

Lawrence Berkeley National Laboratory

LBL Publications

Title

Reinforcement Learning Control for Buildings Co-Optimizing Energy, Comfort, and Indoor Air Quality: An Annual Assessment

Permalink

<https://escholarship.org/uc/item/08f1371p>

Journal

2025 IEEE CONFERENCE ON CONTROL TECHNOLOGY AND APPLICATIONS, CCTA, 00

ISSN

2768-0762

ISBN

979-8-3315-3909-2

Authors

Guo, Fangzhou
Ham, Sang woo
Kim, Donghun

Publication Date

2025-08-27

DOI

10.1109/ccta53793.2025.11151460

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Reinforcement learning control for buildings co-optimizing energy, comfort, and indoor air quality: an annual assessment

Fangzhou Guo, Sang woo Ham, Donghun Kim

Building Technology & Urban Systems Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
{fzguo, SHam, DonghunKim}@lbl.gov

Abstract—Efficient control of Heating, Ventilation, and Air Conditioning (HVAC) systems is crucial for optimizing energy use and maintaining indoor comfort in buildings. Traditional control methods, such as PID control, cannot handle energy use trade-offs among multiple components in the building energy system at a supervisory level. Reinforcement learning (RL) presents a promising solution, offering adaptive and data-driven control strategies that optimize performance over time. However, RL also faces several challenges, including the conflicts encountered in co-optimizing energy savings, occupant comfort, and indoor air quality, and the requirement for extensive interactions with the environment in training. We proposed a flexible simulation platform that integrates a hybrid model for RL training and designed an RL agent to control the entire central HVAC system, focusing on co-optimizing energy consumption, thermal comfort, and indoor air quality (CO₂ and PM_{2.5} concentrations). Finally, we evaluated the RL agent's performance over an annual cycle. Our findings indicate that the RL agent can effectively manage the HVAC system with 14.7% energy savings annually and balance multiple objectives, which demonstrates significant potential for improving HVAC system control and sustainability in buildings.

I. INTRODUCTION

The control of Heating, Ventilation, and Air Conditioning (HVAC) systems is critical in maintaining indoor environmental quality and energy efficiency in buildings. Local control methods, such as proportional-integral-derivative (PID) controllers, cannot handle supervisory-level interactions between different devices and control objectives, leading to suboptimal performance and increased energy consumption. In comparison, advanced supervisory-level control methods, particularly those based on reinforcement learning (RL), offer a promising alternative by leveraging data-driven methods to optimize control policies. RL agents can learn and adapt to the complex interactions within HVAC systems, improving their ability to maintain desired indoor conditions while minimizing energy use. This is crucial in the growing energy demands and the need for sustainable building operations.

Many researchers have explored the application of RL to control HVAC systems, distributed energy resources, windows, and other building components to reduce energy consumption and costs while maximizing occupant comfort [1]–[4]. RL control has several advantages. Firstly, most RL algorithms are model-free, which means they learn optimal policies from interactions with the environment, not optimizing system operation by using models in real-time. Secondly, they are highly effective in tackling problems involving non-

linear dynamics in a building and HVAC system. Thirdly, RL can continue to enhance its performance over time and adapt autonomously to dynamic conditions after being deployed in real buildings, requiring minimal human intervention.

However, the application of RL comes with its own set of challenges. First, RL usually requires a complicated simulation environment for training before deploying in real buildings. The models can be classified by (1) white-box models developed by EnergyPlus [5], [6], TRNSYS [7], Modelica [8]–[10], ordinary differential equations (ODE) [11], [12], or others, (2) gray-box models such as resistance-capacitance (RC) building envelope models [13], and (3) black-box models [14]. White-box models are often used, but they are difficult to develop and also hard to calibrate. Black-box models need a lot of good-quality data to train and are hard to generalize to out-of-sample data. Gray-box models are ideal for some problems but cannot capture all the dynamics. Therefore, we propose a hybrid simulator that simulates the dynamics of each component with different types of models. Second, RL is designed to solve a complicated control problem with multiple objectives, such as co-optimizing energy saving, occupant comfort, and indoor air quality by controlling the chiller, AHU, and others. The decision space is usually very large, and thus training is computationally expensive and time-consuming. In this respect, a real need is to develop a sufficiently accurate, but also relatively simple and very flexible simulation platform for modeling any configurations of buildings and HVAC systems, which will significantly reduce the modeling workload and train the RL efficiently.

To tackle the above challenges, this paper made the following contributions:

- 1) Proposed a flexible simulation platform that integrates a hybrid model for RL training.
- 2) Designed an RL agent to control the entire central HVAC system for co-optimization of energy, thermal comfort, indoor air quality of CO₂ and PM_{2.5} concentration.
- 3) Conducted an annual performance evaluation of the RL and analyzed its control behavior.

II. METHODOLOGY

Training and evaluation of the RL agent are executed in the Python environment. A hybrid simulation model is developed and integrated in Python and then interacts with the RL

agent. The simulation model is developed based on a single-room building in FLEXLAB [15], a test facility at Lawrence Berkeley National Lab. This section introduces the hybrid simulation model and design of the RL controller.

A. Control problem description

This study aims to employ RL to control the central HVAC system of an office room, with the purpose of achieving optimal energy efficiency, good occupant comfort, and healthy indoor air quality. The office room has a size of 56 m², with two exterior walls and a large window facing the south. The HVAC system is composed of a chiller, a gas boiler, and an AHU serving the room. The RL agent is designed to control (1) the temperature setpoint of the chilled water and hot water when cooling/heating is requested by the thermostat, (2) the supply fan and outdoor air damper in the AHU, and (3) an air purifier in the office room.

B. Hybrid model development

A hybrid simulation model is developed to train and evaluate the RL agent, which is comprised of a dynamic HVAC system model, a dynamic building envelope model, and dynamics of indoor moisture (i.e., change of humidity ratio), CO₂, and PM2.5 transfer. After developing and calibrating these models, they are coupled in a Python environment, called a hybrid simulator. The details of each model will follow.

The HVAC system model is developed by the Modelica simulation tool with the Buildings library [16] due to its accuracy, flexibility, dynamic simulation in the continuous time domain, and seamless integration with RL agents in Python via FMU. Fig. 1 shows the network scheme of the dynamic model. In order for the Modelica model to interact with the RL agent and other models in the Python environment, it is first exported as a Functional Mock-up Unit (FMU) [17] and then imported into Python using the PyFMI package [18]. The input and output arrows in Fig. 1 allow the Modelica model to exchange information with other subsystems in the co-simulation.

The building envelope is described by a Resistance-Capacitance (RC) model, a gray-box model that represents the heat flow and heat storage in a building. Compared to other detailed physics-based models, RC models require fewer parameters and less computational power, but still provide sufficiently accurate results. Especially, only a couple of templates (e.g., 2R2C, 3R3C, 4R4C) can cover most configurations of the building envelopes, which is convenient to use for the purpose of building a large simulation platform. This paper adopts a 2R2C model representing the dynamics of the zone (indoor) air (Eq. 1) and thermal mass (walls) (Eq. 2):

$$C_{za} \frac{dT_{za}}{dt} = \frac{T_w - T_{za}}{R_{zw}} + \frac{T_{oa} - T_{za}}{R_{zo}} + \dot{q}_{sol} \beta A_{win} + \dot{Q}_{hvac} + \dot{Q}_{gain} \quad (1)$$

$$C_w \frac{dT_w}{dt} = \frac{T_{za} - T_w}{R_{zw}} + \dot{q}_{sol} (1 - \beta) A_{win} \quad (2)$$

where C and R refer to the capacitance and resistance of the zone air and walls that calibrated based on experimental data, $\dot{q}_{sol}$ is the incident solar radiation on the window surface [W/m²], A_{win} refers to an overall window area of the whole building [m²], and β refers to the fraction of radiation absorbed by the indoor air. $\dot{Q}_{hvac}$ is the sensible heating rate provided by the AHU [W], and $\dot{Q}_{gain}$ is the internal heat gain [W].

The dynamics of indoor moisture, CO₂, and PM2.5 transfer are all modeled by ordinary differential equations. Our calibration shows that they are accurate enough for training and evaluation.

$$\rho_a V \frac{d\omega_{za}}{dt} = \rho_a \dot{V}_{sa} (\omega_{sa} - \omega_{za}) + \dot{S}_\omega \quad (3)$$

$$V \frac{dC_{CO_2,za}}{dt} = \dot{V}_{oa} (C_{CO_2,oa} - C_{CO_2,za}) + 10^6 \cdot \dot{S}_{CO_2} \quad (4)$$

$$V \frac{dC_{PM,za}}{dt} = -(\dot{V}_{oa} + \eta_2 \dot{V}_{ra} + k_d V + CADR) \cdot C_{PM,za} + (\dot{V}_{oa} (1 - \eta_1) (1 - \eta_2)) \cdot C_{PM,oa} \quad (5)$$

where V is the volume of the office room [m³], ρ_a is the air density [kg/m³], $\dot{V}$ refers to the volume flow rate [m³/s], and subscripts za , sa , ra , and oa refer to the zone (indoor) air, supply air, return air, and outdoor air, respectively. In Eq. 3, ω is the humidity ratio [kg H₂O/kg dry air] and $\dot{S}_\omega$ is the moisture generation rate by occupants which approximately equals 2.43×10^{-5} kg/s/person [19]. In Eq. 4, C_{CO_2} is the CO₂ concentration [ppm] and $\dot{S}_{CO_2}$ is the CO₂ generation rate by occupants which approximately equals 5.2×10^{-6} m³/s/person [20]. In Eq. 5, C_{PM} is the PM2.5 concentration [μ g/m³], η_1 is the efficiency of the pre-filter at the outdoor damper, η_2 is the efficiency of the final filter in the supply air duct, k_d is the particle deposition rate [1/s], and $CA DR$ is the clean air delivery rate of the air purifier in the office room [m³/s].

C. RL controller design

In the supervisory control problem of a central HVAC system, multiple control variables exist in the water system, air system, and office room. Additionally, most control variables are continuous. Hence, this study adopts an actor-critic method termed Deep Deterministic Policy Gradient (DDPG) [21] as the RL agent to control the HVAC system. The DDPG algorithm is well-suited for applications involving continuous control actions. It operates within an actor-critic framework represented by artificial neural networks. The actor network determines the control actions based on the system's current state, and the critic network evaluates the actor's performance by estimating a Q-function. Additionally, to encourage exploration and avoid local optima, DDPG also introduces noise (often using an Ornstein-Uhlenbeck process) into the actions during the training process. The pseudo-code of DDPG is shown in Algorithm 1.

The RL controller is designed based on the structure of the DDPG algorithm. It has 12 state variables: the month, time of day, outdoor temperature, outdoor humidity ratio, incident

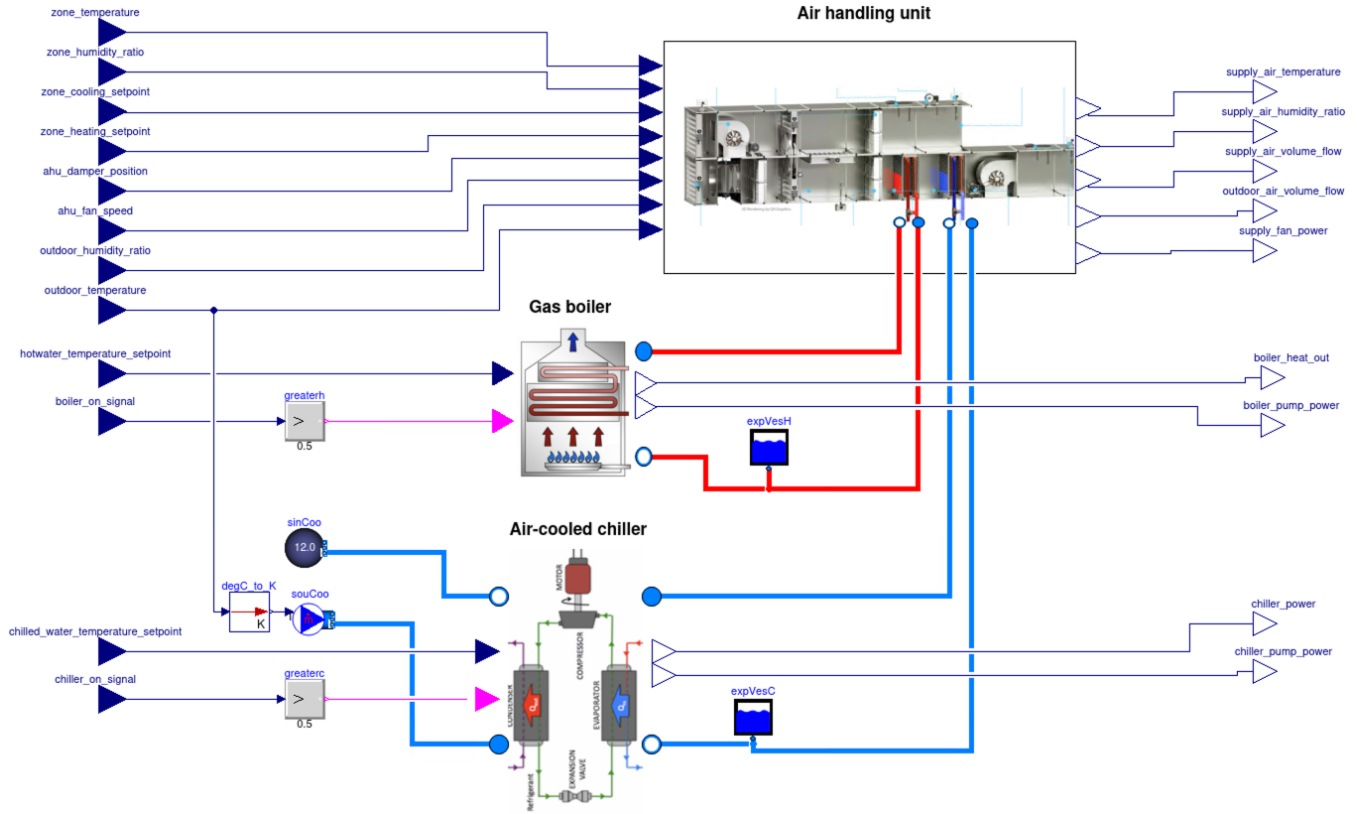


Fig. 1. Network scheme of the central HVAC system represented in Modelica. The HVAC system has an air-cooled chiller, a gas boiler, and an AHU. Several input and output icons are created for co-simulation with the RL agent in the Python environment.

Algorithm 1 DDPG Algorithm

- 1: Initialize critic network $Q(s, a|\theta^Q)$ and actor network $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ
- 2: Initialize target networks Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
- 3: Initialize replay buffer R
- 4: **for** episode = 1 to M **do**
- 5: Initialize an Ornstein-Uhlenbeck process N for action exploration
- 6: Receive initial state s_1
- 7: **for** $t = 1$ to T **do**
- 8: Select action $a_t = \mu(s_t|\theta^\mu) + N_t$ according to the current policy and exploration noise
- 9: Execute action a_t and observe reward r_t and new state s_{t+1}
- 10: Store transition (s_t, a_t, r_t, s_{t+1}) in R
- 11: Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R
- 12: Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
- 13: Update critic by minimizing the loss: $L = \frac{1}{N} \sum (y_i - Q(s_i, a_i|\theta^Q))^2$
- 14: Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$

- 15: Update the target networks:

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'} \end{aligned}$$

- 16: **end for**
- 17: **end for**

solar radiation on the window, outdoor PM2.5 concentration, zone cooling setpoint, zone heating setpoint, zone tempera-

ture, zone humidity ratio, zone CO₂ concentration, and zone PM2.5 concentration. Because the outdoor CO₂ concentra-

tion is usually around 400-500 ppm, relatively stable and much lower than the safety threshold for occupants, it is not included as a state variable. In this study, the outdoor CO₂ concentration is assumed to follow the normal distribution with a mean of 450 ppm and a standard deviation of 10 ppm.

The RL controller has 5 action variables: the supply chilled water temperature setpoint, supply hot water temperature setpoint, supply fan speed, outdoor air damper opening, and ON/OFF of the air purifier. Note that the zone air temperature setpoint is not controlled by RL. The cooling setpoint is 24°C during the occupied time and 26°C during the unoccupied time, and the heating setpoint is 21°C during the occupied time and 18°C during the unoccupied time.

The reward (r_t) of the RL agent has 5 components: the penalty for energy expenditure (r_E), along with violations of the indoor temperature (r_T), relative humidity (r_ϕ), CO₂ concentration (r_{co2}), and PM2.5 concentration (r_{pm}). Except for r_E , all other penalties are only added during the occupied time (6AM-10PM). They are expressed as follows:

$$r_t = \begin{cases} r_E + \alpha_1 r_T + \alpha_2 r_\phi + \alpha_3 r_{co2} + \alpha_4 r_{pm} & (occupied) \\ r_E & (unoccupied) \end{cases} \quad (6)$$

where

$$r_E = -(E_{chiller} + E_{chiller_pump} + E_{boiler} + E_{boiler_pump} + E_{fan} + E_{air_purifier}) \quad (7)$$

$$r_T = -([T_{za} - T_u]_+ + [T_l - T_{za}]_+) \quad (8)$$

$$r_\phi = -[\phi_{za} - \phi_u]_+ \quad (9)$$

$$r_{co2} = \begin{cases} -1 & \text{if } C_{CO2,za} \geq 1000 \text{ ppm} \\ 0 & \text{if } C_{CO2,za} < 1000 \text{ ppm} \end{cases} \quad (10)$$

$$r_{pm} = \begin{cases} -1 & \text{if } C_{PM,za} \geq 10 \mu\text{g}/\text{m}^3 \\ 0 & \text{if } C_{PM,za} < 10 \mu\text{g}/\text{m}^3 \end{cases} \quad (11)$$

In Eq. 7, the energy expenditures include the natural gas consumed by the boiler and electricity consumed by the chiller, chilled water pump, boiler pump, supply fan, and air purifier. In Eq. 8, $T_u = T_{csp} + 0.5^\circ\text{C}$ and $T_l = T_{hsp} - 0.5^\circ\text{C}$, which means a penalty of temperature violation is given when the room temperature is either half a degree Celsius above the cooling setpoint or half a degree Celsius below the heating setpoint. In Eq. 9, $\phi_u = 0.6$, which means a penalty is given when the room relative humidity is above 60%. The α 's in Eq. 6 determines the importance of each penalty term. Based on our previous study through grid search [22], a relatively large penalty is given to the α 's to prevent the controller from violating the thermal comfort and indoor air quality.

III. SIMULATION

A. Simulation setup

The training and test sets for the RL agent are typical meteorological year (TMY) datasets from two locations in Seoul, South Korea. It contains hourly data on outdoor temperature, humidity, solar irradiation, and outdoor PM2.5 concentration from January 1st to December 31st.

The internal heat gain of the office room includes equipment, lighting, and occupant sensible heat generation. In the simulation, the number of occupants in the office is assumed to be 4, and the equipment and lighting full load are assumed to be 865W and 465W, respectively. We also assume that the office building has a routine schedule every day, summarized in Table I.

TABLE I
OCCUPANT, EQUIPMENT, AND LIGHTING SCHEDULES AS A FRACTION OF FULL OCCUPANCY/LOADS.

	6AM-8AM	8AM-5PM	5PM-10PM	10PM-6AM
Occupancy	10%	100%	10%	0
Equipment	40%	90%	40%	40%
Lighting	30%	90%	30%	10%

The RL agent is trained for 100 episodes by interacting with the hybrid simulation model until it converges. After that, its performance is evaluated by the test set. During training and evaluation, some noises are added to the occupancy level and internal heat gain to simulate the actual environment in an office room and also prevent the RL agents from overfitting. The next section shows the evaluation results.

B. Baseline controller design

The performance of the RL controller is compared to a rule-based controller (RBC) as a baseline. Note that RBC is also a supervisory controller, and the variables it controls are the same as those of the RL agent. Table II summarizes its control strategy.

TABLE II
CONTROL STRATEGY OF THE RULE-BASED CONTROLLER (RBC).

	Occupied time (6AM - 10PM)	Unoccupied time (10PM - 6AM)
Supply chilled water setpoint	7°C	7°C
Supply hot water setpoint	65°C	65°C
Supply fan speed	30%	10%
Outdoor damper opening	Enable economizer, or between 20%~35%	0
Air purifier condition	Turn on only when $C_{PM,za} > 10 \mu\text{g}/\text{m}^3$	OFF

IV. RESULTS AND DISCUSSION

This section evaluates the performance of the RL controller and compares it with the rule-based controller, regarding energy expenditures, thermal comfort, and indoor air quality. Additionally, some patterns of the actions generated by the RL controller are identified and discussed. The results demonstrate that the RL controller can efficiently manage multiple control objectives in building energy management, and its learned control strategies are reasonable and consistent with domain expertise.

Fig. 2 compares the electricity and natural gas expenditure (in Korean won, KRW) between the rule-based controller (RBC) and the RL controller. Compared to RBC, the RL controller saves 14.7% of energy expenditure. The saving mainly occurs in the summer and shoulder seasons, but not in the winter season. This is partially because the heating/cooling load values are similar under the RBC and the RL control. However, the chiller COP is significantly affected by the chilled water temperature, while the efficiency of the gas boiler is constant, irrelevant of the hot water temperature. Therefore, the RL controller can increase the chilled water temperature when the cooling load is low to save the chiller power, but it has no means to save the natural gas consumption of the boiler. Another possible energy saving comes from the reduction of the supply fan speed when possible (e.g., both the cooling and heating loads are low).

Fig. 3 shows the indoor conditions of the office under the baseline and the RL control during the occupied time. Except for humidity control, the RL controller performs better than the baseline controller. The solid black boundaries in the figure indicate the thresholds used in the reward function (i.e., a penalty is given if the room air property goes beyond the thresholds). The RL agent basically controls the room air properties within the boundaries for no penalty. Sometimes, the zone air temperature can be close to its upper boundary, because allowing for a higher room temperature in summer can reduce energy usage. However, as a result, the room air properties could occasionally run over the boundaries a little bit, which is usually unavoidable in the RL control. Therefore, in the design of an RL controller, it is suggested to first define a tolerance limit for each property, as shown by the dotted black boundaries in Fig. 3, which is the constraint that should not be violated. Then, the boundaries to penalize the RL agent in the reward function need to be designed more conservatively.

Fig. 4 depicts a few typical relationships between the state and actions of the RL controller. Plots (a) and (b) show the control of the chilled water supply temperature. In plot (a), as the outdoor temperature rises above the cooling setpoint (24°C during the occupied time), the chilled water temperature gradually decreases to meet the increasing cooling load. Also, when the outdoor temperature is lower than the cooling setpoint, sometimes the chilled water temperature needs to be reduced for dehumidification. Plot (b) clearly shows that the outdoor humidity level significantly affects the control

of the chilled water temperature. As the outside becomes more humid, the RL controller reduces the chilled water temperature to dehumidify the processed air more effectively.

Fig. 4c shows the control of the outdoor air damper. A larger damper opening is always preferred to dilute the indoor CO₂ level, but not always beneficial for temperature and humidity control, and may worsen the indoor air quality if the outside is heavily polluted. Note that a hard constraint of 40% opening is placed when the outdoor temperature is less than 10°C, and another hard constraint of 30% opening is placed when the outdoor temperature is less than -10°C. When the room air temperature is higher than both the outdoor temperature and the heating setpoint, increasing the damper opening can help passively cool down the building (this is also called the economizer mode). Since the room temperature is controlled between 20°C and 25°C, the damper has the largest opening at around 18°C. As the outdoor temperature rises, the passive cooling effort becomes less effective. Especially, if the outdoor temperature moves above 25°C, the damper should maintain its minimum opening to reduce the ventilation load. On the other hand, when the outdoor temperature is lower than 18°C, the damper also needs to reduce its opening to avoid over-cooling the room.

In Fig. 4d, the supply fan control reflects the cooling/heating demand requested by the thermostat. When the outside is either very cold or very hot, the fan needs to increase its speed to meet the cooling/heating load. When the economizer mode is enabled at around 20°C, the fan also needs to work harder to bring in more fresh air to cool down the building. Fig. 4e shows that the trend of hot water temperature is very similar to the supply fan in the heating season. When more heating is requested, both the supply fan speed and the hot water temperature increase. Note that the hot water temperature does not affect the boiler's efficiency.

Finally, Fig. 4f shows the relationship between the outdoor PM2.5 level and the proportion of time that the air purifier is switched on. The plot displays that the air purifier turns on more frequently when the outdoor pollution becomes more severe.

V. CONCLUSION

This paper conducted a simulation study to evaluate the annual performance of RL for the co-optimization of energy consumption, thermal comfort, and indoor air quality in a small office building. In this process, we built a simulation platform that integrated dynamic models of the HVAC system, building envelope, indoor moisture transfer, indoor CO₂ dissipation, and indoor particulate transmission. The platform is able to train and evaluate different RL controllers, and the models can be calibrated to pre-train an RL controller before deploying it in a real building. During performance evaluation, we not only compared the performance of the RL controller to a baseline that is commonly adopted in the HVAC control, but also analyzed the control strategies learned by the RL agent. The results demonstrated that the RL controller can be trained to maximize energy saving while maintaining indoor thermal comfort and indoor air

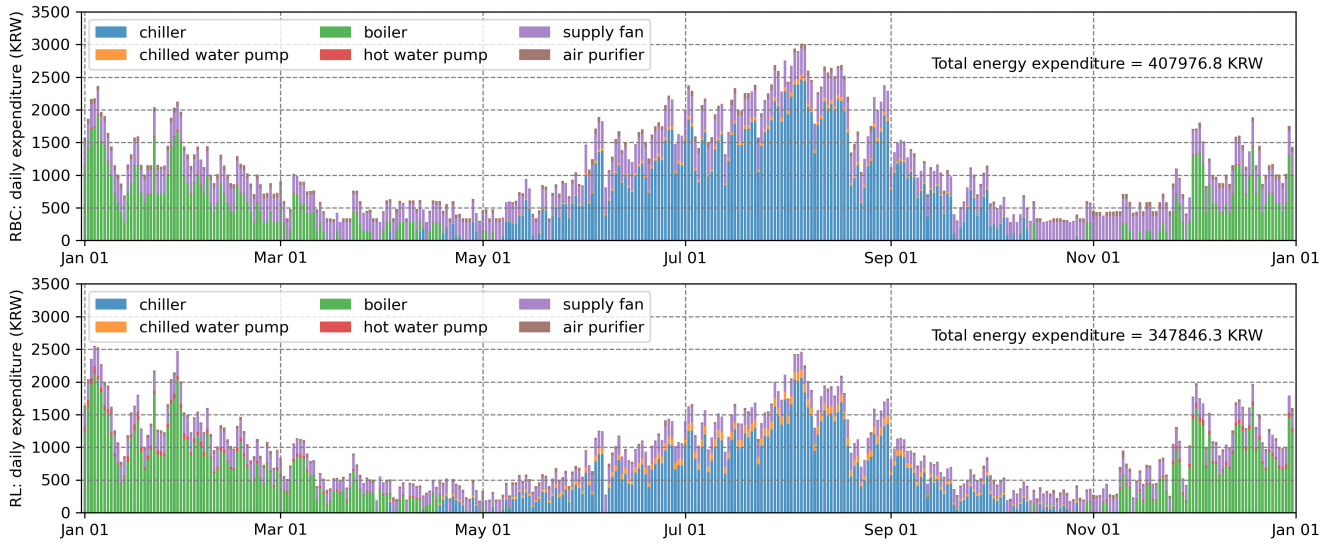


Fig. 2. Annual expenditure (in Korean won, KRW) of electricity and natural gas of the baseline and the RL controller. The RL controller saves 14.7% of energy expenditure.

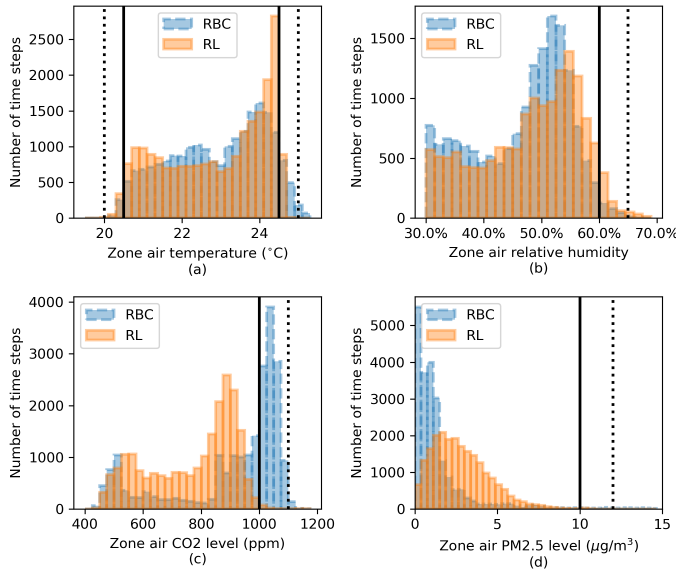


Fig. 3. Temperature, relative humidity, CO₂ level, and PM_{2.5} level in the office room under the baseline and the RL control during the occupied time. Solid black boundary: threshold for no penalty in the reward function. Dotted black boundary: tolerance limit for the control problem.

quality. Additionally, the trends it learned are consistent with the domain knowledge of human experts.

Future work can focus on the following aspects:

- Improve the design of the RL controller, such as allowing the RL to control supply air temperature, including forecast weather data and future electricity price data as state variables in demand response.
- Improve the simulation platform by integrating multi-zone building envelope models (e.g., 3R3C and 4R4C

gray-box models) and other types of HVAC systems (e.g., heat pumps and distributed energy resources).

- Evaluate and compare the performance of DDPG to other RL agents, such as Proximal Policy Optimization (PPO), Asynchronous Advantage Actor-Critic (A3C), and Soft Actor-Critic (SAC).
- Finally, transfer learning is an effective approach for generalizing trained RL agents to different building configurations. Future work could further explore leveraging transfer learning to adapt RL policies to new environments.

ACKNOWLEDGMENT

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20212020800120).

REFERENCES

- [1] M. Han, X. Zhang, L. Xu, R. May, S. Pan, and J. Wu, *A review of reinforcement learning methodologies on control systems for building energy*. Höskolan Dalarna, 2018. [Online]. Available: <https://urn.kb.se/resolve?urn=urn:nbn:se:du-27956>
- [2] M. Han, R. May, X. Zhang, X. Wang, S. Pan, D. Yan, Y. Jin, and L. Xu, "A review of reinforcement learning methodologies for controlling occupant comfort in buildings," *Sustainable Cities and Society*, vol. 51, p. 101748, Nov. 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2210670719307589>
- [3] Z. Wang and T. Hong, "Reinforcement learning for building controls: The opportunities and challenges," *Applied Energy*, vol. 269, p. 115036, July 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0306261920305481>
- [4] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied Energy*, vol. 235, pp. 1072–1089, Feb. 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261918317082>

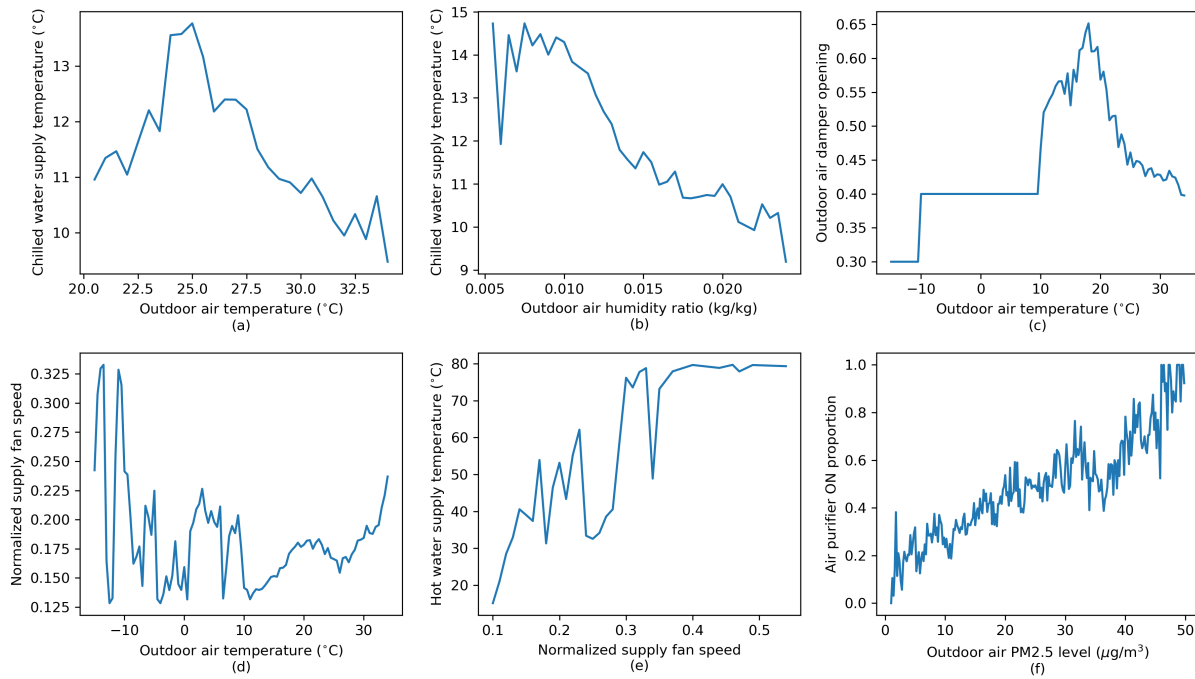


Fig. 4. Some typical behaviors learned by the RL controller. In general, the trends are consistent with the domain knowledge of human experts.

- [5] X. Deng, Y. Zhang, Y. Zhang, and H. Qi, "Towards optimal HVAC control in non-stationary building environments combining active change detection and deep reinforcement learning," *Building and Environment*, vol. 211, p. 108680, Mar. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360132321010702>
- [6] S. Touzani, A. K. Prakash, Z. Wang, S. Agarwal, M. Pritoni, M. Kiran, R. Brown, and J. Granderson, "Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency," *Applied Energy*, vol. 304, p. 117733, Dec. 2021. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0360261921010801>
- [7] G. Gao, J. Li, and Y. Wen, "Deepcomfort: Energy-efficient thermal comfort control in buildings via reinforcement learning," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8472–8484, 2020.
- [8] C. Blad, S. Bøgh, and C. S. Kallesøe, "Data-driven Offline Reinforcement Learning for HVAC-systems," *Energy*, vol. 261, p. 125290, Dec. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360544222021739>
- [9] X. Lu, Y. Fu, and Z. O'Neill, "Benchmarking high performance hvac rule-based controls with advanced intelligent controllers: A case study in a multi-zone system in modelica," *Energy and Buildings*, vol. 284, p. 112854, 2023.
- [10] Y. Zhang and Q. Zhao, "Energy saving algorithm of hvac system based on deep reinforcement learning with modelica model," in *2022 41st Chinese Control Conference (CCC)*, 2022, pp. 5277–5282.
- [11] Y. An, T. Xia, R. You, D. Lai, J. Liu, and C. Chen, "A reinforcement learning approach for control of window behavior to reduce indoor pm2.5 concentrations in naturally ventilated buildings," *Building and Environment*, vol. 200, p. 107978, 2021.
- [12] T. Yang, L. Zhao, W. Li, J. Wu, and A. Y. Zomaya, "Towards healthy and cost-effective indoor environment management in smart homes: A deep reinforcement learning approach," *Applied Energy*, vol. 300, p. 117335, Oct. 2021. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0360261921007431>
- [13] Y. Du, F. Li, J. Munk, K. Kurte, O. Kotevska, K. Amasyali, and H. Zandi, "Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control," *Electric Power Systems Research*, vol. 192, p. 106959, Mar. 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378779620307574>
- [14] S. Xia, P. Wei, Y. Liu, A. Sonta, and X. Jiang, "RECA: A Multi-Task Deep Reinforcement Learning-Based Recommender System for Co-Optimizing Energy, Comfort and Air Quality in Commercial Buildings," in *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. Istanbul Turkey: ACM, Nov. 2023, pp. 99–109. [Online]. Available: <https://dl.acm.org/doi/10.1145/3600100.3623735>
- [15] Lawrence Berkeley National Laboratory, "FLEXLAB: The world's most advanced integrated building and grid technologies testbed," <https://flexlab.lbl.gov/>, 2021, accessed: 2025-01-25.
- [16] M. Wetter, W. Zuo, T. S. Noudui, and X. Pang, "Modelica Buildings library," *Journal of Building Performance Simulation*, vol. 7, no. 4, pp. 253–270, 2014. [Online]. Available: <https://doi.org/10.1080/19401493.2013.765506>
- [17] T. Noudui, M. Wetter, and W. Zuo, "Functional mock-up unit for co-simulation import in EnergyPlus," *Journal of Building Performance Simulation*, vol. 7, no. 3, pp. 192–202, 2014, publisher: Taylor & Francis.
- [18] C. Andersson, J. Åkesson, and C. Führer, "Pyfmi: A python package for simulation of coupled dynamic models with the functional mock-up interface," 2016, publisher: Centre for Mathematical Sciences, Lund University Lund, Sweden.
- [19] American Society of Heating Refrigerating and Air-Conditioning Engineers, *2001 Ashrae handbook : fundamentals (SI)*. ASHRAE., 2001.
- [20] "Standard guide for using indoor carbon dioxide concentrations to evaluate indoor air quality and ventilation," in *ASTM Standard D6245-12*. ASTM International West Conshohocken, PA, 2012.
- [21] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," July 2019, arXiv:1509.02971 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1509.02971>
- [22] F. Guo, S. woo Ham, D. Kim, and H. J. Moon, "Deep reinforcement learning control for co-optimizing energy consumption, thermal comfort, and indoor air quality in an office building," *Applied Energy*, vol. 377, p. 124467, 2025.