Energy and AI 3 (2021) 100043

Contents lists available at ScienceDirect

Energy and AI

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

Deep reinforcement learning for home energy management system control

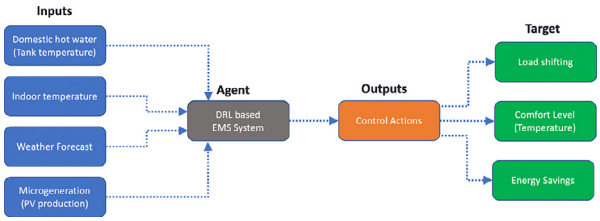
Paulo Lissa [a,b,c,*], Conor Deane [a,b,c], Michael Schukat [a], Federico Seri [a,b,c], Marcus Keane [a,b,c], Enda Barrett [a]

[a] College of Science and Engineering, National University of Ireland, Galway, Ireland
[b] Informatics Research Unit for Sustainable Engineering (IRUSE) Galway, Ireland
[c] Ryan Institute, National University of Ireland Galway, Ireland

HIGHLIGHTS

- Deep Reinforcement Learning-based control handles energy savings and comfort.
- PV self-consumption optimization brings flexibility for energy management systems.
- Deep Reinforcement Learning does not need prior information about the building.
- Home energy systems can have smart control due to new hardware and software.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 6 September 2020
Received in revised form 17 December 2020
Accepted 18 December 2020
Available online 25 December 2020

Keywords:
Deep reinforcement learning
Residential home energy management
Demand response
Autonomous control

ABSTRACT

The use of machine learning techniques has been proven to be a viable solution for smart home energy management. These techniques autonomously control heating and domestic hot water systems, which are the most relevant loads in a dwelling, helping consumers to reduce energy consumption and also improving their comfort. Moreover, the number of houses equipped with renewable energy resources is increasing, and this is a key element for energy usage optimization, where coordinating loads and production can bring additional savings and reduce peak loads. In this regard, we propose the development of a deep reinforcement learning (DRL) algorithm for indoor and domestic hot water temperature control, aiming to reduce energy consumption by optimizing the usage of PV energy production. Furthermore, a methodology for a new dynamic indoor temperature setpoint definition is presented, thus allowing greater flexibility and savings. The results show that the proposed DRL algorithm combined with the dynamic setpoint achieved on average 8% of energy savings compared to a rule-based algorithm, reaching up to 16% of savings over the summer period. Moreover, the users' comfort has not been compromised, as the algorithm is calibrated to not exceed more than 1% of the time out the specified temperature setpoints. Additional analysis shows that further savings could be achieved if the time out of comfort is increased, which could be agreed according to users' needs. Regarding demand side management, the DRL control shows efficiency by anticipating and delaying actions for a PV self-consumption optimization, performing over 10% of load shifting. Finally, the renewable energy consumption is 9.5% higher for the DRL-based model compared to the rule-based, which means less energy consumed from the grid.

1. Introduction

A report from Eurostat [1] shows that energy consumption in the residential sector accounted for approximately 26.1% of total energy consumption in the EU in 2018, where the main consumption was in heating systems and water heating, consuming 63.6% and 14.8% of the total, respectively. Moreover, most of the residential energy consumption is covered mainly by natural gas (32.1%) and electricity (24.7%), while renewables account for just 19.5%. There is a global trend of renewable energy asset expansion, as they represented almost two-thirds

* Corresponding author at: College of Science and Engineering, National University of Ireland, Galway, Ireland.
E-mail address: paulo.lissa@nuigalway.ie (P. Lissa).

https://doi.org/10.1016/j.egyai.2020.100043

of the new net world electricity capacity additions in 2016 and it is expected to increase by 43% between 2017 and 2022, according to studies from the International Energy Agency [2]. To reach the key target set by the EU of at least 32% share for renewable energy this growth has to continue [3]. However, as solar and wind generation rely on weather conditions, challenges due to intermittent generation have to be solved, and solutions for energy management such as demand response and photovoltaic (PV) self-consumption optimization can play a key role in this regard.

Residences with automated actuators and monitored sensors, known as smart homes, can get benefits from advanced machine learning (ML) techniques for energy management and achieve better results in terms of energy savings when compared with scheduling or manual control methods, which in some cases depend on the user's behavior or do not consider all the variables for optimal operation. This can be seen in Barrett and Linder's [4] work, where they have proven that machine learning-based adaptive methods, such as reinforcement learning (RL), can achieve even greater cost reductions in the heating, ventilating, and air conditioning (HVAC) domain, the most relevant load in European residences. Reinforcement learning is a sub-field of ML; it can learn the optimal policy by interacting directly with their environment, choosing actions based on its previous experiences within the domain, with no prior knowledge. After a number of trials and a reward signal indicating the benefit linked to the actions in a particular state, the agent can decide the best action to be taken for a given state. In the context of energy management, the rewards could be based on performing control actions when the cost of energy is low, or the production of energy in the house is high. The target would be, for example, saving energy as much as possible, but also respecting the user's needs.

Reinforcement Learning can be applied through different approaches. Vázquez-Canteli and Nagy [5] presented an extensive review of algorithms and modeling techniques in the demand response domain, where they classified articles according to their application and ability to address the problems of speed of convergence, and curse of dimensionality. Mason and Grijalva [6], in their review about reinforcement learning for autonomous building energy management, also explored different types of RL, concluding that applications of deep reinforcement learning (DRL) algorithms are expected to keep growing due to their increased effectiveness over traditional approaches.

In this work, we propose the use of reinforcement learning to manage and control the heating system and domestic hot water (DHW), with PV self-consumption optimization. The main contributions of this paper include:

- The development of a deep reinforcement learning algorithm for indoor and domestic hot water tank temperature control, aiming to reduce energy consumption by optimizing the usage of PV energy production. The algorithm also provides load shifting, helping the energy grid balance.
- A thorough investigation of the impact of user comfort levels and energy savings is conducted, with performance comparison against conventional control methods.
- A methodology is proposed for a flexible indoor temperature comfort threshold definition, considering a real case study in Ireland.

The rest of this paper is organized as follows: *Related Work* shows the related works regarding reinforcement learning applied to home energy management systems. *Environment Setup* provides information about the simulated environment and the temperature threshold definition. *DRL Home Energy Management System*, describes the algorithm and tests performed. The *Results* section presents all the relevant outputs of our experiments, showing the performance of the algorithm and comparing the methods. Finally, *Conclusions and Future works* recaps the main points of the paper, introducing ideas for future work.

2. Related work

A Home Energy Management System (HEMS) is a system deployed in the home containing both hardware and software that allows users to manage their energy consumption and production, thus enabling participation in the electricity market. These systems are becoming increasingly popular due to the recent advancement in residential metering, monitoring and controlling, which allows a more reliable energy management and new business schemes, such as demand response programs. Shareef et al. [7] presented the evolution of HEMS from works since 1970, showing control methods such as rule-based and artificial intelligence (AI), where the common targets were energy savings, CO_2 emission reductions, and user's comfort. They also pointed out a future trend of self-learning AI techniques for HEMS. Lee and Cheng [8] found that 34% of the papers about energy management systems in the same period belong to the residential sector. Vázquez-Canteli and Nagy [5] categorized papers about reinforcement learning for demand response across four different groups: HVAC and DHW, electrical vehicles, smart appliances, and distributed generation with storage; the first two groups represented more than two-thirds of the recent publications. Q-Learning is the most popular learning method [5,6,9], with deep Q-learning gaining popularity since 2015 due to its capability to handle learning rate issues from large state-action space, known as the curse of dimensionality. After analyzing a number of studies, authors in [6] showed that RL algorithms can improve residential energy efficiency, and although projects can vary significantly, typical savings of approximately 10% and 20% can be achieved for HVAC applications and DHW, respectively.

Regarding the HVAC application, Barrett and Linder [4] proposed a tabular Q-learning RL architecture for occupancy prediction and HVAC control, optimizing user comfort and energy costs achieving 55% and 10% of cost reductions compared to "always on" and "programmable" methods, respectively. Extending upon this research Lissa et al. [10] developed a transfer learning technique which successfully sped up the learning time taken to learn optimal policies in the domain. Cheng et al. [11] applied model-free Q-learning for HVAC control and window systems for natural ventilation, achieving up to 23% of energy savings. Wei et al. [12] applied a data driven DRL, with cost reductions from 20% to 70% when comparing their results with scheduling methods. Nagy et al. [13] used a data-driven approach, outperforming rule-based control by between 5% and 10%. Wang et al. [14] reduced energy consumption of a central HVAC by 5% through a model-free RL method applying recurrent neural networks. Gao et al. [15], Zhang and Lam [16] and Valladares et al. [17] applied DRL for energy optimization and thermal comfort control.

Moving to DRL applied to other home applications, Wan et al. [18] proposed an algorithm to minimize the energy cost of a house with battery energy storage. Liang et al. [19] also worked with energy storage systems, but in coordination with HVAC control. One of the most recent studies in this field belongs to Sangyoon and Dae-Hyun [20], where they proposed a hierarchical DRL for scheduling appliance usage in a single home, including an air conditioner, a washing machine, a PV system, an energy storage device, and an electric vehicle (EV), under a time-of-use pricing model.

The aforementioned model-free methods have proposed solutions for different home energy system architectures. However, to the best of our knowledge, this is the first time that DRL is proposed as a solution for indoor heating and DWH control, aligned with PV self-consumption optimization and a variable setpoint strategy. There are also other control methods for comfort and energy management that are not from the RL domain, such as fuzzy control [21,22] or model predictive control (MPC) [23,24], where building an accurate model for a house can be time-consuming.

Table 1
Construction characteristics.

Construction	*U*-Value (W/m^2)
Wall	0.27
Roof	1.50
Windows	3.00
Ground floor	0.45
Internal partitions	1.80
Doors	1.35

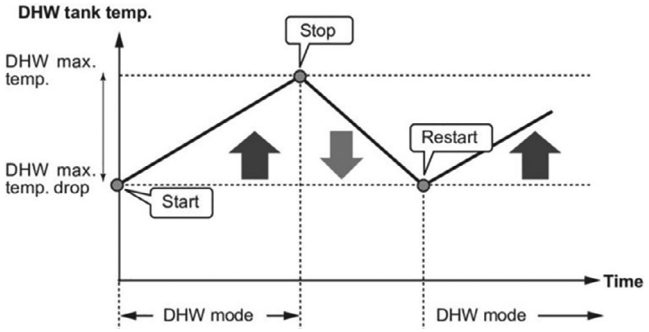


Fig. 1. DHW operation [25].

3. Environment setup

3.1. Building simulator

The first step of the experiment involved the development of the Building Simulator (BS) application which enabled the DRL control algorithms to be applied. The case study is a residential dwelling located on the Island of Inis Mór, Co. Galway, Ireland. The dwelling was built in the1970ı́s and has a total floor area of 110 m^2. In recent years the dwelling has been upgraded, including additional insulation to the walls and roof and installation of an 8.5 kW Mitsubishi heat pump along with a PV panel array consisting of 8 panels, with a total nominal power of 2 kWp. The construction characteristics of the building can be seen in Table 1.

The heating system installed in the dwelling is an 8.5 kW Mitsubishi Electric Ecodan heat pump. The heat pump connects to a 170 L hot water cylinder which is used to store hot water for both space heating and DHW. The DHW mode operates automatically based on the upper and lower limits set by the system installer as seen in Fig. 1. This is influenced by the DHW max temperature and the DHW maximum temperature drop. The hot water is automatically heated once the tank temperature exceeds the DHW max temp drop. For example, if the DHW max temperature is set to 50 °C and the lower threshold is 40 °C, there is a delta temperature of 10 degrees [25].

The BS was formed by developing a building energy model (BEM) extracting key data from the case study dwelling. Sensors were installed in the dwelling which measured indoor temperature °C, total electricity consumption (kWh), total PV production (kWh) and heat pump electricity consumption (kWh). The characteristics obtained from the site survey, along with data collected from each of the sensors, were used as to develop a detailed and calibrated white-box model, using the Integrated Environmental Solution Virtual Environment (IESVE) software [26]. Simulations have been carried out to identify the main parameters and heating transfer dynamics necessary to build a reduced grey-box model. The parameters extracted from the white-box model were indoor air temperature increase and decrease rates for both, DHW and indoor temperatures, considering their behavior during stationary conditions (system off) and when actions are performed.

Considering the heat pump off, the indoor temperature drop ranges from 0.0057 °C to 0.0230 °C every 5 min, depending on how high the difference between them is, if the indoor temperature is higher than the outdoor temperature. DHW temperature only relies on tank insulation, so the temperature drop is around 0.075 °C per 5-min interval for the operational range of 40–55 °C, slightly reducing this value until reaching the indoor temperature level. Regarding heat pump on, the temperature increases on average 0.2 °C and 2.75 °C every 5 min for space heating and DHW function, respectively. The new building simulator grey-box model reads the current environment state every 5 min and estimates the next DWH and indoor temperature values, calculated from the rates established previously. The benefit is that the new simulator can run the environment step-by-step, which allows us to implement our proposed DRL control approach.

3.2. Indoor thermal comfort threshold definition

Demand response programs can take advantage of occupant behavior changes, which can increase flexibility during the control decision process, hence allowing higher energy savings. However, authors in [27,28] stated that thresholds must be identified prior to deploying a strategy. In order to assess the thermal comfort of the occupant during the analysis, an indoor temperature threshold was developed based on a statistical analysis on the measured indoor temperature. The approach stems from Sweetnam et al. [27], who calculate the mean internal temperature (MIT) and upper and lower thresholds for 15-min intervals for each home in a 31-dwelling case-study, calculating the effect of a demand response control system on the heat pump. Upper and lower thresholds are based on calculating the upper and lower deciles of temperature. The outcome of the data analysis is a unified temperature profile bandwidth over a 4-month period which is used to compare temperatures pre- and post instalment of HEMS.

Different approaches can be used to determine the best threshold to handle savings and comfort. For instance, a popular method for assessing building comfort is the predicted mean vote (PMV), which was first proposed by Fanger [29] and is an index that predicts the mean value of the thermal sensation votes of a group of people according to a seven-point sensation scale [30]. Comfort compliance is achieved if the PMV is within a range of $-0.5 < x < 0.5$ as per ASHRAE Standard 55-2017 guidelines [30]. This method is more suitable for application in controlled office environments but less suitable for homes where residents may have other individual preferences for temperature conditions, and the assessment of temperature may depend on other criteria than in an office environment.

To identify the individual preferences for the case study, the homeowner was interviewed on the preferred indoor temperature and occupancy hours, where two key assumptions were identified. Firstly, it is assumed that over the previous heating season months the occupant is satisfied with the indoor temperature when the heating is on. Secondly, through validation with the homeowner, the temperature can be altered for the analysis. Data was gathered at 30-min intervals for each day over three months. In the case of significant outliers or anomalies, data was cleaned by interpolating the temperature based on previous and proceeding timeslot.

For each of the months, the cumulative indoor temperature was calculated for each hour of the day. For example, the indoor temperature value at 00:00 h summed for each day (92 days in total) and the mean temperature of that hour was calculated by Eq. (1):

$$MIT = \frac{\sum_{n=1}^{N} T_{indoor}}{N} \tag{1}$$

Where N is the total number of days for which the data is gathered. The MIT is calculated for every 30 minutes and is outlined in Fig. 2. To measure the upper and lower limits of MIT, the standard deviation (Eq. (2)) is calculated from the mean.

$$\sigma = \sqrt{\frac{\sum |x - MIT|^2}{N}} \tag{2}$$

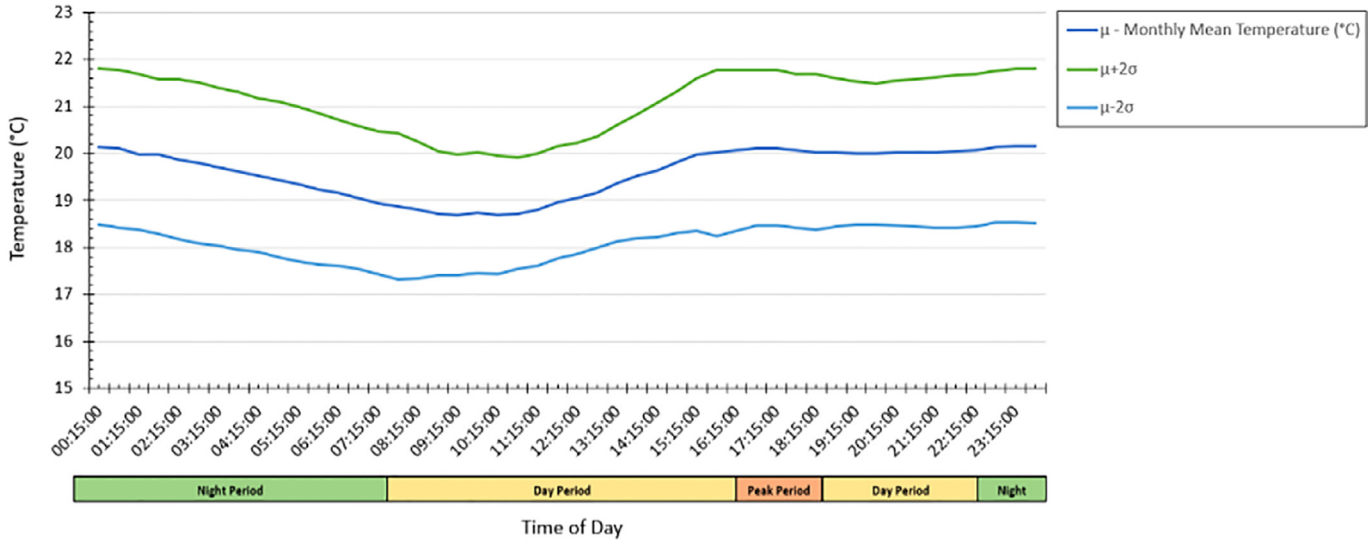


Fig. 2. Temperature threshold.

Thus, two times the standard deviation was formulated by multiplying the standard deviation by 2.

The approach outputted the MIT for each hour of the selected months and the associated lower and upper threshold limits for analysis. It is assumed that changes in indoor temperature which are within this temperature bandwidth are acceptable and can be utilised for the demand response strategy. The MIT and associated upper and lower limits are outlined in Fig. 2.

The proposed threshold is utilized as part of our DRL environment and control optimization in later stages, helping to achieve additional savings without impacting user's comfort.

4. DRL-Based home energy management system

4.1. Markov decision processes and reinforcement learning

The Markov Decision Process (MDP) is basically composed of 4 components: the state space (S), the set of possible actions (A), the probability distributions regarding state transitions ($p(|s,a)$) and the probability distribution governing the rewards received ($q(|s,a)$). Its main property states that the *"future is independent of the past given the present"*, which means that the current state incorporates all the information regarding past states, hence there is no need of keeping a historical record of actions performed or states.

An MDP task usually can be discretized into time periods, where at each period t the agent occupies a state $s_t \in S$, and then chooses an action a_t from the set of all possible actions within the current state. Performing the chosen action results in a state transition to $s_{(t+1)}$ and an immediate numerical reward $R(s_t, a_t)$. The state transition probability $p(s_{t+1}|s_t, a_t)$ governs the likelihood that the agent will transition to state s_{t+1} as a result of choosing a_t in s_t. The numerical reward received upon arrival at the next state is governed by $q(s_{t+1})|s_t, a_t)$ and is indicative as to the benefit of choosing a_t whilst in s_t. Solving an MDP leads to an output policy, which contains information about actions versus states, helping the agent's decisions over the entire learning period.

If the complete environment model is known, the problem can be solved through traditional dynamic programming techniques such as value iteration. However, most real-world problems are not fully observed, which means that the model needs to be either approximated (Model-Based Reinforcement Learning), for instance using statistical techniques, or having its value function or policy directly estimated (Model-Free Reinforcement Learning), where learners attempt to directly approximate a control policy through environmental interactions [4,10,11,14].

MDP models can be applied to HEMS, however dynamic programming techniques can only be used in complete models. If there is a lack of information about the environment, for instance, rewards or transitions probabilities missing, a model-free method such as Q-learning [31] can be used to generate optimal policies. Q-learning is part of the temporal difference methods, having the capability of being able to make predictions incrementally and in an online fashion. The update rule for Q-learning is defined in Eq. (3) and calculated each time a state is reached which is non-terminal.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)] \tag{3}$$

Actions are selected according to the policy π and approximations of $Q^{\pi}(s,a)$ are calculated after each time interval. Different action selection policies can be chosen, such as $\epsilon\text{-}greedy$ and $softmax$, aiming to balance exploration and exploitation and achieve the best results. Over time the actions selected should converge to the optimal, where the agents consistently choose actions which give the greatest amount of cumulative reward. For instance, the $\epsilon\text{-}greedy$ strategy chooses the best action from the policy most of the time, however the agent explores a certain amount of the time governed by ϵ. Let $A'(s) \subseteq A(s)$, be the set of all non-greedy actions. The probability of selection for each non-greedy action is reduced to $\frac{\epsilon}{|A'(s)|}$, resulting in a probability of $1-\epsilon$ for the greedy strategy. The discount factor γ, $(0 < \gamma < 1)$, determines the importance of future rewards, where a value close to 1 assigns a greater weight to it, while a value close to 0 considers only the most recent rewards. Finally, α is a value lower than 1 that represents the learning rate of value estimates over the learning process.

Depending on the algorithm chosen, the set of estimated Q-values, actions, and states can be represented in tabular form (Q-table) or as part of a function approximator. Although tabular methods present good accuracy, they require continuous updating of the value estimates through repeatedly revisiting the states and choosing actions in the environment. As the size of the state-action space grows, the learning time to converge to an optimal policy also increases. This is known as the curse of dimensionality, where each additional state or action variable added increases the problem size exponentially.

One of the possible solutions is to replace the Q-table, using an artificial neural network (NN) to estimate the Q-values of Eq. (3), which can now be simplified. The learning rate α is no longer needed, as this method has a back-propagating optimizer that already has this function. After removing α, the two $Q(s,a)$ terms cancel each other. The

Eq. (4) shows the new update rule after the changes. This technique, known as Deep Reinforcement Learning, can manage larger state-action spaces, thus bringing more possibilities when working with a huge number of variables.

$$Q(s,a) \leftarrow r + \gamma Q(s',a') \tag{4}$$

As the proposed HEMS will manage a variety of variables with a high range of values, and also thinking about future upgrades and additional features we can possibly add to it, we chose an online DRL approach as the core of our solution. The details about the neural network design are described in the following sections.

4.2. HEMS operation and MDP formulation

The proposed HEMS is composed of the elements described in Sub-Section 3.1. The goals are to optimize energy usage and to ensure user's comfort, by controlling the heating and DHW loads. The optimization process is made by performing actions when PV production is higher and considering the variable setpoint from Sub-Section 3.2. The timestep is discretized to 5-min periods, where the system analyzes the current state, performs an action, and moves to another state. The states are independent, as they only rely on weather conditions and building dynamics at the current time, thus following the Markov property. The state-space S comprises the following:

- ot: Outdoor temperature (˚C).
- it: Indoor temperature (˚C).
- tk: DHW tank temperature (˚C).
- pv: PV production (kW).
- hr: Hour of the day.

The DRL algorithm target is to minimize energy consumption, whilst keeping it and tk temperatures within their respective thresholds, as presented in Section 3. The set of actions A is related to the heat pump operation modes, which are space heating on ($HEAT_{ON}$), domestic hot water on (DHW_{ON}), and system on hold ($SYSTEM_{OFF}$). After performing an action, the environment will move to a new state and a reward r_t is calculated. The main target can be divided in three minor tasks, where individual rewards were assigned: indoor temperature control ($r1_t$: Eq. (5)), DHW temperature control ($r2_t$: Eq. (6)), and energy savings ($r3_t$: Eq. (7)).

$$r1_t = \begin{cases} 0, & \text{if } it_{min} < it < it_{max} \\ |it - it_{set}|, & \text{otherwise} \end{cases} \tag{5}$$

$$r2_t = \begin{cases} 0, & \text{if } tk_{min} < tk < tk_{max} \\ |tk - tk_{set}|, & \text{otherwise} \end{cases} \tag{6}$$

$$r3_t = \begin{cases} 0, & \text{if } pv > a_{power} \\ a_{power} - pv, & \text{otherwise} \end{cases} \tag{7}$$

Regarding $r1_t$ and $r2_t$, the reward is 0 if the current temperatures are in between their respective minimum and maximum threshold. The it limits can be seen in Fig. 2, and the tk limits are defined as DHW max. and the maximum temperature drop allowed (Fig. 1). To increase the system flexibility, the limits chosen are from 40° to 55°. This way, the algorithm can either create a buffer of hot water in case of high PV production, or delay a DHW_{ON} action if production is low at the current state. On the other hand, if $r1_t$ or $r2_t$ is out of the boundaries, the rewards are calculated as the distance of the current temperature compared to their central setpoint. Moving to $r3_t$, the reward is 0 when PV production is higher than the energy spent to perform a specific action, which is on average 1.5 kW for heating and 2.0 kW for DHW, considering our modeled environment. If the PV production is low, the reward is defined as the remaining amount necessary to meet the action power. Finally, the three partial rewards are aggregated following Eq. (8) along with the constants y, z, and w, added to balance the targets of comfort (indoor and DHW temperatures) and energy savings.

$$r_t = y \times (r1_t + r2_t \times w) + r3_t \times z \tag{8}$$

As mentioned previously, the state transitions follow the principles of Eq. (4), where the values are approximated by a neural network. The steps involved in the simulation and DRL process is depicted by Algorithm 1.

Algorithm 1 HEMS simulation process.

```
Initialize Building parameters
Initialize Q(s, a) arbitrarily
    Repeat (for each episode)
    Initialize s
    repeat
        Choose a from s using policy from Q(ϵ-greedy)
        Take action (a)
        Update building states (s′)
        Calculate reward (r)
        Q(s, a) ← r + γQ(s′, a′)
        s ← s′
    until s is terminal
end
```

4.3. Neural network design

The neural network designed to estimate the Q-values is similar to the one presented by Wei et al. [12]. It contains an input layer, two hidden layers, and an output layer, as can be seen in Fig. 3. First, each of the variables received from the current state is pre-processed and their values are normalized on a scale from 0 to 1 through the min-max method (Eq. (9)). This practice helps the algorithm to achieve a more stable learning process, as it does not have to deal with numbers with different ranges. For instance, in this experiment outside temperatures range from approximately 4 ˚C to 20 ˚C, while PV production ranges from 0 to 2 kW.

$$value' = \frac{value - min}{max - min} \tag{9}$$

Next, there are two hidden layers, where the Rectified Linear Unit (ReLU) is the activation function chosen, followed by the output layer with possible actions. The losses are calculated by the mean squared difference between the current output and the ideal target values. Moreover, we use a *Gradient Descent Optimizer*, which adjusts weights in the neural network to minimize these losses. The proposed algorithm also incorporates an ϵ-*greedy* policy to improve efficiency [4,10,12].

5. Results

5.1. Experiments and methodology

To validate the proposed DRL algorithm we simulated our test environment using weather data from Weatherbit.io [32], considering information from 1 May 2020 to 31 December 2020, and historical data about PV production from the mentioned Irish residence in the same period. The first step was to create a baseline, so the Mitsubishi's rule-based operational mode from Section 3.1 was replicated, where the controller has to keep indoor temperature between 19 ˚C and 22 ˚C, and DHW temperature between 40 ˚C and 55 ˚C. This baseline is used in later stages for evaluating the efficiency of our method. The second step consisted of adding the dynamic indoor setpoint from Section 3.2 to the rule-based method. Finally, the DRL-based algorithm was deployed, aiming to get benefits from the dynamic setpoint and PV self-consumption optimization.

Over the DRL deployment, a number of different parameters have been tested to determine the best configuration for handling comfort and energy savings. The final NN architecture is composed of 16 neurons in

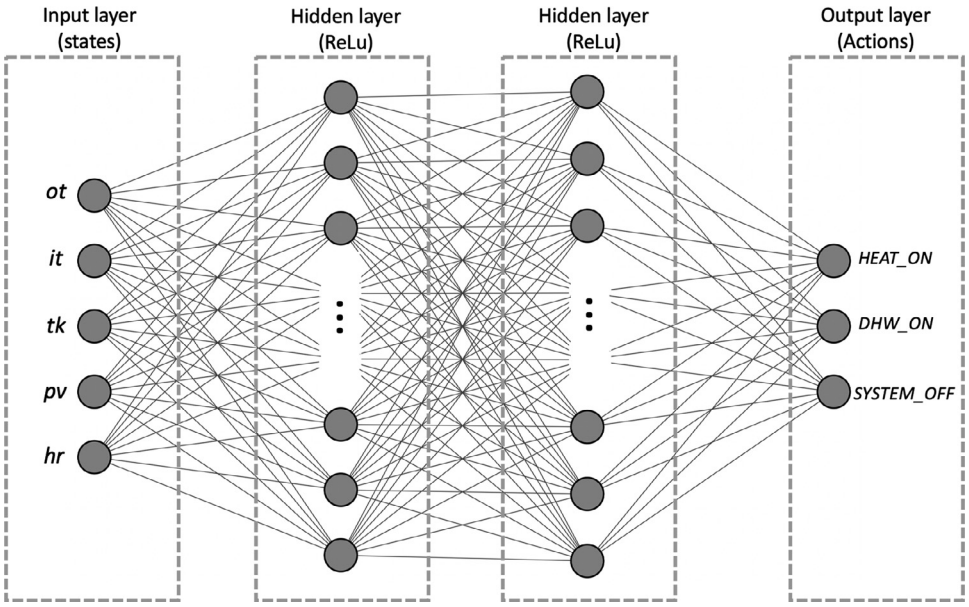


Fig. 3. Neural network design.

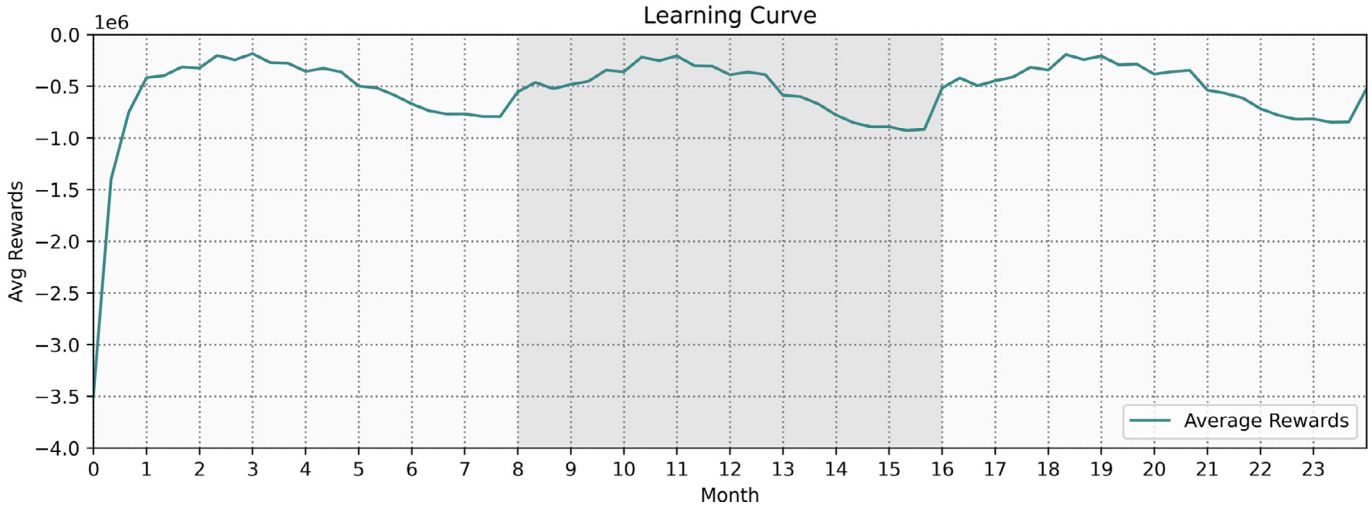


Fig. 4. Learning Curve.

each of the hidden layers, with $\alpha = 0.0001$, $\gamma = 0.95$ and $\epsilon = 0.05$. The y, z, and w reward balancing parameters were defined as 1000, 3.5, and 3.5 respectively. The learning process and algorithm stability happen over the first month of simulation, as can be seen in Fig. 4. At this stage, the performance achieved in terms of comfort is similar to the rule-based method, but it is also performing actions aiming to reduce energy imported from the grid.

After every 8 months of simulation (May to December), the episode is finished, and the environment is restarted. Fig. 4 shows the average accumulated rewards per month over 3 episodes. Note that the average rewards decrease from the middle to the end of each episode, and then start to increase in the subsequent months. This is related to the outside temperatures, because rewards depend on the number of actions executed. Colder temperatures found in winter months means that more heating actions are needed to keep user's comfort, hence lower rewards. On the other hand, warmer months need less heating actions and also have higher PV production, which makes rewards higher.

5.2. Comfort analysis

This section shows the overall comfort performance after the full implementation of the DRL algorithm with a dynamic setpoint. Results show that from June to December both indoor and DHW temperatures are kept in between their respective setpoints most of the time, with the indoor setpoint being violated less than 40 timesteps over the seven months, and less than 420 for DHW. The number is higher for DHW because the heating dynamic during the actions is different. Indoor heating is slow, which makes precision easier to be achieved. On the other hand, DHW heating is very fast, it needs less than 20 min to exceed the maximum setpoint limit, hence if the action of heating water is kept for one more timestep than necessary, there is a higher probability of surpassing the limit. But even in this condition, this temperature deviation represents less than 1% of the timesteps, and it does not exceed 2° above the maximum allowed. Fig. 5 shows the seven months of temperature behavior accumulated in an hourly granularity. Note that from the indoor temperature chart it is possible to see bigger bars from 11 am to 6 pm, which means the algorithm is exploring more of the setpoint limits due to a higher PV production activity. This is not possible to infer from the DHW chart, because its action occurs on average two times a day only, as the tank temperature decrease is slow.

5.3. Energy analysis

Having achieved comfort for both indoor and DHW, the second goal of our proposed DRL-based HEMS is to optimize PV self-consumption, by prioritizing performing actions when production is higher. The dy-

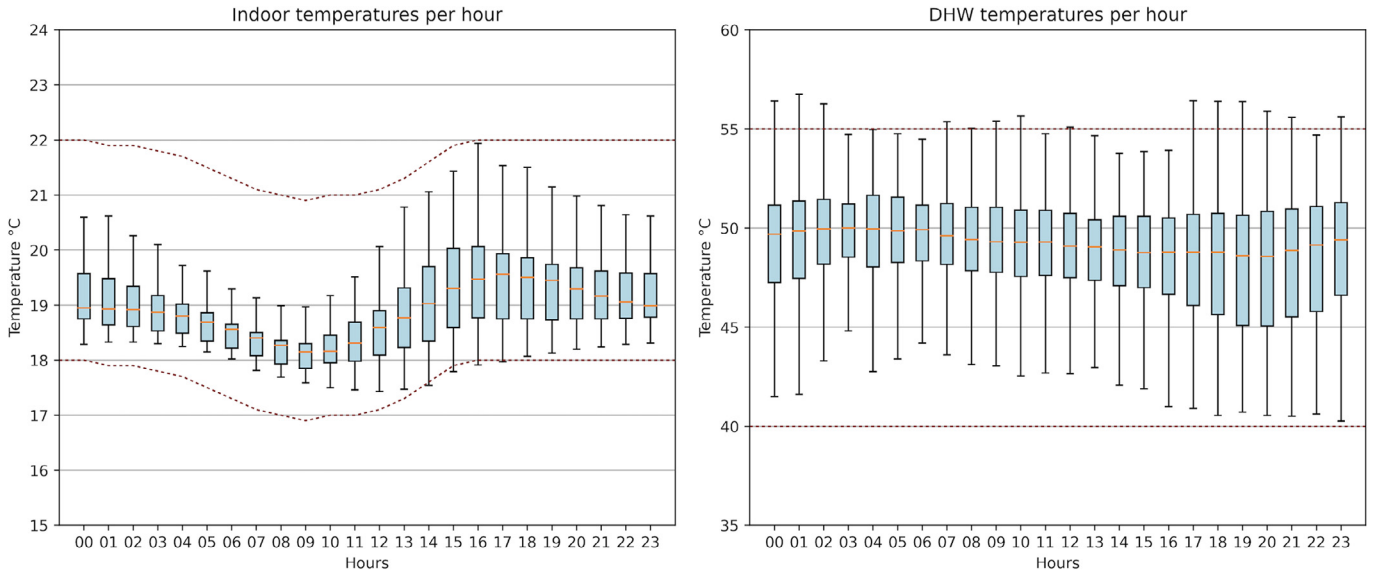


Fig. 5. DRL-based control: Temperature behavior.

Table 2
Algorithms comparison.

Month	Rule-based Static kWh (Baseline)	Rule-based Dynamic kWh (Reduction %)	DRL-based Dynamic kWh (Reduction %)
June	146.15	130.01 (11.5%)	126.55 (13.4%)
July	105.08	94.10 (10.4%)	87.85 (16.4%)
August	111.80	100.47 (10.1%)	93.28 (16.6%)
September	140.76	127.79 (9.2%)	126.07 (10.4%)
October	207.06	196.18 (10.1%)	190.28 (8.1%)
November	247.96	245.88 (0.8%)	241.06 (2.8%)
December	257.95	253.34 (1.7%)	253.99 (1.5%)
Total	1,216.78	1,147.79 (5.7%)	1,119.12 (8.0%)

namic setpoint brings additional flexibility during the decision making process. For instance, the algorithm can opt to anticipate the action and heat the space or tank creating a heating buffer, or it may postpone the action and save energy. There are two main benefits of this optimization: consumers can reduce their energy bills and also help utilities to balance the grid, through a demand-response program of load shifting.

To assess energy savings, we first compared the rule-based methods with and without a dynamic setpoint by estimating the net energy consumption (action power minus PV production), which was assumed to be zero if production is higher than consumption. Although PV production is part of the calculus, both methods do not consider any optimization in this regard, which means the actions are taken exclusively to keep comfort. As a result, the dynamic setpoint approach achieved up to 11% of savings over the months, with a 5.7% reduction average.

In the second analysis, we compared the DRL-based algorithm with a dynamic setpoint against the baseline. The average savings increased to 8%, reaching up to 16% of reductions over the summer months. There is a small energy reduction in November and December, mainly because PV production is low in those months. Also, outdoor temperatures are lower, hence more actions over the day are needed to keep indoor comfort, which reduces flexibility. Table 2 summarizes energy saving across the different algorithms.

To evaluate the DRL-based algorithm capability of adapting its control actions according to PV production, we compared the accumulated number of actions $HEAT_{ON}$ and DHW_{ON} performed over the seven months, grouped by hour, as can be seen in Fig. 6. The rule-based algorithm (dotted-red line) has actions well distributed across the day. After applying DRL with PV optimization (blue line), a number of actions that were originally in ranges from 2 AM to 10 AM and from 5 PM to 9 PM moved to the period from 10 AM to 5 PM. This load shifting represents 10.2% of more activity where PV production is higher.

PV self-consumption optimization is important to apply in situations where the policy is not favorable to the consumer. For instance, in this use case the PV system is connected to the utility grid and, if the energy is not consumed, the excess amount goes to the grid, but the user does not receive credits or incentives for that. In this regard, the proposed DRL-based algorithm was able to use up to 9.5% more of PV energy over the months than the rule-based one, hence consuming less from the grid.

In summary, the proposed algorithm was able to optimize PV self-consumption by choosing actions when production was higher, as shown in Fig. 7. The grey bars represent the accumulated number of actions per hour over the 8 months of the experiment. The green line shows the average PV production and, similarly to the previous analysis, the higher concentration of actions occurs when production is high. Finally, the orange dotted line is the average energy consumed from the grid. Even having more actions from 10 AM to 5 PM, the imported energy is lower compared to the other periods. There is also a slight increase of actions very early in the morning due to the DHW cycle, which happens on average twice a day. For instance, if DHW is turned on around 3 PM, it will probably turn on again 2 AM to keep the temperature to the specified setpoint.

5.4. Balancing comfort and energy savings

An important role of the proposed DRL-based control is to balance users' comfort and energy savings, and this process is ruled by the z value of Eq. (8). To define y, z, and w reward parameters, first we removed $r2t$, $r3t$, y, z and w. As a result, the algorithm only focuses on keeping the indoor temperature in between the setpoints. Then, we added $r2t$ and the parameter w. We started to change the value of w, verifying the impact on the indoor and tank temperatures, until both were in between their respective setpoint ranges. This was necessary because indoor and tank temperatures dynamics are different. Moreover, the setpoint operation zone is 4° for indoor and 15° for DHW, affecting directly the weight of the rewards. After this step, the algorithm achieved a similar performance to the rule-based method in terms of reaching temperature targets. Finally, $r3t$, y and z were added to the formula. The value of y only equalizes the temperature rewards to a similar weight level of

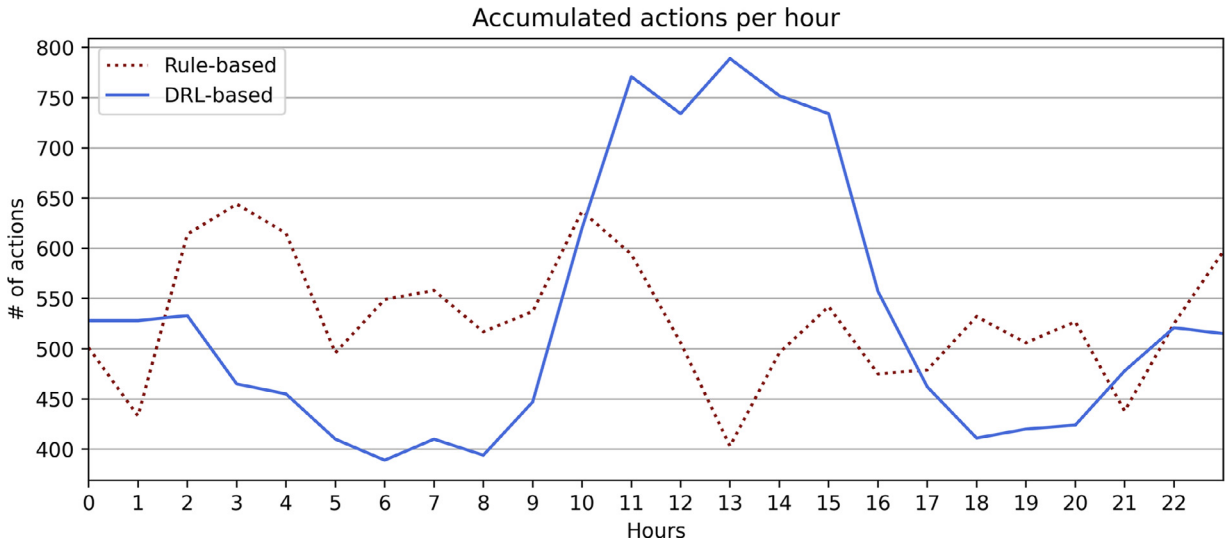


Fig. 6. Actions distribution.

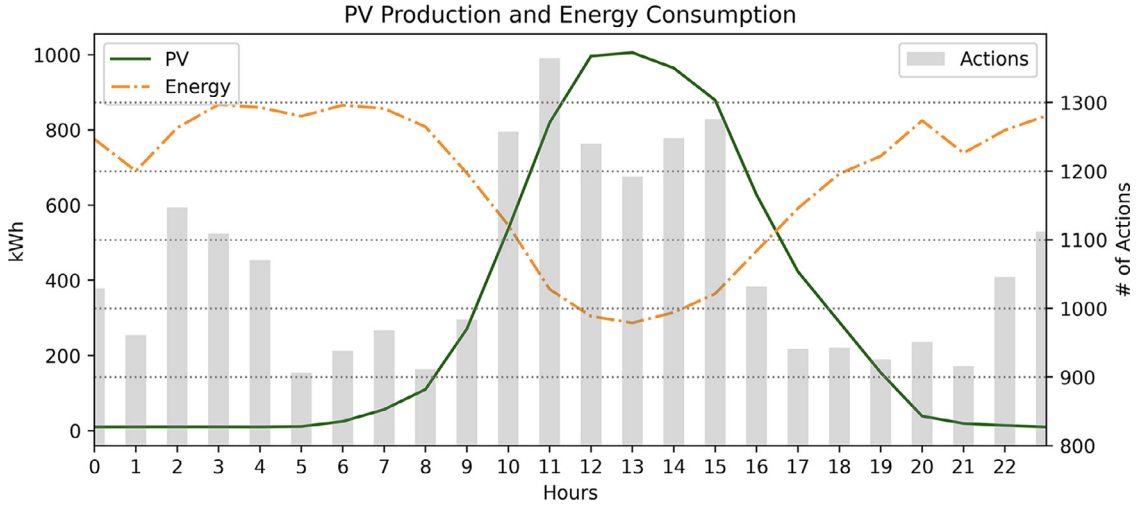


Fig. 7. PV production and Energy consumption.

energy savings. Finally, we started to change z to find the best balance between temperature and energy savings targets.

A z number closer to zero represents an algorithm focused strictly on comfort, and in that case it will not try to optimize energy usage according to PV production. As can be seen in Fig. 8, increasing z makes energy consumption to decrease, but after a certain limit the comfort matters start to be sacrificed. The purple line represents the percentage of timesteps out of indoor and DHW setpoints for different z values, considering accumulated values from June to December. The red line shows the total energy consumption for the same period and specified parameters.

As our first target was to achieve a similar comfort performance of the rule-based control, a z value of 3.5 was chosen. The time out of comfort in this configuration is around just 0.5%, in addition to outperforming inferior z values in terms of energy savings performance. Note that a z greater than 3.5 achieves more energy consumption reduction, but the time out comfort starts to increase quickly, reaching 25% of discomfort rate for a z equal to 6. In a common sense with the user, another z could be selected targeting greater energy savings, such as 4.5, where time out comfort is around 5% and the energy consumption is lower when compared to a z of 3.5. Another option would be also changing the w parameter of Eq. (8), which aims to balance the importance level between indoor and DHW temperatures.

6. Conclusion and future work

This work presented a new DRL approach for a HEMS control, which includes a PV self-consumption optimization and a dynamic setpoint definition for indoor temperature. The primary goal was to keep user's comfort, followed by energy savings and load shifting. Although our algorithm architecture and use case differs from previous works in many ways, the energy savings of up to 16% achieved is in line with values found by authors in [4,6,10,11,14]. Apart from bringing savings, PV self-consumption optimization also contributed to 10.2% of load shifting, by anticipating or delaying heating actions. Furthermore, right after the first month the level of comfort achieved by the proposed algorithm is nearly optimal, which means users would not experiment more than 1% of temperatures out of the specified setpoints range. The proposed DRL-based control also allows balancing energy savings and comfort according to the user's preferences.

For future work, a different approach to get the user's setpoint preferences could be applied. For instance, the user could receive a smartphone app where they can evaluate if the temperature is good or not at a specific time. Another option, if the house already has temperature sensors installed with historical data available, a data-driven statistical or machine learning models could be used to infer the setpoints. Moving to DRL, further analysis and optimization can be carried out consider-

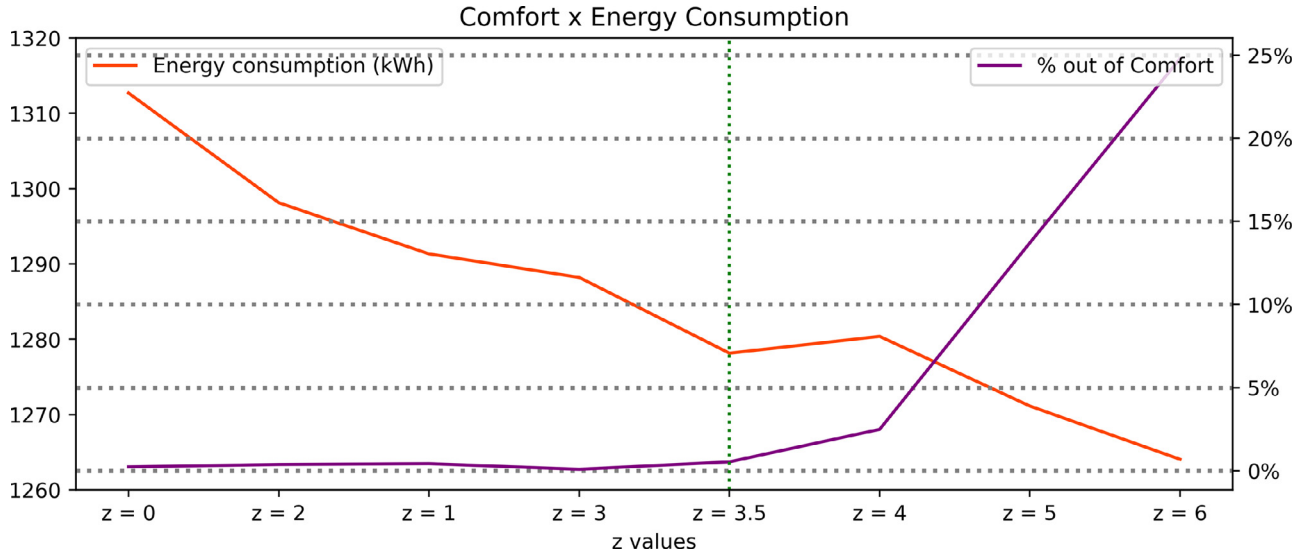


Fig. 8. Balancing comfort and energy savings.

ing other factors, such as dynamic prices or neighborhood energy peak reduction, which can also include a multi-agent or transfer learning approach. Another important topic to observe is about changes in the PV production policies, which can suddenly allow users to sell energy back to the grid. In that case, the algorithm could be upgraded considering selling prices, deciding what is the best time to consume or sell the produced energy. Finally, this work could be extended to other geographic locations with different PV production patterns and loads attached to it.

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.

Acknowledgment

This research work was funded by the European Union under the RESPOND project with Grant agreement No. 768619.

References

[1] EUROSTAT (2018). Energy consumption in households. Https://ec.europa.eu/eurostat/statistics-explained/index.php/Energy_consumption_in_households. Accessed 15 Jun 2020.
[2] (a). International Energy Agency (IEA). Renewables 2017: Analysis and Forecasts to 2022, 2017. https://doi.org/10.1787/re_mar-2017-en.
[3] Comission, E. (2020). 2030 climate & energy framework. Https://ec.europa.eu/clima/policies/strategies/2030_en. Accessed 15 Jun 2020.
[4] Barrett, E., & Linder, S. P. (2015). Autonomous HVAC control. A reinforcement learning approach,. 10.1007/978-3-319-23461-8-1.
[5] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Applied Energy 2019;235:1072–89. doi:10.1016/j.apenergy.2018.11.002. ISSN 0306-2619.
[6] Mason K, Grijalva S. A review of reinforcement learning for autonomous building energy management. Computers & Electrical Engineering 2019;78:300–12. doi:10.1016/j.compeleceng.2019.07.019. ISSN 0045-7906.
[7] Shareef H, Ahmed M, Mohamed A, Al Hassan E. Review on home energy management system considering demand responses, smart technologies, and intelligent controllers. IEEE Access 2018. doi:10.1109/ACCESS.2018.2831917. PP. 1-1.
[8] Lee D, Cheng C-C. Energy savings by energy management systems: areview. Renew Sustain Energy Rev 2016;56:760–77. doi:10.1016/j.rser.2015.11.067.
[9] Han M, May R, Zhang X, Wang X, Pan S, Yan D, Jin Y, Xu L. A review of reinforcement learning methodologies for controlling occupant comfort in buildings. Sustainable Cities andSociety 2019;51:101748. doi:10.1016/j.scs.2019.101748. ISSN 2210-6707.
[10] Lissa P, Schukat M, Barrett E. Transfer learning applied to reinforcement learning-based HVAC control. SN Comput Sci 2020;1:127. doi:10.1007/s42979-020-00146-7.
[11] Yujiao C, Norford Leslie K, Samuelson Holly W, Ali M. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. Energy Build 2018;169:195–205. ISSN 0378-7788.
[12] Wei T, Wang Y, Zhu Q. Deep reinforcement learning for building HVAC control. 2017 54th ACM/EDAC/IEEE design automation conference (DAC). austin; 2017. P. 1–6.
[13] Nagy, A., Kazmi, H., Cheaib, F., & Driesen, J. (2018). Deep reinforcement learning for optimal control of space heating. ArXiv e-prints (2018). arXiv :1805.03777 [stat.AP].
[14] Wang Y, Velswamy K, Huang B. A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. Processes 2017. doi:10.3390/pr503 0046.
[15] Gao, G., LI, J., & Wen, Y. (2019). Energy-efficient thermal comfort control in smart buildings via deep reinforcement learn- ing. ArXiv preprint arXiv :1901.04693.
[16] Zhang Z, Lam KP. Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. In: Proc. 5th ACM int. conf. syst. built environ.; 2018. Pp. 148–157.
[17] Valladares W, Galindo M, Gutiérrez J, Wu W-C, Liao K-K, Liao J-C, et al. Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. Build Environ 2019:155. doi:10.1016/j.buildenv.2019.03.038.
[18] Wan Z, Li H, He H. Residential energy management with deep reinforcement learning. Proc Int Jt Conf Neural Netw 2018:1–7.
[19] Yu L, Xie W, Xie D, Zou Y, Zhang D, Zhixin S, Zhang L, Zhang Y, Jiang T. Deep reinforcement learning for smart home energy management. IEEE Internet of Things Journal 2019. doi:10.1109/JIOT.2019.2957289. PP. 1-1.
[20] Lee S, Choi D-H. Energy management of smart home with home appliances, energy storage system and electric vehicle: a hierarchical deep reinforcement learning approach. Sensors 2020;20:2157. doi:10.3390/s20072157.
[21] Shepherd A, Batty W. Fuzzy control strategies to provide cost and energy efficient high quality indoor environments in buildings with high occupant densities. Build Serv Eng Res Technol 2003;24(1):35–45.
[22] Calvino F, La Gennusa M, Rizzo G, Scaccianoce G. The control of indoor thermal comfort conditions: introducing a fuzzy adaptive controller. Energy Build 2004;36(2):97–102.
[23] Wei T. Design and management for energy-efficient cyber-physical systems. Riverside: UC; 2018.
[24] Wei, T., Chen, X., Li, X., & Zhu, Q. (2018). Model-based and data-driven approaches for building automation and control. 1–8. 10.1145/3240765.3243485.
[25] Electric M. Mitsubishi electric pre-plumber cylinder with FTC4 control system. Travellers lane, Hatfield, Hertfordshire, AL10 8XB, England: Mitsubishi electric; 2018.
[26] IESVE (2019). IES virtual enviroment (IESVE). [Online] Available at: https://www.iesve.com/software.
[27] Sweetnam T, Fell M, Oikonomou E, Oreszczyn T. Domestic demand-side response with heat pumps: controls and tariffs. Build Res Inf 2019;47(4):344–436.
[28] Aghniaey S, Lawrence TM. The impact of increased cooling setpoint temperature during demand response events on occupant thermal comfort in commercial buildings: a review. Energy Build 2018;173(1):19–27.
[29] Fanger P. Thermal comfort: analysis and applications in environmental engineering. Denmark: Danish Technical Press; 1970.
[30] ASHRAE. Standard 55-2017: thermal enviromental conditions for human occupancy. Atlanta, GA: American Soceity for Heating, Refrigeration and Air Conditioning Engineers; 2017.
[31] Watkins C. Learning from delayed rewards, England: University of Cambridge; 1989. Ph.d. dissertation.
[32] (b). Weatherbit.io. https://www.weatherbit.io/. Accessed 30 Apr 2020.