POLITECNICO DI TORINO
Repository ISTITUZIONALE

Energy Management of a Residential Heating System Through Deep Reinforcement Learning

Original
Energy Management of a Residential Heating System Through Deep Reinforcement Learning / Brandi, S.; Coraci, D.; Borello, D.; Capozzoli, A.. - STAMPA. - 263:(2022), pp. 329-339. (Intervento presentato al convegno 13th KES International Conference on Sustainability and Energy in Buildings, SEB 2021 nel 2021) [10.1007/978-981-16-6269-0_28].

Availability:
This version is available at: 11583/2938752 since: 2021-11-18T19:34:10Z

Publisher:
Springer Science and Business Media Deutschland GmbH

Published
DOI:10.1007/978-981-16-6269-0_28

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Springer postprint/Author's Accepted Manuscript (book chapters)

This is a post-peer-review, pre-copyedit version of a book chapter published in Sustainability in Energy and Buildings 2021. The final authenticated version is available online at: http://dx.doi.org/10.1007/978-981-16-6269-0_28

(Article begins on next page)

29 May 2024

Energy Management of a Residential Heating System through Deep Reinforcement Learning

Silvio Brandi[1*], Davide Coraci[1], Davide Borello[1], Alfonso Capozzoli[1]

[1] , Department of Energy "Galileo Ferraris", TEBE Research group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy.
*silvio.brandi@polito.it

Abstract. In this study, a controller based on deep reinforcement learning was tested for a residential building equipped with a radiant heating system. In detail, a Soft Actor-Critic (SAC) algorithm was implemented to optimize the operation of the heating system while ensuring adequate levels of indoor temperature. A probabilistic window opening behavior model was implemented within the simulation framework in order to emulate the interaction of the occupants with the building.
A sensitivity analysis on SAC hyperparameters was carried out to determine the best configuration that was then deployed in four different scenarios in order to analyze the adaptability of the controller to different boundary conditions. The performance of the reinforcement learning agent was evaluated against a baseline strategy which combines rule-based and climatic control.
The developed agent was able to achieve a saving of heating energy provided to the building in the range between 2% and 6% while increasing temperature control performance up to 65% in the four scenarios investigated.

Keywords: building adaptive control, deep reinforcement learning, automated system optimization.

1 Introduction

The energy consumption related to the operation of building systems accounts for 40% of the worldwide energy demand and 36% of CO_2 emissions [1]. Heating, Ventilation and Air Conditioning (HVAC) systems represent the most energy-intensive in buildings and significant improvements have been made in recent years to enhance their energy efficiency [2]. However, the optimal management of these systems is challenging due to the influence of stochastic endogenous and exogenous factors which cause the non-linearity of the control problem [3]. Traditionally, ON/OFF or Proportional-Integrative-Derivative (PID) controllers are the most widely applied bottom-level control system. At the supervisory level Rule-Based Control agents (RBCs) are commonly employed. However, since these strategies are mainly reactive and unable to predict changes in weather or building conditions [4], or to take into account more than one control objective, their implementation results in sub-optimal control policies [5,6].

Model-based control strategies, such as Model Predictive Control (MPC), were explored to overcome such limitations, showing an excellent ability in improving comfort conditions and energy efficiency in buildings [7–9]. However, despite the promising results, MPCs are not widely adopted in real-world applications due to their strong dependence from the accuracy of the underlying model of the system [10] and from the robustness of the optimization algorithm [11]. As a consequence, Reinforcement Learning (RL), specifically Deep-RL (DRL), has emerged as a promising control algorithm due to its model-free approach for the optimization of building performance [12]. Recent works in literature have proven the feasibility in the application of DRL strategies to control supply water temperature [13–15] and indoor temperature setpoint [16,17].
In this paper, an off-policy DRL algorithm named Soft Actor-Critic [18], was implemented to control the supply heating power for a residential building located in Turin, Italy. The experiment was carried out in a simulation environment which combines Python and EnergyPlus. The control agent is designed to reduce heating energy supplied to the building while maintaining the desired indoor temperature values. Moreover, it was implemented a probabilistic model for the operation of the windows (open/close) to simulate the interaction of occupants with the residential building.

2 Methods

Reinforcement learning (RL) can be formalized as a Markov Decision Process (MDP), defined by a four-values tuple, including a set of state S, a set of action A, transition probabilities between the states and a reward function r. The goal of the RL agent is to learn an optimal control policy (π), a mapping between states and actions that maximizes the cumulative sum of future rewards [19]. The problem can be defined by two functions, namely state-value $v_\pi(s)$ and action-value $q_\pi(s)$. These functions determine the optimal policy of the RL agent and are used to show the expected return of a control policy π starting from a specific state or a state, action pair, as follows:

$$v_\pi(s) = E[r_{t+1} + \gamma v_\pi(s') | S_t = s, S_{t+1} = s'] \quad (1)$$

$$q_\pi(s,a) = E[r_{t+1} + \gamma q_\pi(s',a') | S_t = s, A_t = a] \quad (2)$$

where γ [0,1] is the discount factor for future rewards. For $\gamma = 1$ the agent will consider future rewards more important than current ones. Contrarily, for $\gamma = 0$ the agent will give greater importance to immediate rewards. The most widely applied approach among RL algorithms is the Q-Learning. Q-Learning exploits a tabular approach to map the relationships between states and action pairs [13]. These relationships are formalised as Q-values, which are updated according to the following formula:

$$Q(s,a) \leftarrow Q(s,a) + \mu[r_t + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (3)$$

where μ [0,1] is the learning rate, which determines the degree of overwriting of old knowledge with the new one. For $\mu = 1$ new knowledge completely substitutes old knowledge.

Soft Actor-Critic

Deep Neural Networks, and their combination with RL algorithms (i.e., Deep Reinforcement Learning (DRL)) seemed to overcome Q-Learning limitations. Therefore, DRL resulted more suitable for complex problems. In this paper, it was implemented the Soft-Actor Critic (SAC), an off-policy algorithm introduced by Haarnoja et al. [18] capable of handling continuous action spaces. The adopted actor-critic architecture employs two different deep neural networks: the *Actor* maps the current state based on the estimated optimal action, while the *Critic* evaluates the actions by calculating the value-function. The entropy regularization represents a key-feature in SAC, ensuring that the agent is pushed towards the exploration of new policies while avoiding that it gets stuck in sub-optimal behavior [3]. Therefore, in SAC algorithm the objective is to maximize both expected reward and entropy [20] as follows:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \sum_{t=0}^{\infty} [r_t + \alpha H_t^{\pi}] \quad (4)$$

where H is the Shannon entropy term, expressing the attitude of the agent in taking random actions, and α is the temperature parameter, a regularization coefficient that determines the importance of the entropy term against the reward.

3 Case Study and Methodology

The proposed control strategy was developed for a five stories residential building located in Turin, Italy. The building is characterized by a net heated surface of 527 m^2 organized into five different thermal zones, one for each floor. The average transmittance values of the opaque and transparent envelope components are 0.985 and 2.681 W/m^2K respectively. The thermal zones are served by a radiant floor heating system equipped with a variable speed circulation pump. The amount of heat delivered to each zone can be controlled by three-way circulation valves. The objective of the developed DRL controller is to maintain the indoor temperature into an acceptability range, defined between [-1, 1] °C from the desired temperature value of 21°C, while reducing the heating energy provided to the building through the regulation of the heating power supply to each radiant floor. The design value of supply heating power per each floor is respectively equal to 11 kW, 6.5 kW, 5.0 kW, 5.0 kW and 6.5 kW. The interaction between the control agent and the building energy model was simulated through Building Control Virtual Test Bed (BCVTB), that allows the information exchange between the building model (developed in EnergyPlus) and the SAC control agent (developed in Python). Simulation time step and control time step was equally set to 30 minutes.

3.1 Occupancy schedules and modeling of window opening behavior

Different occupancy schedules were implemented for each floor in the building. Moreover, in order to better characterize the behavior of the occupants and, in particular, their interactions with the building a probabilistic model for window opening/closing was implemented within the simulation environment. The model developed in [21]

employs indoor temperature, relative humidity and CO_2 along with outdoor air temperature, relative humidity, wind speed and solar radiation to estimate probabilities of opening and closing of the windows in the building. The model is based on logistic regression which coefficients depends on the day of the week and the time of the day and was implemented in Python. The open/close state of windows was passed as an additional binary "control-signal" (i.e., 0 = closed, 1 = open) to EnergyPlus at each control time step.

3.2 Baseline control logic

The performance of the DRL controller was evaluated against a baseline controller consisting of a combination of rule-based and climatic-based logics for the control of the supply power of the heating system. The RBC controller manages the operation of the radiant heating systems according to the indoor temperature values and occupancy schedules separately for each floor. The heating energy to each floor is supplied two hours before the arrival of the occupants or when the indoor temperature is lower than the lower threshold of the acceptability range during occupancy period. Contrarily, during unoccupied hours or when the indoor temperature is higher than the upper threshold of the acceptability range during occupancy period the heating energy is not supplied. The opening degree of the valves, which determines the fraction of the nominal heating power provided to each floor, is managed through a climatic-based curve implemented in the real building on which this case study is based on. Nominal heating power is provided when outdoor air temperature values fall below 6°C, while when outdoor temperature rises above 19 °C the system is switched off.

3.3 Design of Reinforcement Learning Control Agent

This section discusses the design of the action-space, state-space and reward function of the DRL controller.

Design of the action-space. Since SAC was selected as the control algorithm, the action-space was expressed as a continuous space of 5 different actions, related to the supply power per each thermal zone, expressed in kW:

$$At = [A_{ground\ floor}, A_{first\ floor}, A_{second\ floor}, A_{third\ floor}, A_{fourth\ floor}] \quad (5)$$

The supply power was limited between 0 and the design value for each floor.

Design of the state-space. The state-space is composed of 26 observed variables, reported in **Table 1** with their lower and upper bounds, and the time step at which they are evaluated. The variables chosen are feasible to be collected in a real-world implementation and provide the necessary information required by the agent to predict immediate future rewards.
Observations were scaled in the (0,1) range according to a min-max normalization in order to be fed to the neural network.

Design of the reward function. The reward function was formulated as a linear combination of two competing terms. The first term is related to the heating energy supplied to the building expressed in kWh that is directly proportional to the control action. The second term is defined as quadratically proportional to the difference between the measured indoor temperature (T_i) and the desired setpoint (T_{SP}).

These terms were combined through the introduction of two weight factors (δ and β) that determine respectively the relative importance of heating energy consumption and indoor temperature requirements.

Table 1. State-space variables

Variable	Timestep	Min Value	Max value	Unit
Hour of the day	t	1	24	H
Day of the week	t	1	7	-
Outdoor Air Temperature	t	-8	32	°C
Direct Solar Radiation	t	0	1100	W/m^2
Time to Occupancy start	t	0	10	H
Time to Occupancy end	t	0	15	H
Indoor $\Delta T_{ground\ floor}$	t, t-1, t-2, t-4	-5	10	°C
Indoor $\Delta T_{first\ floor}$	t, t-1, t-2, t-4	-5	10	°C
Indoor $\Delta T_{second\ floor}$	t, t-1, t-2, t-4	-5	10	°C
Indoor $\Delta T_{third\ floor}$	t, t-1, t-2, t-4	-5	10	°C
Indoor $\Delta T_{fourth\ floor}$	t, t-1, t-2, t-4	-5	10	°C

The resulting reward function depends by the presence of the occupants and it is expressed as follows:

$$R = \begin{cases} -\delta * \sum_{i=1}^{N} E_{HEATING,i} - \beta * (T_{SP} - T_i)^2, & if\ OCC = 1 \\ -\delta * \sum_{i=1}^{N} E_{HEATING,i}, & if\ OCC = 0 \end{cases} \quad (6)$$

where $E_{HEATING,i}$ is the energy provided to each floor and N is the number of floors.

3.4 Training and Deployment Phase

Training phase. The performance of the DRL agent is highly influenced by several hyperparameters. To assess their influence, a sensitivity analysis was performed to select the value of the following hyperparameters: discount factor γ, learning rate μ, weight factors of the reward terms β and δ, batch size, number of neurons per hidden layer and number of training episodes. The different tested configurations are shown in **Table 2**. A training episode lasts 61 days and includes two months, from 1 November to 31 December, for a total of 2928 control time-steps. The weather file used in this phase refers to the heating season 2018/2019 for Turin, Italy.

Deployment phase. The best configuration of hyperparameters, retrieved from the sensitivity analysis performed during the training phase, was deployed in four different

scenarios to evaluate the adaptability of the learned control policy. The deployment period last one episode including two months, from 1 January to 28 February, considering the same weather file as in the training phase. The proposed scenarios are the following:

- Scenario 1: this is the base case with no implemented changes in the controlled environment. This scenario aims to evaluate the adaptability of the control agent to different patterns of outdoor conditions.
- Scenario 2: in this scenario the indoor temperature setpoint was decreased from 21 °C to 20 °C to assess the performance of the agent in satisfying different temperature requirements from the one assumed in the training phase.
- Scenario 3: in this case was evaluated the performance of the agent considering thermal transmittance U_w and the solar factor g of windows reduced to 1.1 W/m^{2}K and 0.33 respectively.
- Scenario 4: in the last scenario it was assessed the adaptability of the agent considering the internal mass doubled to rise the thermal inertia and internal heat capacity of the building.

The best trained agent was deployed statically, then it was not updated during the deployment and was used as static function. This process requires less computational time at the cost of a lower capability to adapt to changes in the controlled system [14].

Table 2. Tested hyperparameters configurations for the DRL controller during the training phase

Configuration	γ	μ	β	δ	Batch size	Neurons	Episodes
1	0.9	0.001	1	0.1	256	256	10
2	0.95	0.001	1	0.1	256	256	10
3	0.99	0.001	1	0.1	256	256	10
4	0.9	0.001	1	0.5	256	256	10
5	0.9	0.001	1	0.1	512	256	10
6	0.9	0.001	1	0.1	128	256	10
7	0.9	0.0001	1	0.1	128	256	10
8	0.9	0.0001	1	0.1	256	256	25
9	0.9	0.001	1	0.1	256	256	25
10	0.9	0.0001	1	0.01	256	256	25
11	0.9	0.0001	5	0.1	256	256	25
12	0.9	0.0005	1	0.1	256	256	25
13	0.9	0.0001	10	0.1	256	256	25
14	0.9	0.0001	1	0.1	256	128	25
15	0.9	0.0001	1	0.1	256	512	25

4 Results and Discussion

In order to consider the influence of the hyperparameters on the DRL control logic performances, a sensitivity analysis was performed. Two metrics were used to compare the different hyperparameters configuration: the energy saving with respect to the baseline and the cumulative sum of temperature violations during the occupancy hours. These metrics were summed up at the end of the training episode. The temperature violations, evaluated in °C, were calculated as the absolute difference between the indoor temperature and the lower or upper limit of the temperature acceptability range [19, 21], when the internal temperature was lower or higher than these limits. **Fig. 1** shows the cumulative sum of temperature violations (on y-axis, defined on a logarithmic scale for the sake of legibility) for the last training episode as a function of the energy saving. The performance of the baseline is reported with black dashed lines that divides the plot into four quadrants. The left-bottom quadrant includes the configurations in which the DRL agent reduced both the supplied heating energy and the temperature violations.

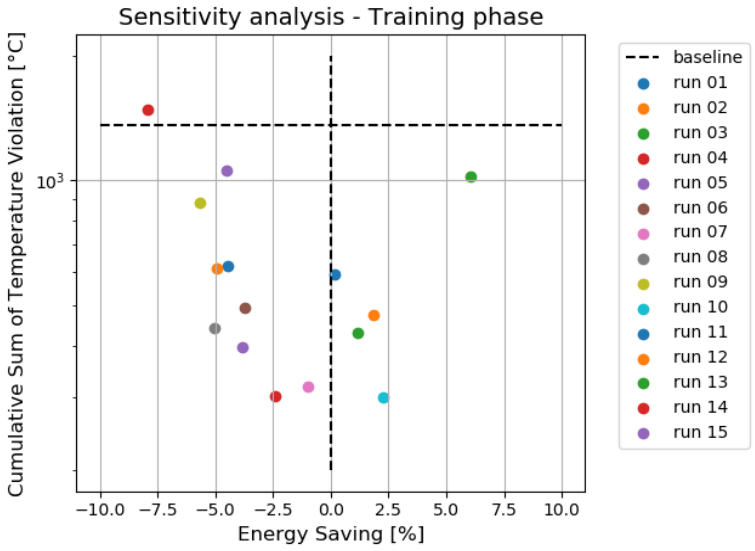

Fig. 1. SAC control agent performance in the last episode of the training phase

The configuration eight (i.e. run 08 in the figure) showed the best trade-off between energy saving (-5%) and temperature violations (-65%). The agent trained with this configuration was successively statically deployed in the four deployment scenarios. **Table 3** shows the results obtained for the DRL agent in the four deployment scenarios, considering the energy savings and the reduction of the cumulative sum of temperature violations with respect to the RBC control selected as baseline. In all scenarios, the

DRL controller leads to a reduction of heating energy supplied and temperature violations. The SAC control logic achieves the highest energy saving (i.e., about 5%) in the fourth scenario without reducing the temperature violations with respect to the baseline. In contrast, in the third scenario the DRL controller shows the highest reduction of temperature violations (60 %) with the lowest energy saving with respect to the baseline (around 2%). Overall, the SAC control agent ensures better performance than the baseline. In addition, the definition of a carefully designed state-space allows the developed agent to adapt to each scenario even if statically deployed, avoiding control instability issues and reducing the computational time.

Table 3. Performance comparison between DRL and RBC agents for all deployment scenarios

Scenario	Energy consumption [MWh]		Temperature violations [°C]2	
	DRL agent	Baseline	DRL agent	Baseline
1	20.6	21.2	592.2	1447.8
2	22.6	23.6	395.3	1158.6
3	20.0	20.3	603.2	1553.1
4	20.8	22.0	505.6	508.9

Fig. 2 reports the comparison between the SAC and RBC controllers in the first scenario during 5 days of the deployment period. The figure shows the indoor temperature and supply power patterns for the ground floors. The adaptive control agent is capable to reduce the temperature violations and energy supplied through an optimal management of the heating system. The SAC controller optimizes the pre-heating phase. In particular, the developed agent switches-ON the heating system later than the baseline, reducing the corresponding energy supplied and ensuring that indoor comfort requirements are met.

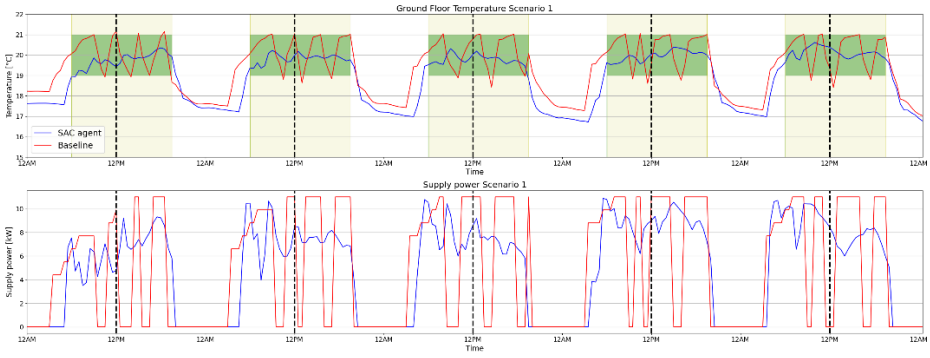

Fig. 2. Comparison between SAC and baseline controllers in Scenario 1

5 Conclusions

In this paper, a DRL agent was implemented to control the heating power supplied to each floor zone of a residential building. To represent the occupants' behavior as close as possible to the reality, it was adopted a model presented in literature based on logistic regression to simulate the windows opening and closing. The control agent was designed to enhance the energy efficiency while maintaining the indoor temperature within an acceptability range. A sensitivity analysis on the hyperparameters of the SAC algorithm was performed during the training phase to choose the DRL agent ensuring the best performance with respect to the baseline controller. The best trained agent was able to reduce the cumulative sum of temperature violations by more than 65%, while ensuring a reduction in the heating energy supplied. Furthermore, the agent was found effective in adapting to modification in the controlled system such as weather conditions, indoor temperature requirements and physical characteristics even if statically deployed. In particular, the developed controller reduced the heating energy supplied up to a maximum of 6% compared to the baseline controller, while ensuring better indoor temperature conditions.

Future works will be focused on the aspects of reproducibility and standardization of the developed controller, since it could perform differently in other buildings or HVAC systems. Moreover, the evaluation of indoor thermal comfort could be performed, by introducing parameters such as Predicted Percentage of Dissatisfied (PPD) and Predicted Mean Vote (PMV) in the reward function.

References

1. Yang, L.; Nagy, Z.; Goffin, P.; Schlueter, A. Reinforcement learning for optimal control of low exergy buildings. *Appl. Energy* **2015**, *156*, 577–586, doi:10.1016/j.apenergy.2015.07.050.
2. Martinopoulos, G.; Papakostas, K.T.; Papadopoulos, A.M. A comparative review of heating systems in EU countries, based on efficiency and fuel cost. *Renew. Sustain. Energy Rev.* **2018**, *90*, 687–699, doi:10.1016/j.rser.2018.03.060.
3. Coraci, D.; Brandi, S.; Piscitelli, M.S.; Capozzoli, A. Online Implementation of a Soft Actor-Critic Agent to Enhance Indoor Temperature Control and Energy Efficiency in Buildings. *Energies* **2021**, *14*, doi:10.3390/en14040997.
4. Wang, Z.; Hong, T. Reinforcement learning for building controls: The opportunities and challenges. *Appl. Energy* **2020**, *269*, 115036, doi:10.1016/j.apenergy.2020.115036.
5. Mechri, H.E.; Capozzoli, A.; Corrado, V. USE of the ANOVA approach for sensitive building energy design. *Appl. Energy* **2010**, *87*, 3073–3083, doi:10.1016/j.apenergy.2010.04.001.
6. Aghemo, C.; Virgone, J.; Fracastoro, G. V.; Pellegrino, A.; Blaso, L.; Savoyat, J.; Johannes, K. Management and monitoring of public buildings through ICT based systems: Control rules for energy saving with lighting and HVAC services. *Front. Archit. Res.* **2013**, *2*, 147–161, doi:10.1016/j.foar.2012.11.001.
7. Ma, Y.; Borrelli, F.; Hencey, B.; Packard, A.; Bortoff, S.; Sturzenegger, D.; Gyalistras,

D.; Morari, M.; Smith, R.S.; Oldewurtel, F.; et al. Model Predictive Control of thermal energy storage in building cooling systems. *IEEE Trans. Control Syst. Technol.* **2009**, *24*, 1–12, doi:10.1109/CDC.2009.5400677.

8. Oldewurtel, F.; Parisio, A.; Jones, C.N.; Gyalistras, D.; Gwerder, M.; Stauch, V.; Lehmann, B.; Morari, M. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy Build.* **2012**, *45*, 15–27, doi:10.1016/j.enbuild.2011.09.022.
9. Sturzenegger, D.; Gyalistras, D.; Morari, M.; Smith, R.S. Model Predictive Climate Control of a Swiss Office Building: Implementation, Results, and Cost–Benefit Analysis. *IEEE Trans. Control Syst. Technol.* **2016**, *24*, 1–12, doi:10.1109/TCST.2015.2415411.
10. Killian, M.; Kozek, M. Ten questions concerning model predictive control for energy efficient buildings. *Build. Environ.* **2016**, *105*, 403–412, doi:10.1016/j.buildenv.2016.05.034.
11. Wang, Y.; Boyd, S. Fast model predictive control using online optimization. *IEEE Trans. Control Syst. Technol.* **2010**, *18*, 267–278, doi:10.1109/TCST.2009.2017934.
12. Hong, T.; Wang, Z.; Luo, X.; Zhang, W. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy Build.* **2020**, *212*, 109831, doi:10.1016/j.enbuild.2020.109831.
13. Ahn, K.U.; Park, C.S. Application of deep Q-networks for model-free optimal control balancing between different HVAC systems. *Sci. Technol. Built Environ.* **2020**, *26*, 61–74, doi:10.1080/23744731.2019.1680234.
14. Brandi, S.; Piscitelli, M.S.; Martellacci, M.; Capozzoli, A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy Build.* **2020**, *224*, 110225, doi:10.1016/j.enbuild.2020.110225.
15. Zhang, Z.; Chong, A.; Pan, Y.; Zhang, C.; Lam, K.P. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy Build.* **2019**, *199*, 472–490, doi:10.1016/j.enbuild.2019.07.029.
16. Lu, S.; Wang, W.; Lin, C.; Hameen, E.C. Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE RP884. *Build. Environ.* **2019**, *156*, 137–146, doi:10.1016/j.buildenv.2019.03.010.
17. Park, J.Y.; Nagy, Z. HVACLearn: A Reinforcement Learning Based Occupant-Centric Control for Thermostat Set-Points. In Proceedings of the Proceedings of the Eleventh ACM International Conference on Future Energy Systems; Association for Computing Machinery: New York, NY, USA, 2020; pp. 434–437.
18. Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. Soft Actor-Critic Algorithms and Applications. *arXiv* **2018**.
19. Sutton, R.S.; Barto, A.G. Reinforcement Learning: An Introduction. *MIT Press Cambridge* **1998**, doi:10.1016/S0140-6736(51)92942-X.
20. Pinto, G.; Piscitelli, M.S.; Vázquez-Canteli, J.R.; Nagy, Z.; Capozzoli, A. Coordinated Energy Management for a cluster of buildings through Deep Reinforcement Learning. *Energy* **2021**, 120725, doi:https://doi.org/10.1016/j.energy.2021.120725.
21. Andersen, R.V.; Olesen, B.W.; Toftum, J. Modelling window opening behaviour in Danish dwellings. *12th Int. Conf. Indoor Air Qual. Clim. 2011* **2011**, *2*, 963–968.