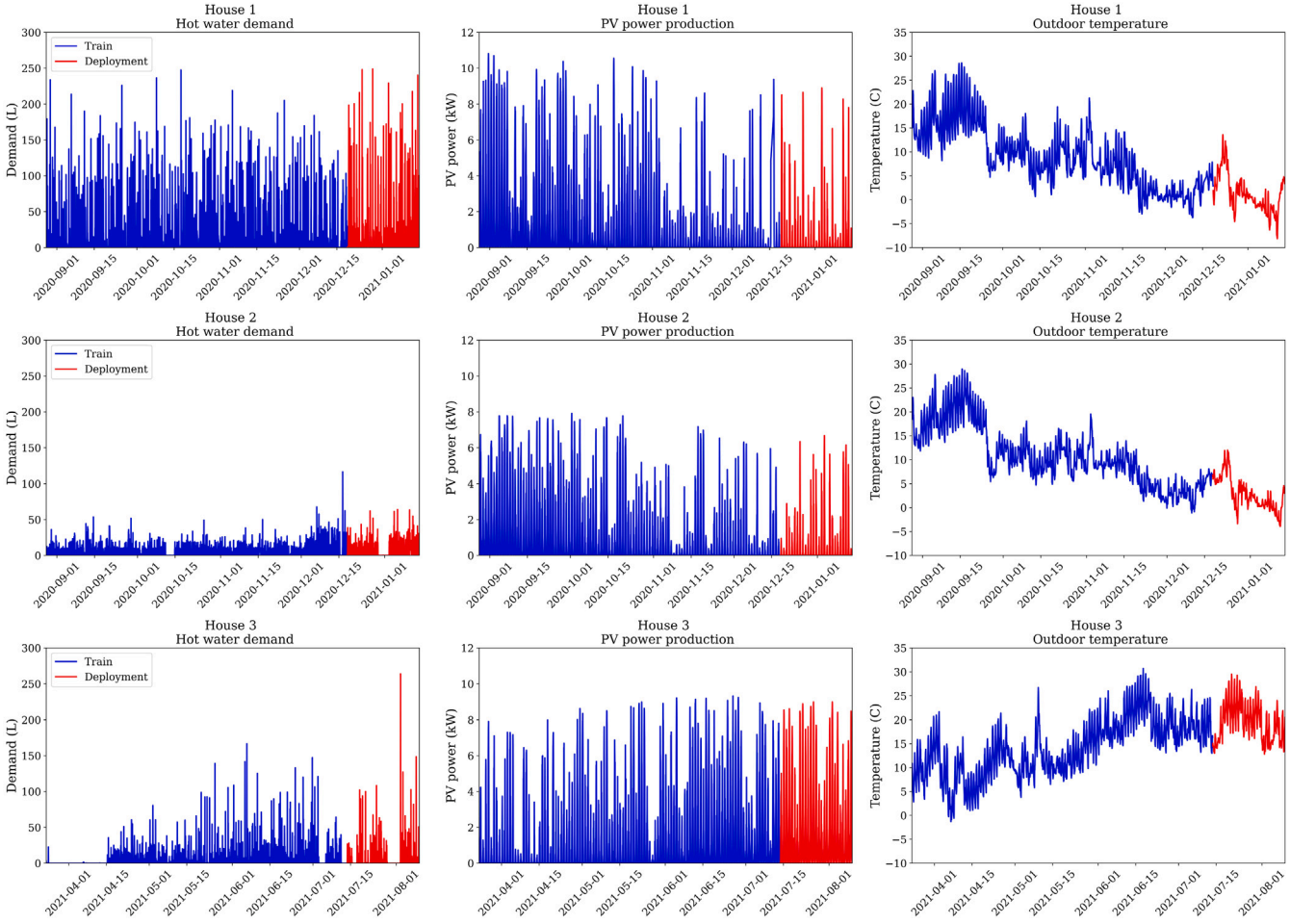


Fig. 9. Hourly hot water demand, PV power production and outdoor air temperature on the case study houses.

Table 3
Selected parameters for the agent.

Hyper-parameter	Value
Agent type	Double deep-Q network
Learning rate	0.003
Batch size	24
Update frequency	4
Memory	48 × 168
Discount factor	0.9

when only hot water production is required. The agent is supposed to learn that during the warm weeks there is less energy demand, and the variations of the hot water tank temperature only depends on the hot water demand. The overview of datasets show that there is a good diversity between the case studies, and between train and deployment phases. These variations provide a great opportunity to examine how the agent can generalize its knowledge and adapt itself to different situations, such as different hot water use behaviors.

To better explore the diversity in hot water use behavior between the case study houses, boxplots of their hourly hot water use data are also presented in Fig. 10. Datasets from other residential buildings [50] show that hot water use pattern usually has two major peaks, one in the morning and the other in the evening. Regarding that the monitoring campaign in this study has been during COVID-19 pandemic, the monitored data over these three houses show some differences with the normal pattern of residential buildings. For example, the peak of average demand for the Houses 2 and 3 is located at the middle of the day, while in the normal situation occupants are at work on this time and no peak is expected. Also the hot water use pattern in House 2 shows a quite uniform demand between 7 A.M. and 9 P.M., which indicates that the occupants have been spending most of their time at home. These differences indicate that the hot water use behavior over the case studies is more stochastic and less predictable than the normal behavior that the agent has observed during the offline training. This

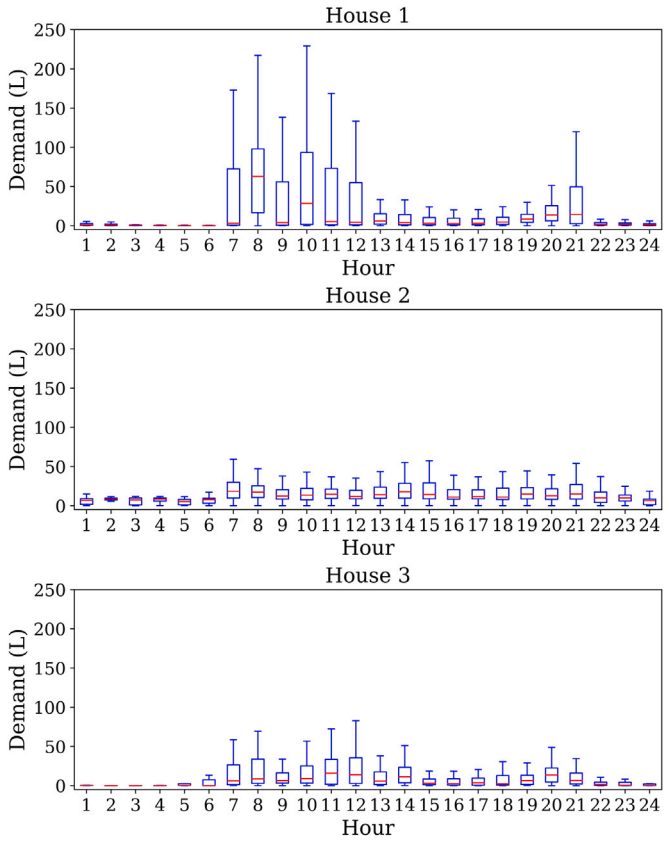


Fig. 10. Boxplots of hourly hot water demand in case study buildings.

abnormal occupant behavior on case study houses allows to assess the adaptation potential of the agent to a behavior that has never observed before.

3.2. Hyper-parameters

The RL framework include a number of hyper-parameters that should be selected based on the specific problem and desired objectives [60]. The main hyper-parameters in this framework include specifications of agent (e.g. Learning rate, Batch size, Update frequency, Memory), weights of the reward function, and also the length of look-back vector for some specific states that are expected to have a higher importance for the target system to be controlled. The look-back vectors that are of specific interest in this study are the number of previous hours of hot water demand, PV power production and indoor air temperature to be included in the state vector. Regarding that for each set of hyper-parameters all the phases of offline training, online training and deployment should be repeated, only a few of hyper-parameters could be evaluated. Therefore, the hyper-parameters of the agent are selected based on the experience from our primary study [50], as presented in Table 3.

One of the important aspects of RL is the trade-off between exploration and exploitation [23]. To maximize the reward, the agent tries to select actions that has previously experienced and are expected to return a higher reward, which is called exploitation. On the other hand, it is still possible that the action with expected highest reward would not be the best action, so it is required that sometimes the agent randomly selects an action during the training phase to better explore the environment, which is called exploration. One of the commonly-used methods to make a balance between exploration and exploitation is the ϵ-greedy method, in which a small probability of ϵ is specified and the agent performs exploration when a random value between 0 to 1 would be higher than specified value for ϵ. In this study, it is desired that during the training phase the agent performs higher exploration (more random actions) at the beginning and then gradually reduces the exploration to near zero. Therefore, a linear decay is established for exploration, where the ϵ linearly decays from 0.9 to 0.0001 at each time step over the first 12 weeks.

Table 4
Selected weights for reward function.

Weight	Associated term	Value
a	Energy	1
b	Hot water use comfort	20
c	indoor air temperature comfort	10
d	Hygiene	10

Weights of the reward function are selected based on the relative importance of each term in the reward. The selected weights are indicated in Table 4. A weight of 1 is selected for energy term, because it is multiplied by the net energy usage, which is in the range of 0–4 kW. The agent is supposed to reduce energy usage, without violating the comfort and hygiene aspects. Higher weights are selected for these terms to highly penalize the agent if any of these aspects are violated. The weight of hot water comfort is a bit higher that of space heating, because the hot water use behavior is more stochastic and therefore the agent should be more conservative towards the hot water use comfort.

3.3. Reward evolution

The evolution of reward over the training phase should be monitored to evaluate if the agent has found an optimal control policy to minimize reward function. Fig. 11 presents the weekly-averaged reward over the offline training, as well as online training on each of the houses. It should be noted that energy reward in this framework is not avoidable, and therefore, depending on the heat pump capacity, variations of reward function up to −5 are due to the power use of heat pump. Considering the weights presented in Table 4, reward values lower than −10 (more negative values) indicate that the comfort or hygiene terms are also violated. As can be seen from the first diagram, there are 5 periods during the offline training phase, where the value of reward reaches to −10 or below. In these periods, the agent has been trying to minimize the energy reward by turning OFF the heat pump, but however due to a low hot water tank temperature it has violated comfort or hygiene terms. After each violation and receiving a high penalty, it has learned that it should increase energy usage to avoid the violation of other terms. After the last violation around week 377, reward value is almost stable. The value of reward function during the online training on the target houses is always above -10, and shows a good stability. This indicates that the agent has gained enough experience during the intensive offline training phase, which has guaranteed an optimal policy since the first week of training phase on each target house. The fast convergence on the target houses, in spite of abnormal hot water use behavior, shows that the variations included in the offline training phase (variations of the system sizes, hot water use pattern, weather conditions, etc.) have provided a generalizable knowledge for the agent, and ensured the transferability to the other houses.

3.4. Visual assessment of the proposed framework

3.4.1. Performance of the agent during the offline training

As shown in Figs. 6 and 7, offline training phase was performed for 10 years, each year on a different city, and with different sizes of system. It is interesting to have a closer look at the offline training phase to see if the agent could preserve the occupant comfort with such variations. Fig. 12 presents the boxplots of hot water tank and indoor air temperatures during the offline training phase. It can be seen that

Table 5
Comparison of performance between RL-OSD and rule-based control methods in three case studies during the deployment phase.

	House 1			House 2			House 3		
	RL-OSD	RC	RE	RL-OSD	RC	RE	RL-OSD	RC	RE
Energy use (MWh)	1.14	1.6	1.23	0.73	1.22	0.8	0.06	0.24	0.15
Violation of DHW comfort (%)	8.1	0	0	5	0	0	1.7	0	0
Average temperature of DHW comfort violations (°C)	38.9	–	–	39	–	–	38	–	–
Number of space heating comfort violations (h)	153	0	0	84	0	0	29	0	0
Average temperature of space heating comfort violations (°C)	24.3	–	–	22.2	–	–	23.5	–	–
Average heat pump COP	2.71	2.3	2.68	2.7	2.3	2.75	3.6	3.1	4

Table 6
Summary of performance of Long-time deployment scenario (RL-OLD) with other control methods.

	RL-OLD	RC	RE
Energy use (MWh)	5.17	7.6	5.4
Violation of DHW comfort (%)	5.3	0	9.2
Average temperature of DHW comfort violations (°C)	38.7	0	38.5
Number of space heating comfort violations (h)	1015	548	532
Average temperature of space heating comfort violations (°C)	23.8	19.8	19.9

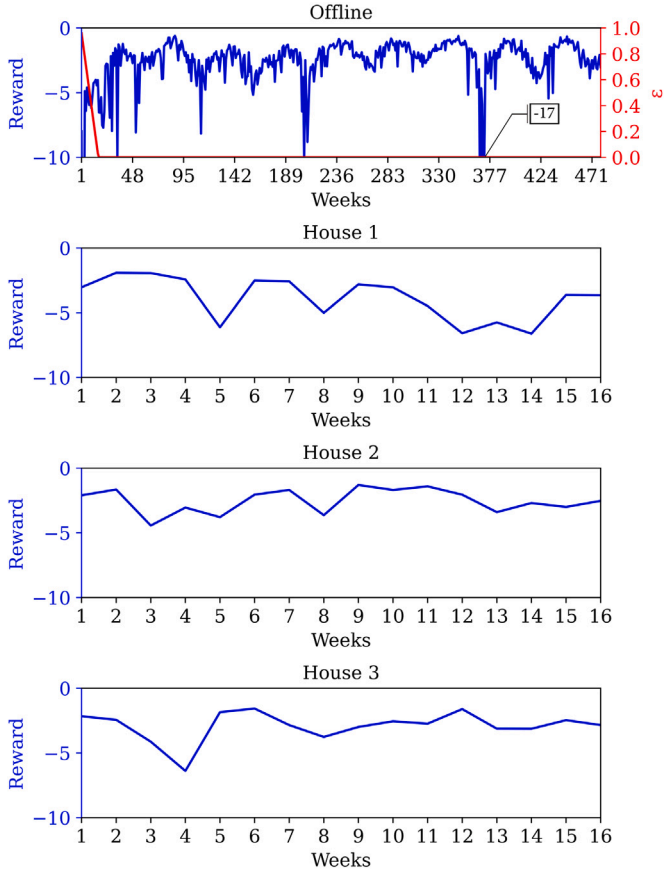


Fig. 11. Evolution of reward over the offline training stage and online training stages in each house.

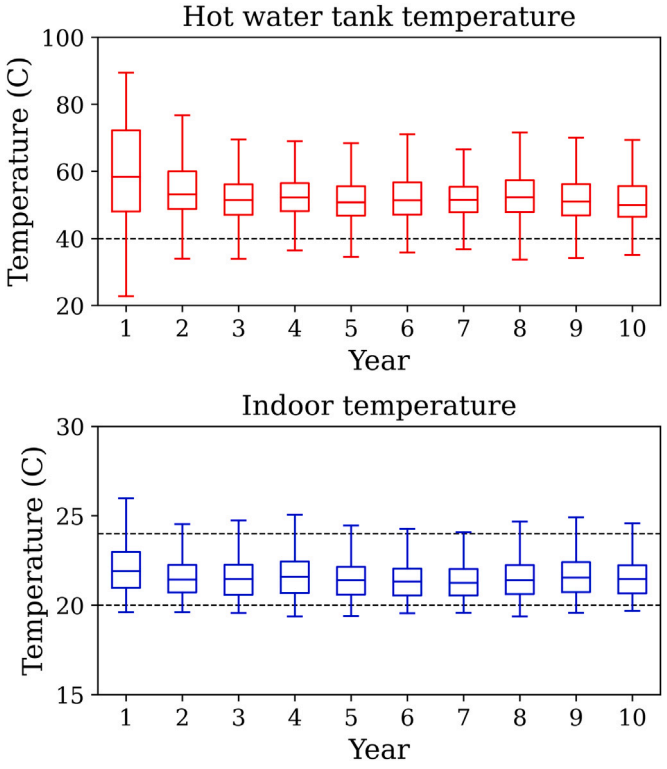


Fig. 12. Boxplots of yearly hot water and indoor air temperature versus comfort limits.

there is a higher variance in hot water and indoor air temperatures over the first year, which is due to the lack of experience by the agent, as well as performing random actions during the exploration phase. From the second year, the hot water and indoor air temperatures show a lower variance, close to the comfort limits, which indicates that in only few hours the occupant comfort is slightly violated. Also the average hot water and indoor air temperatures are higher over the first year. It shows that at the beginning the agent has been trying to preserve occupants comfort by spending more energy, but from the second year it has learned to further reduce temperatures and save more energy while respecting occupant comfort. Overall, from this figure it can be seen that although several parameters (weather, solar radiation, occupant behavior and system sizes) vary from year to year in the environment, the agent performance is stable since the second year. This indicates the adaptation potential of RL to the potential variations that can happen from building to building in a wide-spread implementation.

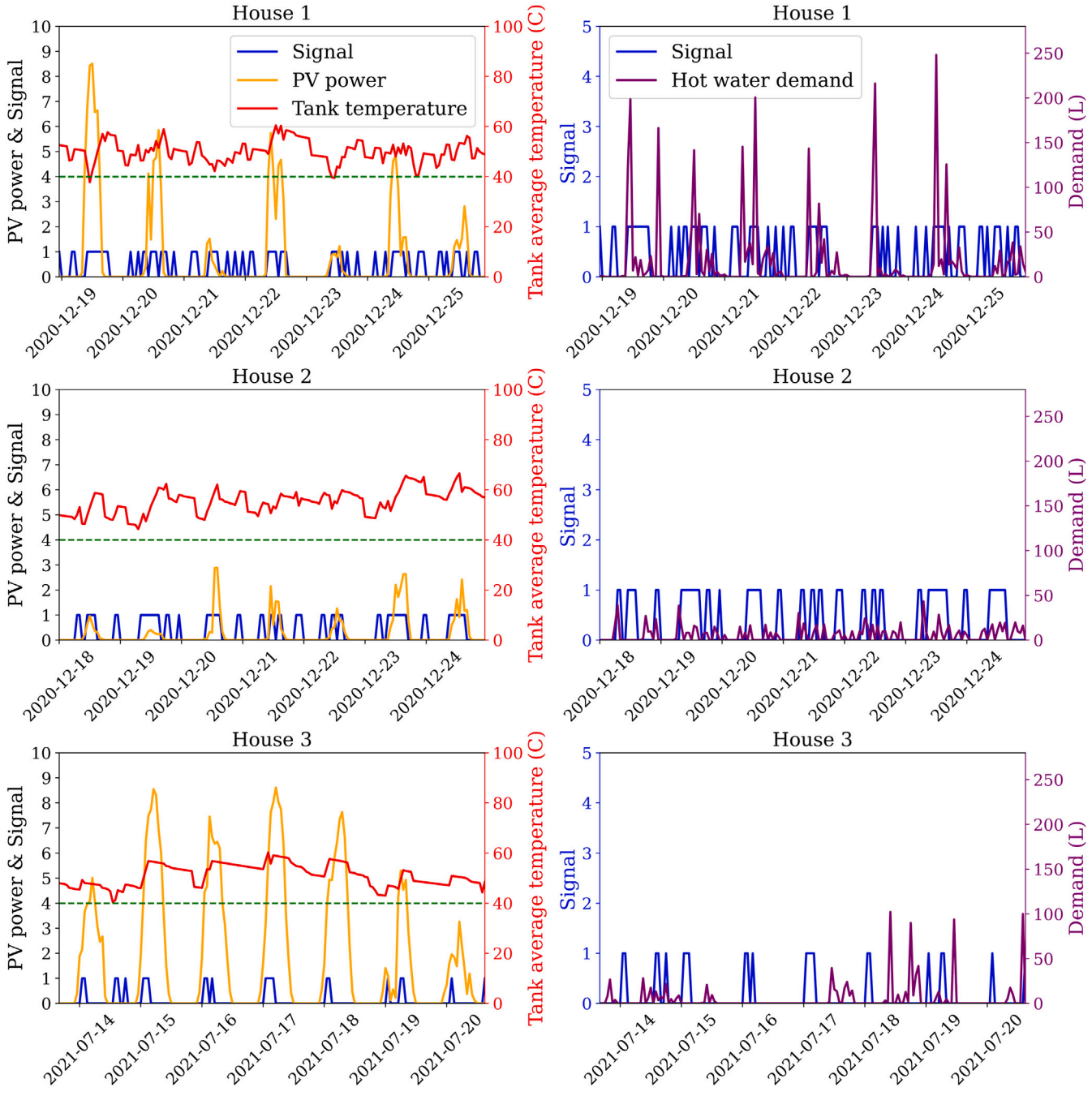


Fig. 13. Adaptation of control signal to the PV power production and hot water demand in RL-OSD scenario.

3.4.2. Performance of the RL-OSD

A major capability of the RL agent is adaptation to stochastic parameters, which in this problem are mainly PV power production and hot water demand. To visualize the adaptation potential of the agent, Fig. 13 presents the control signal versus PV power production and hot water demand. As can be seen in this Figure, the agent mostly turns ON the heat pump when PV power is available. This is more clear in case of House 3. In this house, the deployment phase has been during the summer, with a higher PV power production and lower energy demand, which enables the agent to harvest most of the required energy from PV panels. Hot water tank temperature is also visualized to assess how the agent has adapted to the hot water use behavior to preserve the comfort aspect. It can be seen that the agent has successfully learned the hot water use behavior, because even in case of high volume demands, e.g. in House 1, the agent has always kept the hot water tank temperature above the comfort limit.

Previous studies have usually used the self-consumption of PV power as an evaluation metric for their proposed control approach [61]. However, it should be noted that in this study a higher self-consumption can be caused by the higher energy use of heat pump, e.g. by operating with a lower COP, which is not desired. Rather, in this study the share of PV power production in total power consumption of heat pump is

Table 7
Comparison of performance between RL-OSD and RL-DD in three houses during the deployment phase.

	House 1		House 2		House 3	
	RL-OSD	RL-DD	RL-OSD	RL-DD	RL-OSD	RL-DD
Energy use (MWh)	1.14	1.05	0.73	0.59	0.06	0.14
Violation of DHW comfort (%)	8.14	17.1	5	3.9	1.7	0
Average temperature of DHW comfort violations (°C)	38.9	37.9	38.9	30.8	38	–
Number of space heating comfort violations (h)	153	142	84	126	29	136
Average temperature of space heating comfort violations (°C)	24.3	24.2	22.2	24.2	23.5	24.9

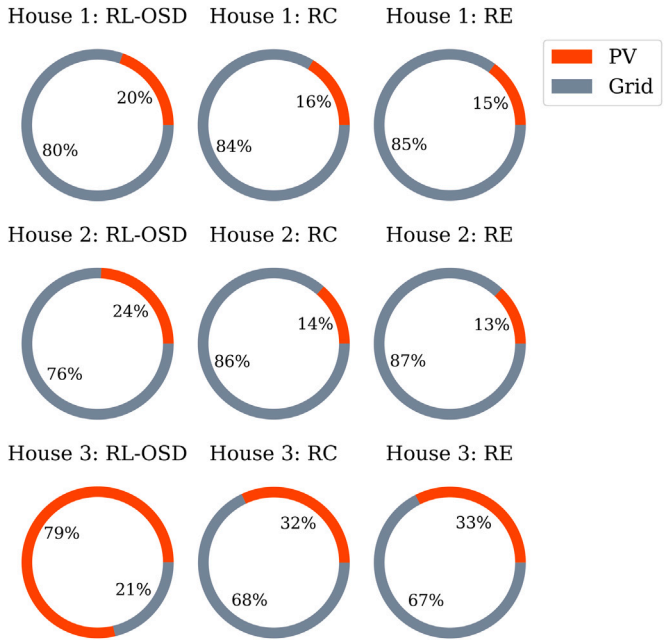


Fig. 14. Contribution of PV power production in heat pump power consumption in RL-OSD scenario.

used for comparison. As shown in Fig. 14, in all Houses, RL-OSD has obtained a higher share of energy consumption from PV panels. This share is much higher in case of House 3, as the deployment phase in this house has been during the summer, with much higher PV power production and lower energy demand. These results indicate that the proposed framework would provide a higher energy saving in regions with a high solar radiation.

To evaluate the comfort aspect during the deployment phase, boxplots of indoor air and hot water tank temperatures by different control methods are shown for House 1 in Fig. 15. Due to the high number of plots only one house is presented, and the other houses show a similar performance. Regarding that the range of comfort in case of indoor air temperature is quite narrow, it can be easily violated. Therefore, all of the methods show some violations of comfort. RL-OSD shows more violations than the rule-based methods, but the violations are less than 2 °C and happen in few hours, which therefore can be ignored. In case of the hot water tank temperature, similarly, RL-OSD shows very slight violations that can be ignored. Interestingly, violations by RL-OSD are even less than RE, which is due to the fact that RL-OSD tries to save energy by adapting to the occupant behavior, while RE tries to do so only by lowering the hot water tank temperature, regardless of occupant behavior.

Boxplots of Legionella concentration over the deployment phase are shown in Fig. 16. As expected, both of rule-based methods maintain a lower concentration than the RL-OSD method, because they are over-conservative. While RL-OSD method is less conservative, it has always respected the hygiene aspect as the maximum concentration is less than 4500 CFU/L, which is much less than the risky limit of 500,000 CFU/L, placed for single-family residential houses [44]. It shows that RL-OSD has learned to maintain hygiene aspect while avoiding over-necessary heating of the tank.

3.4.3. Performance of RL-OLD

Fig. 17 presents the performance of the agent over the long-time deployment (RL-OLD scenario). As can be seen, there are a lot of variations in hot water use behavior of occupants over this period, including a sudden decrease for one month, and an absence period. Also this period includes a good diversity in outdoor air temperature, as it includes cold months at the beginning and hot months at the end. These diversities are valuable to assess how the agent will adapt to the possible changes in environment over a longtime deployment. As shown in this Figure, Although there are significant variations in hot water use behavior, the agent has always kept the hot water tank temperature above comfort temperature of 40 °C. There is an increase in the temperature of pressurized hot water tank from the middle of May (2021–05). This is because in this period there is a higher PV power production, a lower demand for space heating, and at the same time a sudden decrease in hot water demand. Therefore, hot water tank temperature is increased as the agent is trying to get the best use of PV power production by storing the surplus energy in the hot water tank. Indoor air temperature is also within the comfort limit, with slight violations of less than 2 °C. Legionella concentration is also always below the risky limit, while it is higher during the cold season and lower during the warm season, when extra energy is stored by over-heating the tank.

3.4.4. Performance of RL-DD

Fig. 18 compares the performance of the agent which is also trained on the target house (RL-OSD), versus the agent which is only trained offline and has never observed the behavior of that specific house (RL-DD). The training phase of RL includes some randomness, mainly due to the exploration phase. Therefore, even if two agents are trained with exactly same specifications, they can have slightly different performances. So the slight differences between these two agents should not be considered as the consequence of the lack of an online training phase in RL-DD. The two agents, thus, show very similar performance on Houses 1 and 2. The only significant difference is on House 3, where

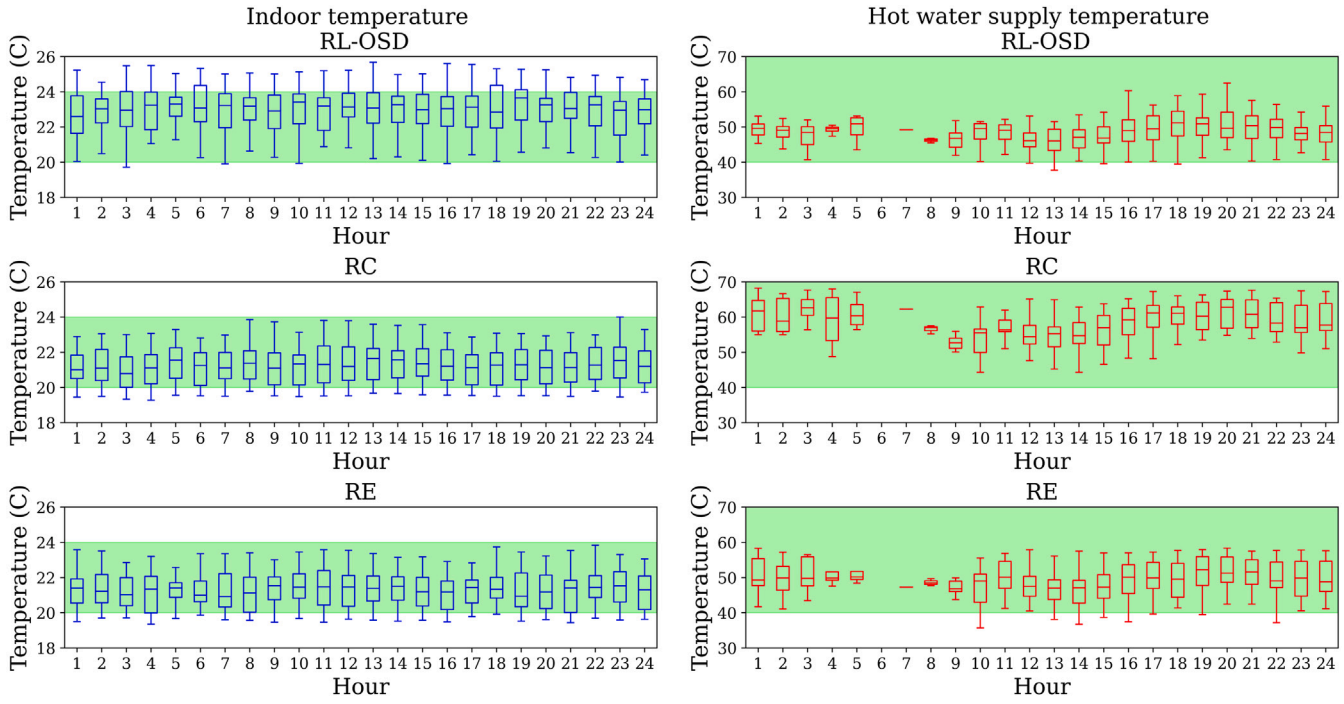


Fig. 15. Boxplots of indoor air and hot water tank temperatures by three control methods in House 1 in RL-OSD scenario.

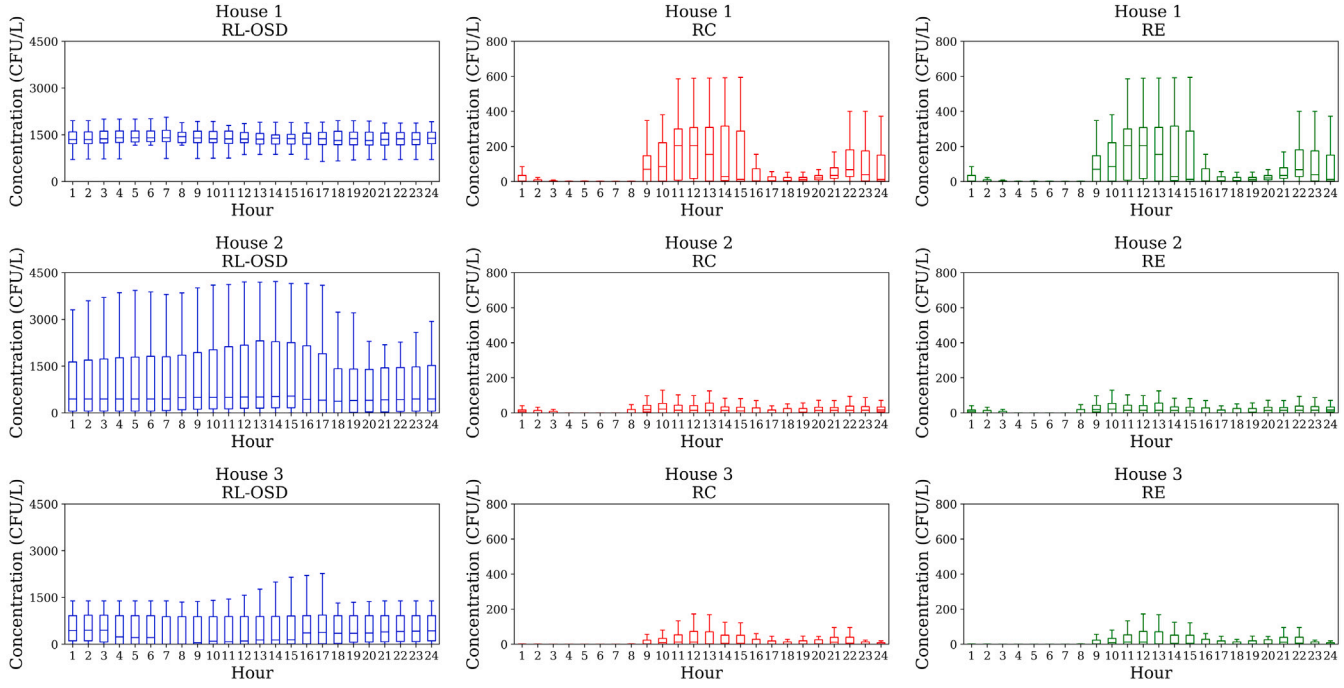


Fig. 16. Boxplots of Legionella concentration in tank by three control methods over three case studies.

the RL-DD agent has kept a higher hot water tank temperature than the RL-OSD agent. This is because the RL-OSD has observed and learned the specific behavior of occupants on House 3, and is better adapted to their behavior than the offline trained agent whci has only observed the stochastic-based hot water use behavior. These diagrams show that while the RL-DD agent has never seen the specific parameters of the target house (weather conditions, occupant' behavior, etc.), it can still maintain the comfort and hygiene aspects.

3.5. Quantified assessment of the proposed framework

3.5.1. RL-OSD versus baseline methods

Table 5 presents the performance metrics of RL-OSD versus RC and RE methods during the deployment phase. RL-OSD consumes least energy in all of the Houses, with the lowest energy use on House 3 where a higher PV power was available. In this house, RL-OSD has provided an energy saving of 60% compared to the RE method, indicating the great potential of RL-OSD for seasons with a high solar energy potential. To

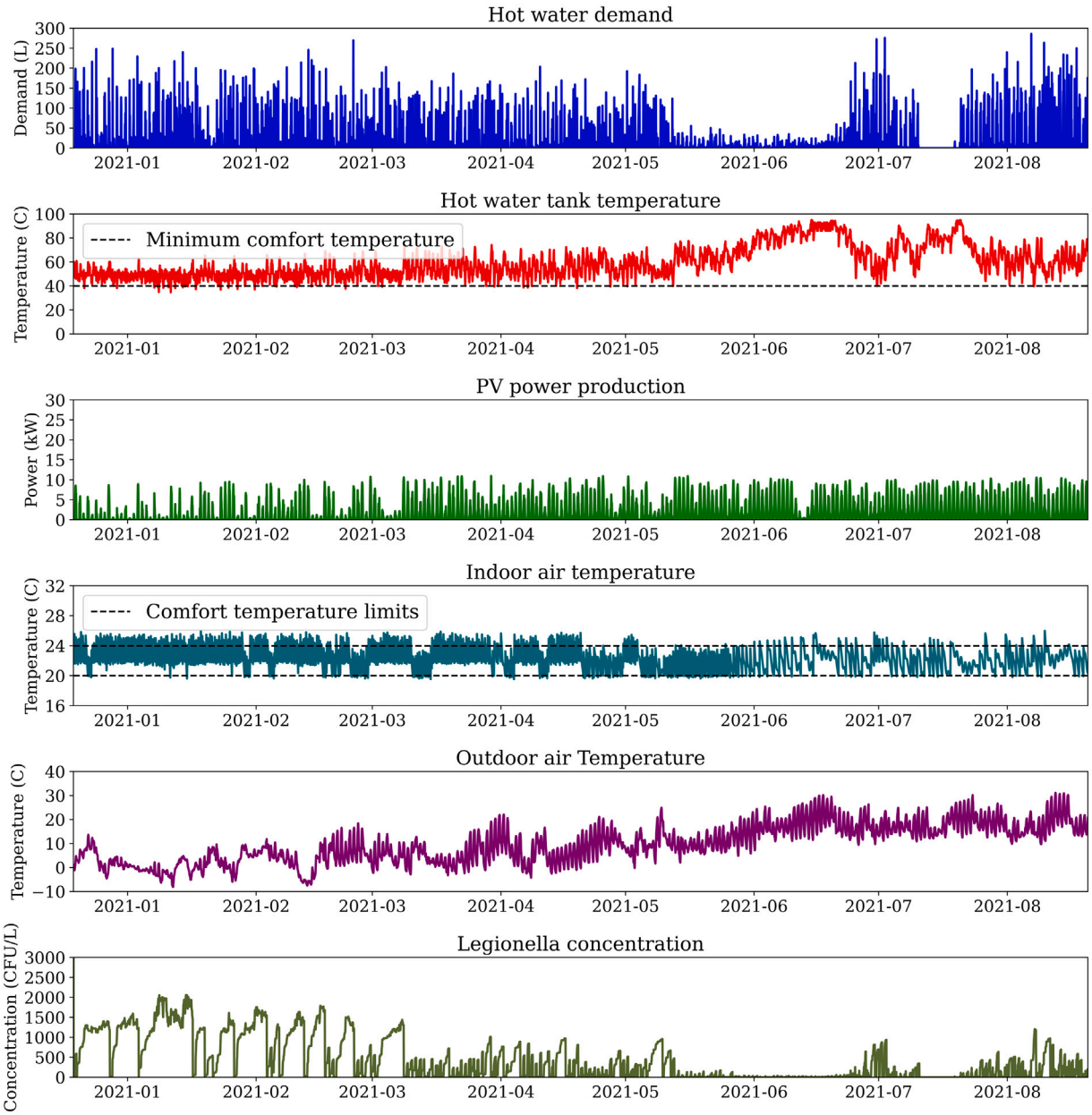


Fig. 17. Performance of the RL-OLD agent during long-time deployment on House 1.

quantify the comfort aspect of the RL-OSD method, the percentage of total hot water demand which was met with a temperature less than comfort level is also indicated in this table. At the worst case, which has happened in House 1, 8% of total demand is violated. The average temperature of these violations is 38.9 °C, which is very close to the comfort limit of 40 °C. In case of space heating, although RL-OSD has violated the comfort limits during a few hours, the average of violations is very close to the comfort limits. In Houses 2 and 3, this average is in comfort limits because some of the violations has been less than 20 °C and some other more than 24 °C. It can be therefore considered that in all of the houses RL-OSD has properly maintained the occupant comfort. The average COP of heat pump by RL-OSD is always equal or lower than by RE. It proves that the energy saving by RL-OSD is not only achieved by lowering the hot water tank temperature (and therefore increasing the COP), but by properly scheduling of heating cycles to profit more from PV power production.

3.5.2. *RL-OLD versus baseline methods*

Table 6 presents the metrics of RL-OLD scenario. These metrics show that over the long-time, even without any other online training, the agent has provided an energy saving while maintaining the occupant comfort and water hygiene. This scenario was presented to prove the performance of the trained agent over a long time deployment without any further online training. But if it is technically possible in practice, sequential or continuous training of the agent will probably provide a higher energy saving.

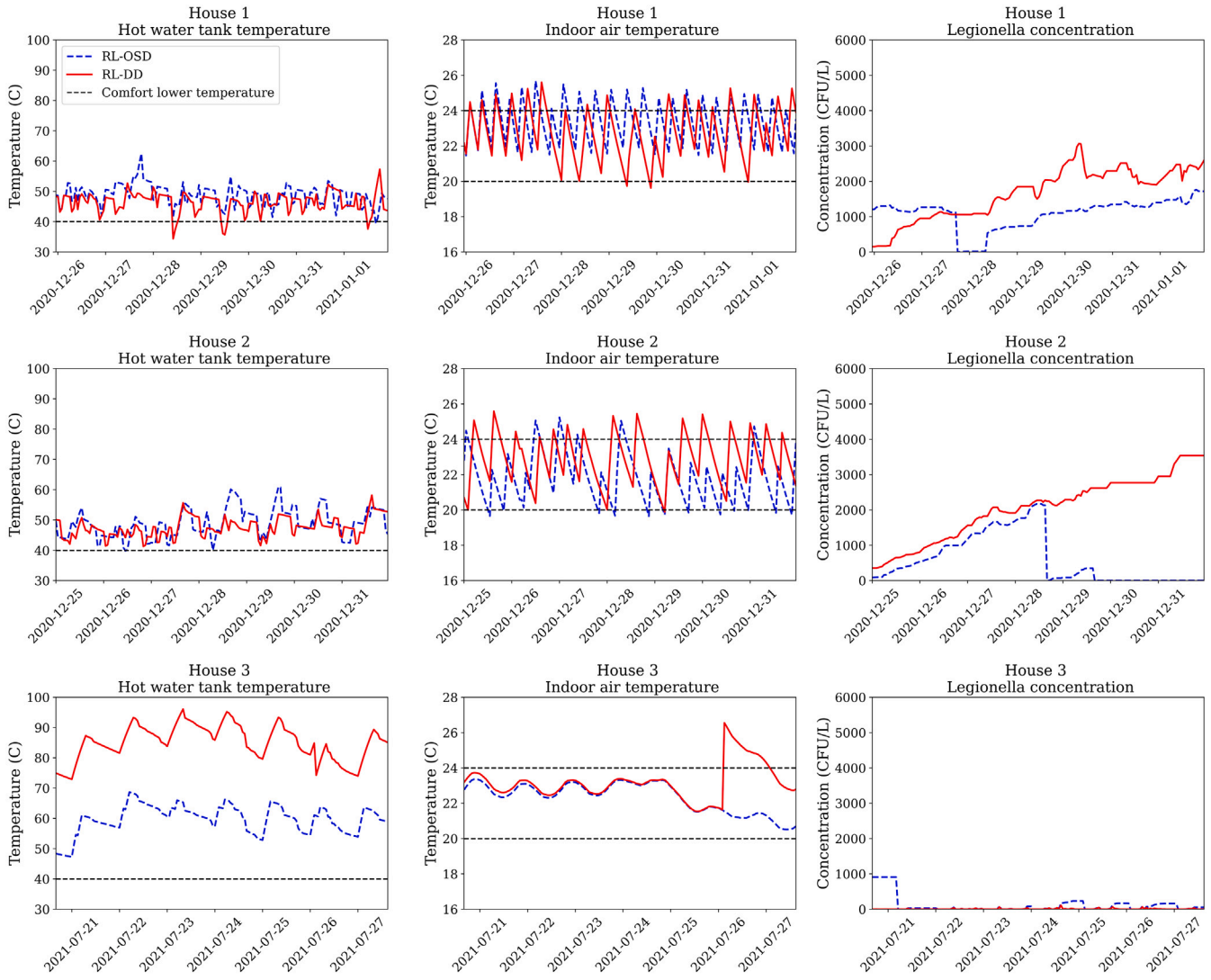


Fig. 18. Performance of RL-DD agent during direct deployment.

3.5.3. *RL-OSD versus RL-DD*

Table 7 represents the energy use and comfort metrics between RL-OSD versus RL-DD. The performance of two scenarios are quite similar in Houses 1 and 2, while RL-OSD performs better in House 3. This shows that the conditions (occupant behavior, weather conditions, etc.) of Houses 1 and 2 are more similar to what the agent has observed during the offline training phase, while House 3 has quite different conditions from offline training. Therefore, in this house an online training has further improved the agent performance in RL-OSD.

3.6. *Conclusion*

There are several stochastic parameters such as occupant behavior, renewable energy potential, and weather condition, that increase the complexity of developing an optimal control method for residential energy systems. Among them, occupant behavior is of significant concern, as it is highly stochastic, specific to each building, varies in time, and therefore very challenging to model and predict. This study proposes a data-driven and model-free control method based on Reinforcement Learning, that can learn these stochastic parameters by itself, and maintain an optimal operation. The agent in this framework also takes into account the hygiene aspect of hot water and learns how to save energy saving while maintaining the water hygiene. The goal of the learning agent is to save energy while maintaining the health and comfort of occupants. The energy system evaluated in this study was a PV-assisted air-source heat pump for space heating and hot water production, though this framework is easily adjustable to other systems. A two-step training method is proposed, including an offline phase integrating stochastic hot water use behavior to provide an initial experience for the learning agent, and an online phase to learn and adapt to the behavior of the target house. The framework was evaluated for three houses in different regions of Switzerland. For these case study houses, weather and solar radiation data were collected from nearby weather stations, and hot water use data was experimentally monitored to evaluate the framework on the real-world behavior of occupants. The following main conclusions can be drawn from this study:

- The proposed framework (RL-OSD scenario) achieved 7% to 60% energy-saving compared to an energy saving rule-based method (RE), and 28% to 75% compared to the common practice rule-based method (RC), without violating the occupant comfort and water hygiene.
- The agent properly learned the variations of PV power production in each building and adapted the heating cycles to the PV power production to get the best use of free solar energy (As can be seen in Fig. 13). The proposed framework could therefore provide a substantial energy saving in House 3, where a higher PV power production was available.

- Evaluation of direct deployment scenario (RL-DD) indicated that the stochastic-based intensive offline training provides a generalizable knowledge for the agent, and therefore it could still outperform rule-based methods even without any online training on the target houses. As expected, the agent that was also trained on the target houses (RL-OSD scenario) indicated slightly better performance. It shows that, if enough computational power is available, the stochastic-based offline training can be further extended by including many possible conditions that can happen in reality (e.g., a sudden change in occupant behavior and weather conditions, change of system components, etc.), which makes it possible to directly implement the trained agent on several houses without any need for online training. It will significantly facilitate the transferability of the proposed framework to other buildings.
- Evaluation of long-time deployment scenario (RL-OLD) indicated that the agent could provide a satisfactory performance over a long time, and further sequential or continuous training is not necessary, which would further facilitate the experimental implementations.

With the increasing complexity of residential energy systems, rather than hard-programming the expert knowledge as a rule-based or model-based control method, it is possible to let the agent to learn the optimal control method by itself in each specific building . In this study, experimentally measured data was used in simulations to provide a realistic while safe environment to perform a primary test of the agent performance. Considering the promising results of this step, in next step of this research project, a Reinforcement Learning control framework will be experimentally implemented to observe the agent performance in practice, and address the technical challenges in the wide-spread implementation of Reinforcement Learning-based control methods in practice.

CRediT authorship contribution statement

Amirreza Heidari: Conceptualization, Methodology, Software, Validation, Visualization, Writing – original draft. **François Maréchal:** Methodology, Supervision, Writing – review & editing. **Dolaana Khovalyg:** Funding acquisition, Project administration, Resources, 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.

References

[1] Majcen D. Predicting energy consumption and savings in the housing stock. A+BE| Archit Built Environ 2016;(4):1–224.

[2] Sun K, Hong T. A simulation approach to estimate energy savings potential of occupant behavior measures. Energy Build 2017;136:43–62.

[3] Gill ZM, Tierney MJ, Pegg IM, Allan N. Low-energy dwellings: the contribution of behaviours to actual performance. Build Res Inform 2010;38(5):491–508. http://dx.doi.org/10.1080/09613218.2010.505371, arXiv:https://doi.org/10.1080/09613218.2010.505371.

[4] Han M, Zhao J, Zhang X, Shen J, Li Y. The reinforcement learning method for occupant behavior in building control: A review. Energy Built Environ 2021;2(2):137–48.

[5] Li J, Yu ZJ, Haghighat F, Zhang G. Development and improvement of occupant behavior models towards realistic building performance simulation: A review. Sustainable Cities Soc 2019;50:101685.

[6] Hong T, Chen Y, Belafi Z, D'Oca S. Occupant behavior models: A critical review of implementation and representation approaches in building performance simulation programs. In: Building simulation, Vol. 11. Springer; 2018, p. 1–14.

[7] Harputlugil T, de Wilde P. The interaction between humans and buildings for energy efficiency: A critical review. Energy Res Soc Sci 2021;71:101828.

[8] Yue T, Long R, Chen H. Factors influencing energy-saving behavior of urban households in Jiangsu Province. Energy Policy 2013;62:665–75. http://dx.doi.org/10.1016/j.enpol.2013.07.051, URL https://www.sciencedirect.com/science/article/pii/S0301421513006940.

[9] REmap 2030: A renewable energy roadmap. 2014.

[10] Grid-integrated distributed solar: addressing challenges for operations and planning. 2016.

[11] Dengiz T, Jochem P, Fichtner W. Impact of different control strategies on the flexibility of power-to-heat-systems. In: Transforming energy markets, 41st IAEE international conference, Jun 10-13, 2018. International Association for Energy Economics; 2018.

[12] Kondziella H, Bruckner T. Flexibility requirements of renewable energy based electricity systems–a review of research results and methodologies. Renew Sustain Energy Rev 2016;53:10–22.

[13] Sethi M, Tripathi R, Pattnaik B, Kumar S, Khargotra R, Chand S, Thakur A. Recent developments in design of evacuated tube solar collectors integrated with thermal energy storage: A review. Mater Today: Proc 2021.

[14] https://www.waermepumpe.de/presse/zahlen-daten/, accessed: 2021-10-1.

[15] Leppin L. Development of operational strategies for a heating pump system with photovoltaic, electrical and thermal storage. 2017.

[16] Camacho EF, Alba CB. Model predictive control. Springer science & business media; 2013.

[17] Fiorentini M, Wall J, Ma Z, Braslavsky JH, Cooper P. Hybrid model predictive control of a residential HVAC system with on-site thermal energy generation and storage. Appl Energy 2017;187:465–79.

[18] Halvgaard R, Poulsen NK, Madsen H, Jørgensen JB. Economic model predictive control for building climate control in a smart grid. In: 2012 IEEE PES innovative smart grid technologies (ISGT). IEEE; 2012, p. 1–6.

[19] Mady AE-D, Provan G, Ryan C, Brown K. Stochastic model predictive controller for the integration of building use and temperature regulation. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 25; 2011.

[20] Smarra F, Jain A, De Rubeis T, Ambrosini D, D'Innocenzo A, Mangharam R. Data-driven model predictive control using random forests for building energy optimization and climate control. Appl Energy 2018;226:1252–72.

[21] Hosseinloo AH, Ryzhov A, Bischi A, Ouerdane H, Turitsyn K, Dahleh MA. Data-driven control of micro-climate in buildings: An event-triggered reinforcement learning approach. Appl Energy 2020;277:115451.

[22] Schreiber T, Netsch C, Eschweiler S, Wang T, Storek T, Baranski M, Müller D. Application of data-driven methods for energy system modelling demonstrated on an adaptive cooling supply system. Energy 2021;230:120894.

[23] Brandi S, Piscitelli MS, Martellacci M, Capozzoli A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. Energy Build 2020;224:110225.

[24] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT Press; 2018.

[25] 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.

[26] Liu S, Henze GP. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 2: Results and analysis. Energy Build 2006;38(2):148–61.

[27] 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.

[28] Cheng Z, Zhao Q, Wang F, Jiang Y, Xia L, Ding J. Satisfaction based Q-learning for integrated lighting and blind control. Energy Build 2016;127:43–55.

[29] Zou Z, Yu X, Ergan S. Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. Build Environ 2020;168:106535.

[30] Valladares W, Galindo M, Gutiérrez J, Wu W-C, Liao K-K, Liao J-C, Lu K-C, Wang C-C. Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. Build Environ 2019;155:105–17.

[31] 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.

[32] Heidari A, Maréchal F, Khovalyg D. An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: a model-free reinforcement learning approach. Applied Energy 2022;312:118833.

[33] Correa-Jullian C, Droguett EL, Cardemil JM. Operation scheduling in a solar thermal system: A reinforcement learning-based framework. Appl Energy 2020;268:114943.

[34] Ali A, Kazmi H. Minimizing grid interaction of solar generation and DHW loads in nZEBs using model-free reinforcement learning. In: International workshop on data analytics for renewable energy integration. Springer; 2017, p. 47–58.

[35] Lissa P, Deane C, Schukat M, Seri F, Keane M, Barrett E. Deep reinforcement learning for home energy management system control. Energy AI 2021;3:100043.

[36] AlphaZero: Shedding new light on chess, shogi, and go. 2022, https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go, accessed: 2022-10-2.

[37] Van Kenhove E, De Backer L, Janssens A, Laverge J. Simulation of Legionella concentration in domestic hot water: comparison of pipe and boiler models. J Buil Perform Simul 2019;12(5):595–619.
[38] Booysen M, Engelbrecht J, Ritchie M, Apperley M, Cloete A. How much energy can optimal control of domestic water heating save? Energy Sustain Dev 2019;51:73–85.
[39] Mirnaghi M, Panchabikesan K, Haghighat F. Application of data mining in understanding the charging patterns of the hot water tank in a residential building: a case study. In: IOP conference series: Materials science and engineering, Vol. 609. IOP Publishing; 2019, 052038.
[40] Carlson KM, Boczek LA, Chae S, Ryu H. Legionellosis and recent advances in technologies for Legionella control in premise plumbing systems: a review. Water 2020;12(3):676.
[41] Krawczyk M, Petruzzelli M, et al. Legionella 2003: An update and statement by the association of water technologies. Assoc Water Technol 2003;26.
[42] Taghdiri S. Airborne dispersion and plume modeling of Legionella bacteria. Arizona State University; 2014.
[43] Sharaby Y, Rodríguez-Martínez S, Oks O, Pecellin M, Mizrahi H, Peretz A, Brettar I, Höfle MG, Halpern M. Temperature-dependent growth modeling of environmental and clinical Legionella pneumophila multilocus variable-number tandem-repeat analysis (MLVA) genotypes. Appl Environ Microbiol 2017;83(8):e03295–16.
[44] van Amerongen G, Lee J, Suter J-M. Legionella and solar water heaters. 2013.
[45] Van Kenhove E, De Backer L, Delghust M, Laverge J. Coupling of modelica domestic hot water simulation model with controller. In: Building simulation 2019, 16th IBPSA international conference and exhibition, Vol. 16. International Building Performance Association (IBPSA); 2020, p. 924–31.
[46] Ryu M, Chow Y, Anderson R, Tjandraatmadja C, Boutilier C. CAQL: Continuous action Q-learning. 2019, arXiv preprint arXiv:1909.12397.
[47] Quillen D, Jang E, Nachum O, Finn C, Ibarz J, Levine S. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. In: 2018 IEEE international conference on robotics and automation (ICRA). IEEE; 2018, p. 6284–91.
[48] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30; 2016.
[49] Tensorforce: a TensorFlow library for applied reinforcement learning. 2021, URL https://tensorforce.readthedocs.io/en/0.6.5/index.html.
[50] Heidari A, Marechal F, Khovalyg D. An adaptive control framework based on reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems. In: Journal of physics: Conference series, Vol. 2042. IOP Publishing; 2021, 012006.
[51] Gelažanskas L, Gamage KA. Forecasting hot water consumption in dwellings using artificial neural networks. In: 2015 IEEE 5th international conference on power engineering, energy and electrical drives (POWERENG). IEEE; 2015, p. 410–5.
[52] Delorme-Costil A, Bezian J-J. Forecasting domestic hot water demand in residential house using artificial neural networks. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). IEEE; 2017, p. 467–72.
[53] for Standardization IO. ISO 7730 2005-11-15 ergonomics of the thermal environment: Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria. International standards, ISO; 2005, URL https://books.google.ch/books?id=p3YcoAEACAAJ.
[54] Ritchie M, Engelbrecht J, Booysen M. A probabilistic hot water usage model and simulator for use in residential energy management. Energy Build 2021;235:110727.
[55] Organization WH, et al. WHO housing and health guidelines. 2018.
[56] Ormandy D, Ezratty V. Health and thermal comfort: From WHO guidance to housing strategies. Energy Policy 2012;49:116–21.
[57] Quero S, Párraga-Niño N, Garcia-Núñez M, Pedro-Botet ML, Gavaldà L, Mateu L, Sabrià M, Mòdol JM. The impact of pipeline changes and temperature increase in a hospital historically colonised with Legionella. Sci Rep 2021;11(1):1–7.
[58] Gooroochurn M, Visram A. Maximization of solar hot water production using a secondary storage tank. J Clean Energy Technol 2019;7(1).
[59] Melius J, Margolis R, Ong S. Estimating rooftop suitability for PV: a review of methods, patents, and validation techniques. 2013.
[60] Zhang Z, Chong A, Pan Y, Zhang C, Lu S, Lam KP. A deep reinforcement learning approach to using whole building energy model for hvac optimal control. In: 2018 building performance analysis conference and simbuild. 3, 2018, p. 22–3.
[61] Vanhoudt D, Geysen D, Claessens B, Leemans F, Jespers L, Van Bael J. An actively controlled residential heat pump: Potential on peak shaving and maximization of self-consumption of renewable energy. Renew Energy 2014;63:531–43.