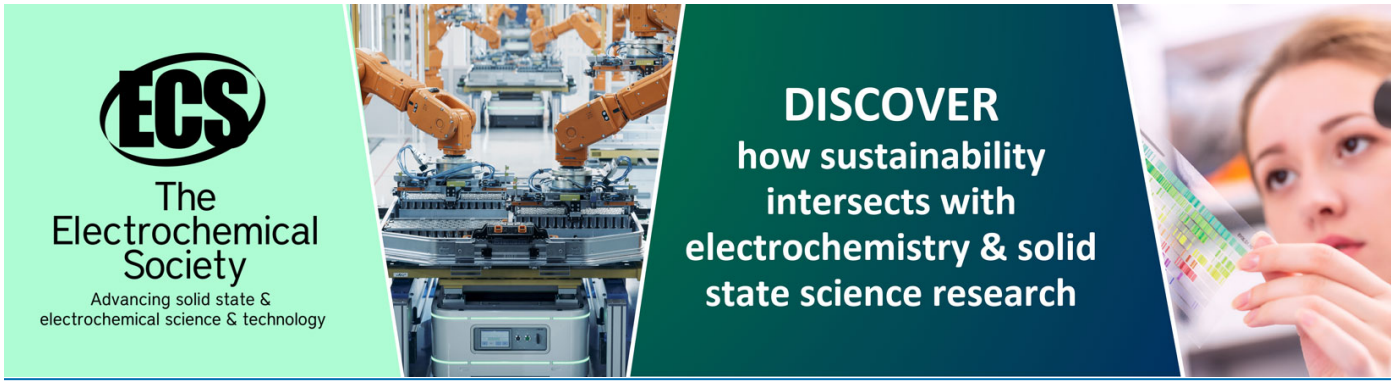
# An adaptive control framework based on Reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems





# An adaptive control framework based on Reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems

**Amirreza Heidari, Francois Marechal, Dolaana Khovalyg**
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
E-mail: amirreza.heidari@epfl.ch

**Abstract.** A major challenge in the operation of water heating systems lies in the highly stochastic nature of occupant behavior in hot water use, which varies over different buildings and can change over the time. However, the current operational strategies of water heating systems are detached from occupant behavior, and follow a conservative and energy intensive approach to ensure the availability of hot water any time it is demanded. This paper proposes a Reinforcement learning-based control framework which can learn and adapt to the occupant behavior of each specific building and make a balance between energy use, occupant comfort and water hygiene. The proposed framework is compared to the conventional approach using the real-world measurements of hot water use behavior in a single family residential building. Although the monitoring campaign has been executed during home lockdown due to COVID-19, when the occupants exhibited a very different schedule and water use related behavior, the proposed framework has learned the occupant behavior over a relatively short period of 8 weeks and provided 24.5% energy use reduction over the conventional approach, while preserving occupant comfort and water hygiene.

## 1. Introduction

Energy demand for hot water production has not changed notably over the generations of buildings, and therefore accounts for about 40%-50% in modern buildings [1]. Although there has been improvements in the design of water heating systems, a major challenge for energy saving is their conservative operational strategies which derives from the highly stochastic nature of hot water demand [2]. Hot water demand is strongly correlated with occupants´ behaviour and demographics [3], therefore it shows high variations between different buildings and over the time [4]. Another driver of high energy use in water heating systems is Legionella, a water-born bacteria which can grow in water with temperature between 20°C and 50°C. Accordingly, a common practise in the operation of hot water systems is to keep the storage tank temperature between 65°C to 75°C, while the required temperature for the occupant comfort is reported to be 40°C [5].

Reinforcement learning (RL) has recently gained popularity for building energy management as it can easily adapt to the time-varying occupant behavior, environmental conditions, renewable energy potential and changes in system characteristics. Although several studies have implemented RL on different aspects of building energy management, little attention has been paid to the hot water production. A related study has addressed the application of model-based RL to make a balance between energy use and comfort [6]. However, their framework was based on data-driven models, which reduces the generalization potential, and also did not consider the Legionella issue, which is a serious concern

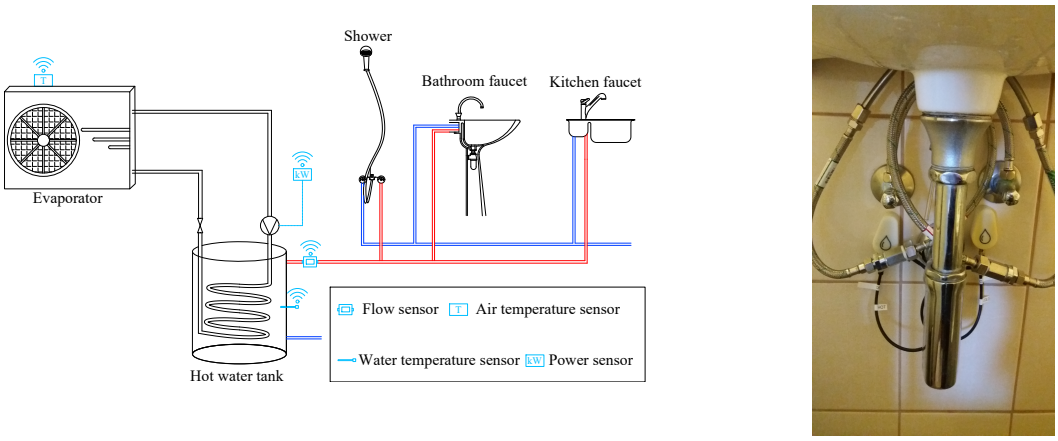
in Switzerland in recent years [7]. Therefore, this paper aims to propose a control framework for water heating systems with the following main specifications:

- **Adaptive:** Continuously adapts to the time-varying occupant behavior and environment conditions to reduce energy consumption while keeping the occupant comfort and water hygiene.
- **Model-free:** No thermodynamic or data-driven model is used within the framework to increase the transferability potential.
- **Hygiene-aware:** Takes into account the periodical sterilization of Legionella bacteria.
- **Sensor efficient:** Relies on minimum number of sensors to increase economic feasibility and robustness.
- **Off-site learning stage:** An off-site learning stage is included in the framework, to ensure the minimum disturbance of occupants comfort and to reduce learning period over the target system.
- **Double deep Q-learning:** While most of the studies are based on deep Q-learning, the proposed framework in this study is based on the double deep Q-learning, which is developed to solve the issue of overestimating action values by deep Q-learning.

## 2. Methodology

### *2.1.* ***Proposed framework***

The proposed framework aims to learn the optimal policy based on 4 different sensors, as shown in Figure 1a, including a flow sensor to monitor the hot water use behavior, a temperature sensor at the middle of the tank to monitor the tank average temperature, an air temperature sensor to monitor the ambient temperature of the evaporator, and a power sensor to monitor the energy use of the heat pump. To monitor the actual hot water use behavior, the hot water usage of a single family residential building is monitored for 14 weeks (28th August 2020 to 4th December 2020). To monitor the demand, LoraWan-based low power IoT sensors from Droople company [8] have been installed on all the water taps. Then, the flow rate of all end uses were summed to represent the flow at the tank outlet. The measurement at all end uses is to be used for future research, and for the proposed framework a single sensor at the tank outlet would be enough. The system to be controlled (referred to as ***environment*** in RL literature) in this study, which includes an air source heat pump and storage tank, was modeled in TRNSYS. The RL agent is a double deep Q-network developed in Python and integrated to the TRNSYS model, which allows the agent to perform actions on the developed model and receive the states consequently.




(a) (b)

Figure 1: (a) Required sensor layout to implement the proposed control framework (b) Experimental monitoring setup using wireless Droople sensors

Design of state, action and reward in the proposed framework are shown in Table 1. In this table, *Demand interval* is the demand in 5 liters intervals (e.g. interval 2 means the demand has been

between 5 to 10 liters), $Tamb$ is the outdoor air temperature, $Tstorage$ is the storage tank temperature at the middle, $Hours\ from\ overheat$ is the number of hours from last instance of heating above 60°C, and parameters a,b and c adjust the importance of each reward term, which are respectively set to 1, 2 and 0.8. To ensure water hygiene, the framework follows the rule that the hot water tank should be heated to 60 °C at least for 11 minutes once a day [4]. The reward function consists of three terms including the ***Energy reward*** to punish the agent for higher energy consumption, ***Comfort reward*** to punish the agent if a demand is met with a temperature lower than 40 °C, and a ***Hygiene reward*** to punish the agent if it exceeds more than 24 hours from the last overheat. As a result, agent always tries to make a balance between energy use, comfort and hygiene in the system. Similar to the proposed RL framework, rule-based control method was also developed in Python and coupled with TRNSYS model of hot water system.

Table 1: Design of state, action and reward

| | |
|---|---|
| ***State*** | $Demand\ interval_{t-1}$,..., $Demand\ interval_{t-6}$<br>$Tamb_{t-1}$,..., $Tamb_{t-6}$<br>$Tstorage_{t-1}$<br>$Hour\ of\ day$<br>$Day\ of\ week$ |
| ***Action*** | *ON or OFF* |
| ***Reward*** | $Energy\ reward = -a \times Energy$<br>$Comfort\ reward = -b \times max(40 - T_{storage}, 0)$ or 0 if no demand<br>$Hygiene\ reward = -c \times max(Hours\ from\ overheat - 24, 0)$<br>$Total\ reward = Energy\ reward + Comfort\ reward + Hygiene\ reward$ |

The proposed framework unfolds over three sequential stages as shown in Figure 2. The framework starts with an ***Off-site train phase*** to reduce learning period and probability of occupant discomfort. In this phase, a transient model of the system is integrated with the stochastic hot water use model to mimic the occupant behavior [9]. Therefore it provides a virtual environment for the agent to gain a prior experience without disturbing occupants comfort. The second stage is then the ***On-site train phase***, in which the pre-trained agent is implemented over the target building. To mimic the target system, a model with different parameters is used in this stage together with the experimentally recorded hot water use data to represent the behavior of real occupants. It is assumed that the agent is trained once in the lab, with no prior knowledge of the target buildings. Then the pre-trained agent can be implemented on different buildings with a heat pump water heating system. Therefore, the size of the system used in ***On-site train phase*** is considered to be different from the ***Off-site train phase*** to indicate that the pre-trained agent is very flexible and can quickly learn and adapt to the target building with different system parameters and occupant behavior. The third stage is the ***Deployment phase***, in which the agent is no longer learning and only controlling the system based on observed state at each time step, and therefore can be implemented on a low computational power system. The parameters used in each model are presented in Table 2.

Table 2: Parameters used in heat pump model in off-site train and on-site train phases

| **Parameter** | **Off-site train model** | **On-site train model** |
|---|---|---|
| Heat pump compressor power (kW) | 0.4 | 0.45 |
| Heat pump nominal heat rejection (kW) | 1.62 | 1.8 |
| Storage tank volume (L) | 270 | 300 |

### *2.2.* ***Base-line approach***

To evaluate the performance of proposed approach, its operation is compared with the most common control approach as the base-line. In this common practice, known as rule-based control, controller turns ON the heat pump when the tank temperature is less than a low-threshold, and turns it OFF when the tank temperature reaches a high-threshold. The main difference between the proposed approach

and the conventional model is that the former continuously learns and adapts to the variations of occupant behavior and climatic conditions to minimize energy usage, while the latter follows some rigid rules and is detached from these variations.

In this study, according to Booyen et al.[5], 65 °C and 75 °C are considered as the low and high thresholds. It should be noted that the occupants desire a certain flow rate with a certain temperature at the end use, which they obtain by mixing the hot and cold water streams. Consequently, if the supply temperature of hot water increases, the occupants mix a lower flow rate of hot stream with a higher flow rate of cold stream to obtain the same desired flow rate and temperature after mixing. To address this point and make a fair comparison between the two control approaches, it is assumed that both cases need to provide the same flow rate after mixing, with the same desired temperature of 40 °C. Considering the cold water temperature of 10 °C, the required flow rate of hot water to produce the desired flow rate after mixing with 40 °C temperature is calculated at each time step based on the supply temperature of hot water, and this flow rate is what should be provided by each system.

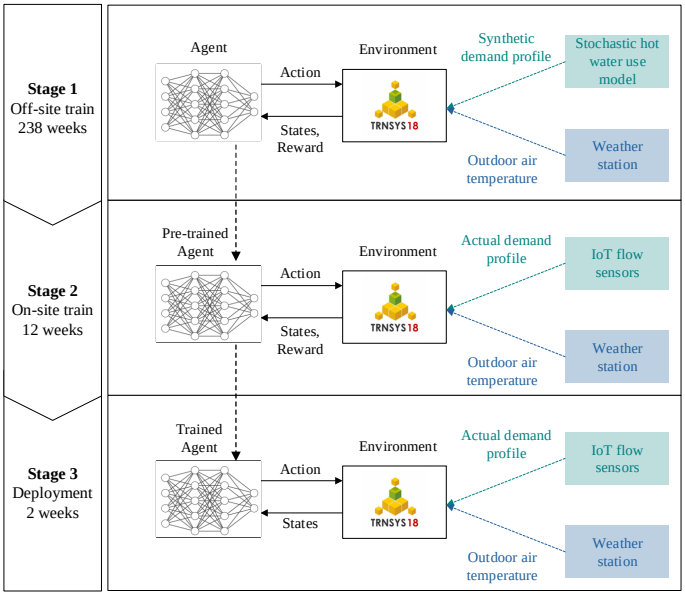



Figure 2: Stages of the proposed control framework

## 3. Results and Discussion

The ***Off-site train phase*** includes 5 years (238 weeks) of synthetic data to ensure that the agent has gained enough experience and will converge quickly when it is applied on the target house. Then, the ***On-site train phase*** and ***Deployment phase*** include 12 weeks and 2 weeks of the actual data of the target house, respectively. Interestingly, the ***On-site train phase*** and ***Deployment phase***, starting from 28th August 2020, are both during the COVID19-imposed home office period. Consequently, the occupants of the case study building (including two adults and two children) were mostly working from home, resulting in a very different hot water use behavior compared to the normal demand pattern that the agent has experienced during the ***Off-site train phase***. It will further highlight the adaptation potential of the proposed control framework.

### *3.1.* ***Evolution of reward***

The average of reward over each episode is a good indicator to evaluate the goodness of the learning process. If the reward does not converge it shows that the agent has failed in achieving the optimal control policy. To this purpose, the convergence of the reward function term-by-term is shown in

Figure 3. As shown in this Figure, the energy and comfort terms show a stable behavior from the beginning, meaning that the agent has already learned the optimal policy for energy and hygiene aspects over the ***Off-site train phase*** and quickly adapts to the target house. The comfort term shows more variations compared to the other terms (between the weeks 5 to 7), which is due to the variations of occupant behavior. The range of temperature violations, however, is less than one degree. The total reward, therefore, converges after 8 weeks of training over the target house. It shows that the agent has successfully gained enough experience over the ***Pre-train phase*** and can be adapted to the new cases quickly.

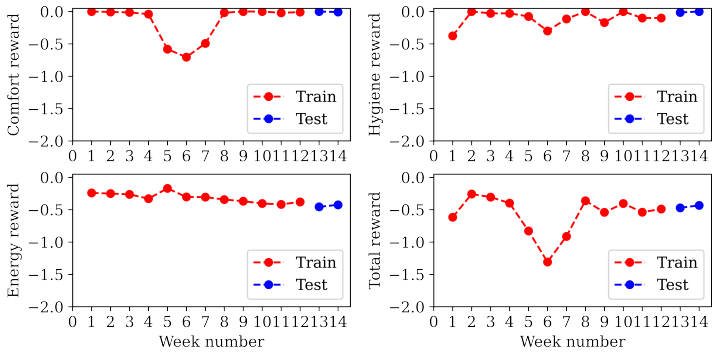



Figure 3: Evolution of total reward as well as each consisting term (energy, hygiene, comfort) over the train and test phases

### *3.2.* ***Performance of on-site training versus off-site training***

Table 3 shows the energy use, the percentage of demand which was met with a violated temperature (a temperature lower than comfort level of 40 °C), and the average temperature of the violations. It is worth to compare the performance of off-site and on-site training, as the former includes the hot water use behavior over the normal periods, while the latter includes the monitored hot water usage of actual building over the COVID-19 period. While the agent has already gained a previous knowledge when it starts the on-site training, as could be seen the percentage of demand met with violated temperature during the on-site training phase is higher than the off-site phase. This is because both of the on-site and off-site phases start with 2 weeks of random decision making (exploration). Considering that the on-site training phase is longer, the ration of violated demands over total demand would be lower. However, the average of violated temperatures in on-site phase (36.7 °C) is closer to the comfort level than the one of off-site phase (29.2 °C), indicating the higher experience of agent over the on-site phase. The average violations of 2.3 °C from the comfort level also shows that even during the on-site training phase, in which the agent is still learning, the occupant comfort is well preserved. It highlights the importance of off-site training with a stochastic model.

### *3.3.* ***Performance of RL versus conventional approach***

As shown in Table 3, the proposed framework provides a reduction in energy use by 39% over the ***On-site train phase*** and 24.5% over the ***Deployment phase*** compared to the conventional approach. Although during the ***On-site train phase*** the agent is still learning the occupants' behavior of actual building, it can provide a significant energy saving over the conventional approach. During the ***Deployment phase***, as the learning process is already completed, the agent violates the comfort of occupants over less than 1% of the demand with an average temperature of 39.3 °C, which is negligible. While the ***On-site train*** and ***Deployment phases*** have been during the COVID19-imposed home office period, the results show that the agent has successfully adapted to this abnormal behavior of occupants, which highlights the potential of an adaptive controller.

Table 3: Comparison of metrics between conventional approach and proposed framework

| | **Energy use (kWh)** | **Percentage of demand with violated temperature (%)** | **Average of violated temperatures (°C)** |
|---|---|---|---|
| Base-line (During train phase) | 1043.4 | 0 | - |
| Base-line (During deployment phase) | 196 | 0 | - |
| RL (During off-site train phase) | 19946.1 | 2.4 | 29.2 |
| RL (During on-site train phase) | 634.7 | 8.05 | 36.7 |
| RL (During deployment phase) | 148.1 | 0.92 | 39.3 |

### *3.4. **Visualization of RL performance***

Figure 4 shows the performance of the proposed framework during the ***Deployment phase***. Comparison of the control signal versus the demand shows that the agent has properly learned the occupant behavior, by turning OFF the heat pump when no demand is expected in upcoming hours and turning it ON before or during the demand instances. Therefore, as indicated by the variations of the storage tank temperature, the agent has learned how to preserve the occupant comfort as none of the demand instances are supplied with a temperature below 40 °C. Agent also preserves the water hygiene by overheating the tank one time per day, so the total hours from last overheat is always equal or below 24 hours. As the occupant comfort has a higher priority than energy for the agent, it does not strictly match the signal to the peaks of outdoor air temperature, and sometimes has to ensure the occupant comfort by charging the tank when the outdoor air temperature is lower to achieve higher COP (Coefficient of Performance) of the heat pump.


Figure 4: Performance of the proposed RL framework (A): Control signal versus hot water demand (B): Control signal versus outdoor air temperature (C): Variations of tank temperature (D): Cumulative hours from last sterilization

## 4. Conclusion

This study proposed a Reinforcement learning-based control framework which is able learn and adapt to the occupant behavior and make a balance between energy use, comfort and hygiene in heat pump water heating systems. The proposed framework is compared to the conventional control approach

which is detached from the occupant´s behavior. Although the monitoring campaign has been during home lockdown due to COVID-19, with an unusual water use behavior and less structured schedule of occupants, results show that the proposed framework has learned the occupant behavior in a relatively short period of 8 weeks, and reduced the energy use by 24.5% over the ***Deployment phase***, while preserving the occupant comfort and water hygiene.

## References

[1] Anna Marszal-Pomianowska, Chen Zhang, Michal Pomianowski, Per Heiselberg, Kirsten Gram-Hanssen, and Anders Rhiger Hansen. Simple methodology to estimate the mean hourly and the daily profiles of domestic hot water demand from hourly total heating readings. *Energy and Buildings*, 184:53–64, 2019.

[2] Amirreza Heidari, Nils Olsen, Paul Mermod, Alexandre Alahi, and Dolaana Khovalyg. Adaptive hot water production based on supervised learning. *Sustainable Cities and Society*, page 102625, 2020.

[3] Dane George, Nathaniel S Pearre, and Lukas G Swan. High resolution measured domestic hot water consumption of canadian homes. *Energy and buildings*, 109:304–315, 2015.

[4] MJ Booysen, JAA Engelbrecht, MJ Ritchie, Mark Apperley, and AH Cloete. How much energy can optimal control of domestic water heating save? *Energy for Sustainable Development*, 51:73–85, 2019.

[5] MJ Booysen, JAA Engelbrecht, MJ Ritchie, Mark Apperley, and AH Cloete. How much energy can optimal control of domestic water heating save? *Energy for Sustainable Development*, 51:73–85, 2019.

[6] Hussain Kazmi, Fahad Mehmood, Stefan Lodeweyckx, and Johan Driesen. Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems. *Energy*, 144:159–168, 2018.

[7] Federal Food Safety and Veterinary Office. Legionella control in buildings, 2020.

[8] Droople sa, switzerland, 2021. URL `https://droople.com`.

[9] MJ Ritchie, JAA Engelbrecht, and MJ Booysen. A probabilistic hot water usage model and simulator for use in residential energy management. *Energy and Buildings*, 235:110727, 2021.