Demonstration of Intelligent HVAC Load Management With Deep Reinforcement Learning

By Yan Du, Fangxing Li, Kuldeep Kurte, Jeffrey Munk, and Helia Zandi

BUILDINGS ACCOUNT FOR 40% OF TOTAL PRIMARY energy consumption and 30% of all CO_2 emissions worldwide. A large portion of building energy consumption is due to heating, ventilation, and air-conditioning (HVAC) systems. In the summer, for example, more than 50% of a building's electricity consumption is used for cooling. With proper energy management, buildings can provide load shifting, peak shaving, frequency regulation, and many other demand response services.

Many existing approaches for building energy management are model based and require the modeling of the complex thermal dynamics of the HVAC system and its interaction with the ambient environment. The development of such a model may introduce measurement and prediction errors, which may undermine the control performance. In addition, models developed for one building may not generalize well for other buildings or unseen operation environments.

In contrast to the model-based approaches, model-free algorithms require no prior knowledge of the physical model, such as the thermal-dynamic model in the HVAC control case; rather, they learn the model through estimation and exploration. One representative model-free approach is reinforcement learning (RL). As shown in Figure 1, an agent (e.g., a demand response controller) interacts with an environment (for instance, the building). At each control time

Digital Object Identifier 10.1109/MPE.2022.3150825
Date of current version: 19 April 2022

1540-7977/22©2022IEEE

Real-World Experience of Machine Learning in Demand Control

interval, the agent observes the state of the environment and takes an action, making the environment transit from one state to another. The environment can send feedback in the form of a reward to the agent. The agent improves its action based on that reward.

This process repeats until the environment reaches the terminating state. The goal of the agent is to learn an action policy that maximizes its cumulative reward from all states. In this learning process, the agent does not rely on prior knowledge to make decisions, but it gradually formulates an optimal action policy through intelligent "trial-and-error." As a result, the model-free approach has greater flexibility in solving control and optimization problems with unknown models or partial observabilities.

Deep RL, which is a combination of deep learning and RL, is a more recently developed approach. The key idea behind deep RL is the adoption of a deep neural network (DNN) to let the agent learn an optimal policy. As shown in Figure 1, the DNN is a neural network with multiple hidden layers. With the given input as the state, the DNN can output either the estimated value of the actions or the optimal action (depending on the specific deep RL approach) at the current state due to the strong feature extraction ability of its multiple hidden layers. Deep RL is model free since the DNN is trained through interacting with and without prior knowledge of the environment. In addition, the deep RL approach also has high generalization abilities for new environments. A well-trained DNN can be regarded as a function with fine-tuned parameters. Whether a state as input is seen before or not, a well-trained DNN can always generate an output.

In this article, we introduce an end-to-end workflow for developing a deep RL-based residential HVAC controller that can control multiple zones, where the zones represent different floors in a house. In particular, we describe two deep RL approaches for HVAC control, present the evaluation of the deep RL-based HVAC control strategies through simulation studies, and discuss the deployment of a deep RL-based HVAC control approach in a real-world residential house. Finally, we analyze the deployment results and provide conclusions.

By optimally designating the setpoint of the HVAC system, the energy cost can be minimized while keeping the indoor temperature within the user's comfort range.

The Deep RL Approach for HVAC Control

Schematic Description of Deep RL for the HVAC Control Problem

A critical premise for applying deep RL approaches is that the problem under investigation is a time-sequential decision-making process. At each time step, the current state is only related to the previous state, and the optimal decision can be made based only on the current state information. This is the case for the optimal control of a multizone HVAC system. The indoor temperature at the current state is only related to the parameters at the previous time interval, and it is not affected by the indoor temperature at earlier intervals. By optimally designating the setpoint of the HVAC system, the energy cost can be minimized while keeping the indoor temperature within the user's comfort range. Without losing generality, in the following discussion, we assume that all HVAC zones need heating. Also, a zone can simply be regarded as a floor in a house in this study.

When applying deep RL approaches, four essential elements should be first defined: the state (s), action (a), state transition probability (p), and reward (r). In the context of a multizone residential HVAC control problem, the state includes the following factors:

- ✔ the time of day
- ✔ the current indoor temperature
- ✔ the current outdoor temperature
- ✔ a 6-h look-ahead outdoor temperature series for planning
- ✔ the current retail price
- ✔ a 6-h look-ahead retail price series
- ✔ the lower bound of the user comfort level
- ✔ the maximum retail price within the next 6 h
- ✔ the length of time to reach the next price peak.

The action is the setting of the HVAC system setpoint. It can be either discretely or continuously adjusted within a certain range. The reward is defined as the negative sum of the energy consumption cost and comfort violation cost for the control interval, and the comfort violation cost is calculated based on how many degrees the indoor temperature deviates from the user comfort level. The environment is the entire building or house including the HVAC system.

This process of applying deep RL for HVAC control is illustrated in Figure 1. Note that we do not define the state transition probability for the process. The *probability* refers to the probability of transitioning to a specific next state after taking an action at the current state. If the state transition probability model is known, the HVAC control problem can be explicitly formulated and solved analytically. However, obtaining an accurate state transition probability model for the HVAC control problem is not a trivial task. This is because the thermal-dynamic model of buildings with HVAC systems is related to a variety of parameters, including resistances and capacitors from different building components; weather factors, such as outdoor temperature and solar irradiance; and so on.

Therefore, as previously mentioned, a model-free approach like RL is more suitable for solving the HVAC control problem. Further, the building models can vary, and we need a more generalized and robust HVAC control approach that can work efficiently

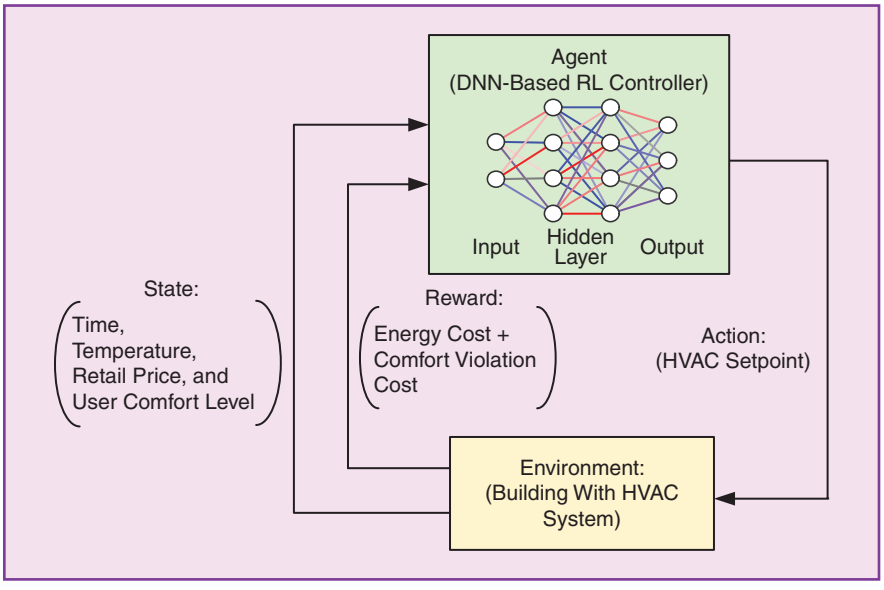

figure 1. The RL-based approach for HVAC control. DNN: deep neural network.

in different building environments. Deep RL, with the powerful embedded DNN providing more generalization and adaptability, is an ideal approach for achieving a flexible and intelligent HVAC control strategy. More details of the deep RL approach will be discussed in the next sections.

Deep-Q Network for HVAC Control

The deep-Q network (DQN) is a combination of Q-learning and a DNN. The main idea behind the DQN is to build a neural network to estimate the optimal control action in a discrete domain. The neural network functions like a complex lookup table, with the input being a state and the output being the action values for all of the possible control actions at the current state. In the DQN, this action value is called the Q-*value*, where a higher value indicates a better or more effective action. This DQN approach selects the action with the highest Q-value evaluated by the neural network.

The DQN approach adopts a stabilization strategy to train two neural networks simultaneously, namely, the target and behavior networks. The function of the target network is to provide stable-labeled samples for the behavior network to learn. The DQN converges when the outputs from the two networks are close to each other. An overview of the DQN approach for HVAC control is shown in Figure 2(a).

Deep Deterministic Policy Gradient for HVAC Control

The deep deterministic policy gradient (DDPG) for HVAC control is specially designed for solving problems with continuous variables. In the described DQN control, the neural network outputs all of the possible action values, and the number of action values generated is limited. As a result, the algorithm processes discrete actions. In the case of HVAC control, given the control action as the setpoint and an example range for the setpoint from 20 to 22 °C, we need to discretize the proposed range. For instance, if we set the step size as 1 °C, then the potential action set becomes {20 °C, 21 °C, 22 °C}, and the algorithm learns to choose from the three actions. In contrast, the DDPG algorithm does not require the action space to be discretized. Given the setpoint range as [20 °C, 22 °C], the algorithm can generate one continuous number within the range. The ability to handle a continuous action space makes the DDPG algorithm more suitable for solving problems where the control action is a continuous variable.

The DDPG algorithm can be regarded as an extension of the DQN algorithm. In the algorithm, there are two types of neural networks applied: the actor and critic networks. The actor network outputs a deterministic control action based on the given current state. The critic network outputs the Q-value based on both the state and action provided by the actor network. The actor network is further updated by maximizing the Q-value under the current policy, and the critic network is updated by minimizing the mean square error of the Q-value, which is the same as the DQN algorithm. In

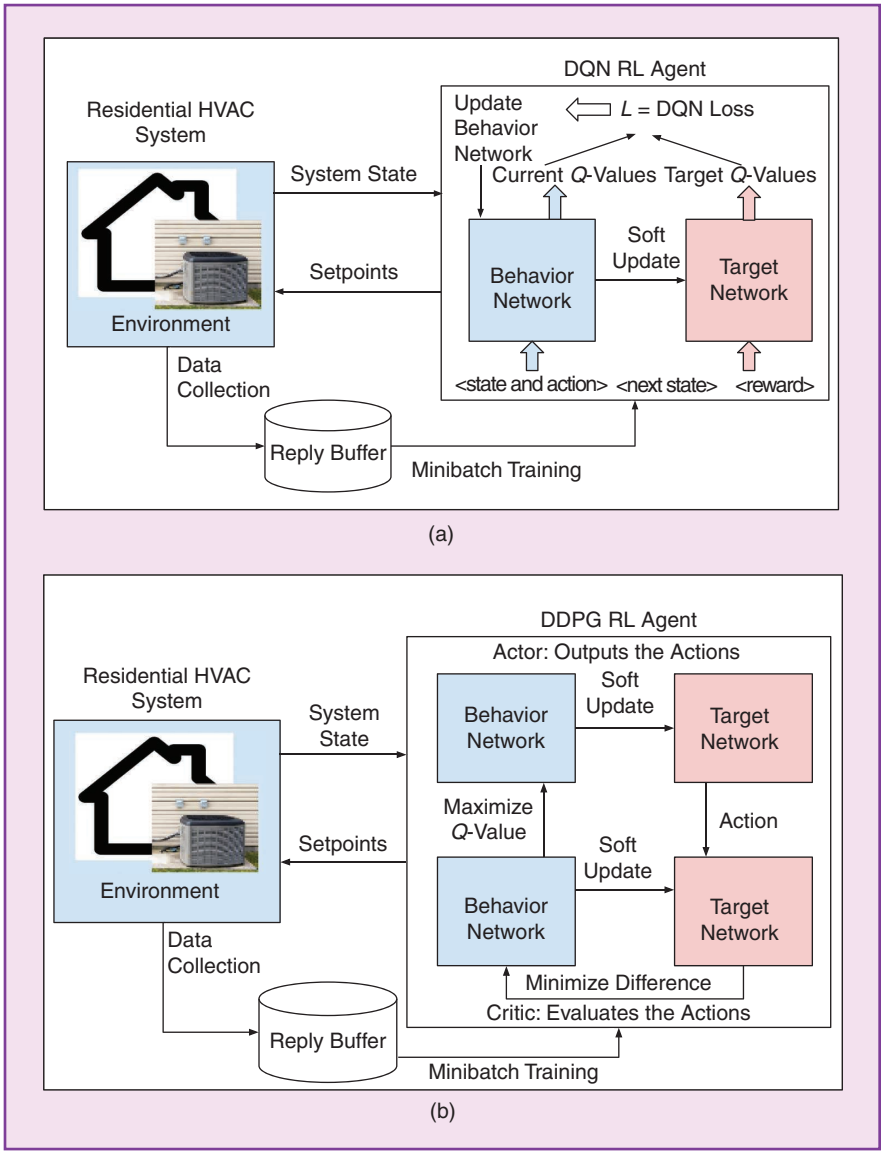

figure 2. A multizone HVAC control framework with (a) a DQN and (b) a DDPG. DDPG: deep deterministic policy gradient.

addition, the DDPG algorithm adopts a stabilization strategy by implementing the behavior and target networks for both the actor and critic networks, which is similar to the DQN algorithm. An overview of the DDPG approach for HVAC control is shown in Figure 2(b).

Evaluation of a Deep RL-Based HVAC Control Strategy Through Simulation

Our main objective is to evaluate the performance of the deep RL-based HVAC control strategy in a real-world environment. However, one cannot deploy an RL algorithm directly in a real-world environment and allow it to learn from scratch for two main reasons:

- ✔ RL-based approaches are essentially trial-and-error methods (albeit with high intelligence) in which an RL algorithm interacts with the environment (i.e., the building or house) and learns from it based on the reward (i.e., operation cost) from the environment. Depending on the application, an RL algorithm might need long experience to learn how to behave optimally.
- ✔ During the initial learning phase, an RL algorithm tends to take random actions to explore and understand the environment. However, homeowners will not be happy if RL designates random setpoints on their thermostats. Therefore, instead of directly deploying an RL algorithm from scratch in a real-world environment, like a building or house, we train and validate it in a simulation environment as a starting point, allowing for faster development and overcoming the challenge of training RL in a real-world situation. Once we are satisfied with its performance in the simulation, we can deploy the pretrained RL model (the trained and deployable RL algorithm) in a real house.

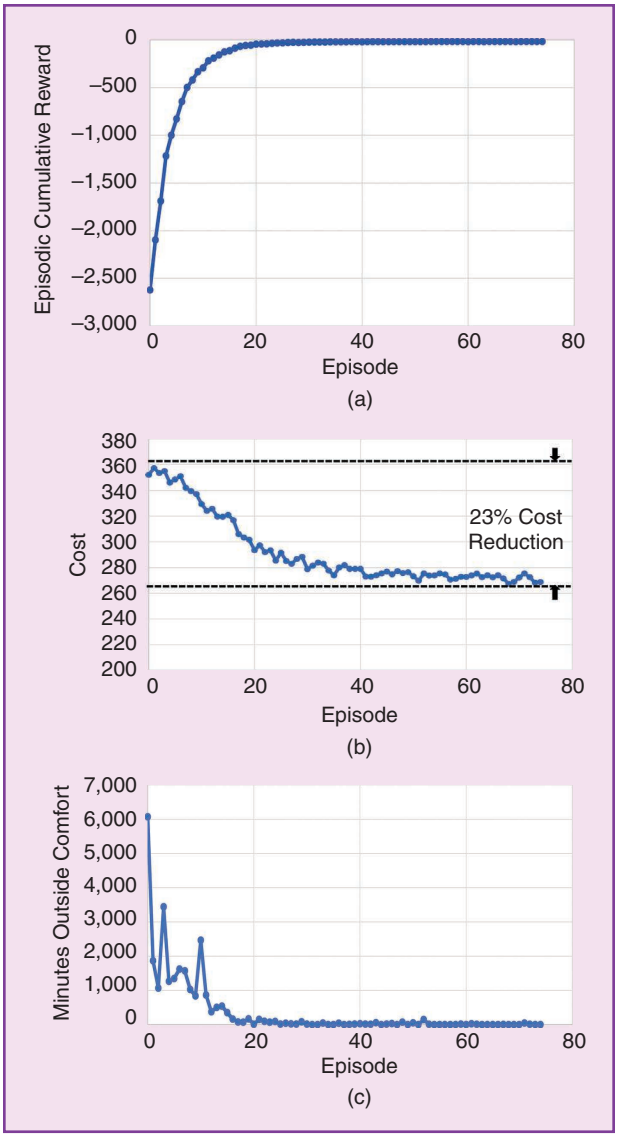

figure 3. The training performance of the DQN algorithm with the discrete control strategy: the (a) episodic cumulative reward, (b) cost of operation, and (c) minutes outside the comfort level.

In this section, we introduce the training and validation of a deep RL-based multizone residential HVAC control strategy on a simulation testbed with real-world data. Performance comparisons with benchmark control strategies demonstrate the efficiency and generalization ability of model-free deep RL approaches.

Simulation Setup

The simulated HVAC building model requires weather-related and price data. The weather data are taken from typical meteorological year (TMY) data from 2019 to 2020 in Knoxville recorded by the National Renewable Energy Laboratory. For price data, the time-of-use price signals with a peak price at US$0.25/kWh and an off-peak price at US$0.05/kWh are applied. These input data sets are applied to a building simulation software testbed for the following training and validation of deep RL-based HVAC control strategies.

Training and Validation of the DQN for HVAC Control

We first present the simulation result of the DQN algorithm for multizone residential HVAC control. For training the algorithm, the Knoxville TMY data from 21 December 2019 to 10 March 2020 were utilized. The simulation step of the HVAC thermal dynamics is 1 min, and for every 5 min, the algorithm provides a setpoint control action. The user comfort level is set to 20–22.22 °C (i.e., 68–72 °F). The state information used in the DQN control includes all of the elements as listed in the "Deep RL Approach for HVAC Control" section. The DQN approach generates a discrete control action by adjusting the setpoint with a fixed step within the user comfort level.

The DQN algorithm was trained for 75 episodes with these settings. As mentioned earlier, we used approximately three months of data from 21 December 2019 to 10 March 2020 for training the DQN algorithm. In this training process, a single training episode is considered complete after the DQN algorithm has explored the three-month data. Next,

we repeated the same process for several episodes until the DQN algorithm converged. Within each episode, at every control step, the DQN algorithm observes the environment through various features (called *states*) and then takes actions that are implemented in the environment. For example, our DQN algorithm provides a setpoint as the action, which leads the environment to evolve into the next state, and then the same procedure repeats. At the end of each control step, the environment provides the reward in response to the DQN algorithm's action. This reward could be the electricity cost incurred due to the DQN algorithm's action. Based on this reward, the DQN algorithm then adjusts its Q-table. Readers may refer to Figure 2 and the relevant description for more details on DQN algorithm training.

It is important to evaluate the DQN algorithm's training performance. Throughout the training, we keep track of the episodic average reward, total electricity cost of operating HVAC using the DQN for three months, and comfort violations that count the minutes during which the indoor temperature violated the user comfort level. The average reward, cost of operation, and minutes outside of the user comfort level during the training are shown in Figure 3. In Figure 3(a), the episodic cumulative reward gradually increases as the training proceeds and stabilizes in the end. A 23% cost reduction is observed by the end of the training session in Figure 3(b). The minutes out of comfort also show a decreasing trend and remain at zero by the end of the training in Figure 3(c). These observations confirm the convergence of the DQN-based HAVC control.

The control performance of the pretrained DQN algorithm is further validated in unseen scenarios: two half-month Knoxville TMY data sets from 1–20 December 2019 and 11–31 March 2020, which were not used during the training stage. A fixed-setpoint control strategy is designed as a baseline case for a comparison with the DQN-based HVAC control strategy. In the baseline case, the heating setpoint is always kept at 20 °C.

For brevity, we present the indoor temperature variations for operating the pretrained DQN-based HVAC controller on 14–15 December in Figure 4. An important observation here is that the DQN control has learned a preheating strategy. The DQN control preheated the zones before the peak price. This was beneficial during the initial hours of the peak period where the power consumption was zero, as shown by the power consumption graph in Figure 4. This is how the DQN control achieved cost savings. The associated daywise energy cost comparison between the DQN control and baseline cases for 1–20 December is shown in Figure 5. We observed that the DQN approach achieved a >32% cost savings over the fixed-setpoint baseline for both the 1–20 December and 11–31 March data. Full details of the entire duration of 1–20 December and 11–31 March are not plotted due to space limits.

Training and Validation of the DDPG for HVAC Control

For the DDPG algorithm, the Knoxville TMY data from 1–30 November 2019 are utilized for training. The control interval of the DDPG algorithm is 60 min. The state information input to the DDPG algorithm includes the current indoor temperature for each zone, outdoor temperature, and retail price as well as the lower bound of the user comfort level (Table 1). The DDPG algorithm directly generates a deterministic, continuous setpoint for each zone of the building's HVAC system. The range of the setpoint is designated to be the same as the user comfort temperature zone.

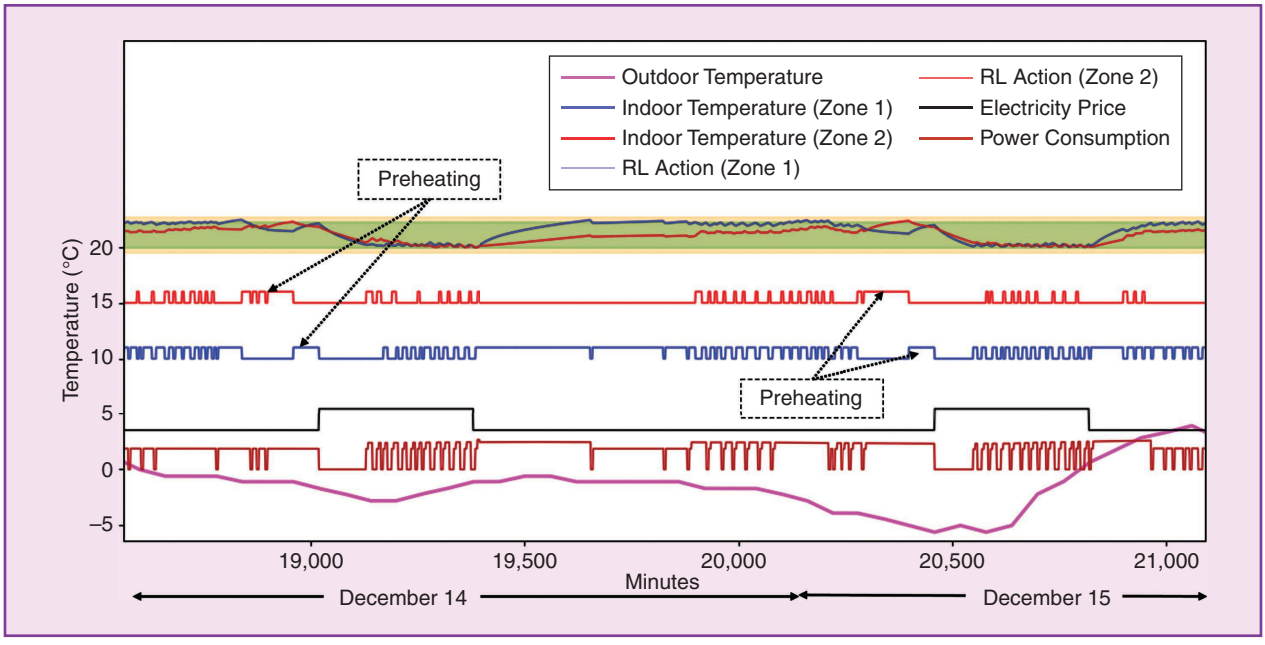

figure 4. The validation of indoor temperature variation for 14 and 15 December with the pretrained DQN model.

We may conclude that the DDPG control can effectively solve unseen physical environments and provide an efficient and flexible HVAC control strategy after its offline training.

The DDPG algorithm is trained for 300 episodes. After the training, the pretrained DDPG algorithm is further validated in two unseen scenarios: 1) with the Knoxville TMY data from 1–20 January 2020 and 2) with the same weather data as the first scenario but with 10 building models that have different thermal mass parameters from the simulation testbed. The DDPG approach is compared with two benchmark control strategies: 1) a rule-based case, where the temperature is set at the lowest during the peak price hours and the highest during the off-peak price hours to achieve the preheating effect to reduce energy costs, and 2) a fixed-setpoint case, where the setpoint is always at the highest value of the setpoint range to avoid violation of user's comfort level.

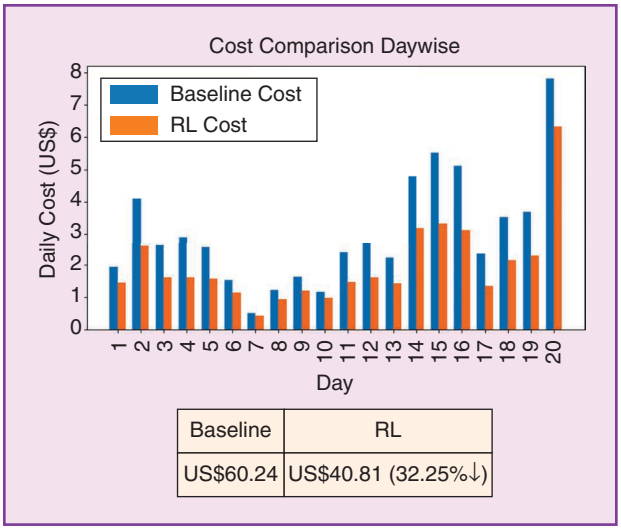

figure 5. A daywise cost comparison of the DQN-based HVAC control with the fixed-setpoint baseline for 1–20 December 2019.

table 1. The daily user comfort level.

Time Period	0:00–6:00	6:00–12:00	12:00–18:00	18:00–24:00
User comfort level: lower bound (°C)	18	17	18	19
User comfort level: upper bound (°C)	20	19	20	21

table 2. A performance comparison of three HVAC control approaches.

Control Approach	DDPG	Rule Based	Fixed Setpoint
Total energy cost (US$)	55.21	39.08	71.48
Minutes out of user comfort level	48	2,617	0
Average comfort violation (°C)	0.13	1.85	0

In the first scenario, the final optimized results of the DDPG algorithm and benchmark cases are shown in Table 2 in which the total energy cost is the accumulated energy cost over the 10 test days. The average comfort violation shows the average value, in degrees, by which the indoor temperature is lower than the setpoint. Table 2 shows that the rule-based case has the lowest total energy cost because of its temperature setting logic based on the peak or off-peak price. However, this control strategy may result in a severe comfort violation because it always designates the temperature setpoint to the lowest value at peak price hours. In contrast, since the temperature is always set at the highest value in the fixed-setpoint case, there is no temperature violation of the user's comfort level. Meanwhile, the energy cost is also the highest among the three control strategies.

Since the temperature is always set at the highest value in the fixed setpoint case, there is no temperature violation of user's comfort level. Meanwhile, the energy cost is also the highest among the three control strategies. The setpoint settings and the associated indoor temperature variations based on the three approaches are illustrated in Figure 6. For each approach, its control strategies for zones 1 and 2 share similar patterns. Therefore, for the sake of simplicity, we plot only the control strategies of each approach in zone 1 in the figure and the control results in the first five days as representative values.

In all parts of Figure 6, the hour-by-hour yellow rectangular bars represent the user comfort level as acceptable temperature ranges, which correspond to Table 1. In Figure 6(a), it can be observed that the DDPG-based control will designate the setpoint at a relatively low value during the peak price hours and at a relatively high value during the off-peak hours. As such, the DDPG-based control can achieve the preheating effect and reduce energy costs during the winter.

Figure 6(b) shows the results of the rule-based case, in which the control strategy designates the setpoint at the lowest value during peak price hours and the highest value during off-peak hours. When the outdoor temperature is extremely low, this control strategy results in severe indoor

temperature violation (i.e., between 5,000 and 6,000 min in the figure), mainly due to the low setpoint. In contrast, the DDPG-based control strategy does not place the setpoint at the lowest possible value even during peak price hours to avoid comfort violations during the time of extremely low outdoor temperatures. These comparisons show that, once well trained, the DDPG control has learned the impacts of the price signal and outdoor temperature on the reward, and it develops an intelligent setpoint control strategy to accommodate both the price peak and low outdoor temperature.

Figure 6(c) shows the fixed-setpoint case, so-called because this control strategy always sets the temperature at the highest value. Thus, the indoor temperature also remains at the highest level among the three control strategies. Consequently, this fixed-setpoint control leads to the highest energy cost.

Note that, when calculating temperature or comfort violations, only the time when the indoor temperature is below the lower bound is counted. The reason is that this is a heating scenario, and a low indoor temperature is considered an unbearable violation, while a high indoor temperature is acceptable to residential HVAC users.

In the second scenario, the pretrained DDPG control strategy is validated with 10 unseen building models with different thermal mass parameters to demonstrate its generalization ability. Table 3 shows a comparison of the energy costs and temperature violations for the DDPG control and two aforementioned benchmark controls. As the table shows, like in scenario 1, the rule-based control gives the lowest energy cost, while the fixed-setpoint control gives the fewest violations. The pretrained DDPG control achieves a balanced HVAC control strategy that results in a relatively lower energy cost and fewer

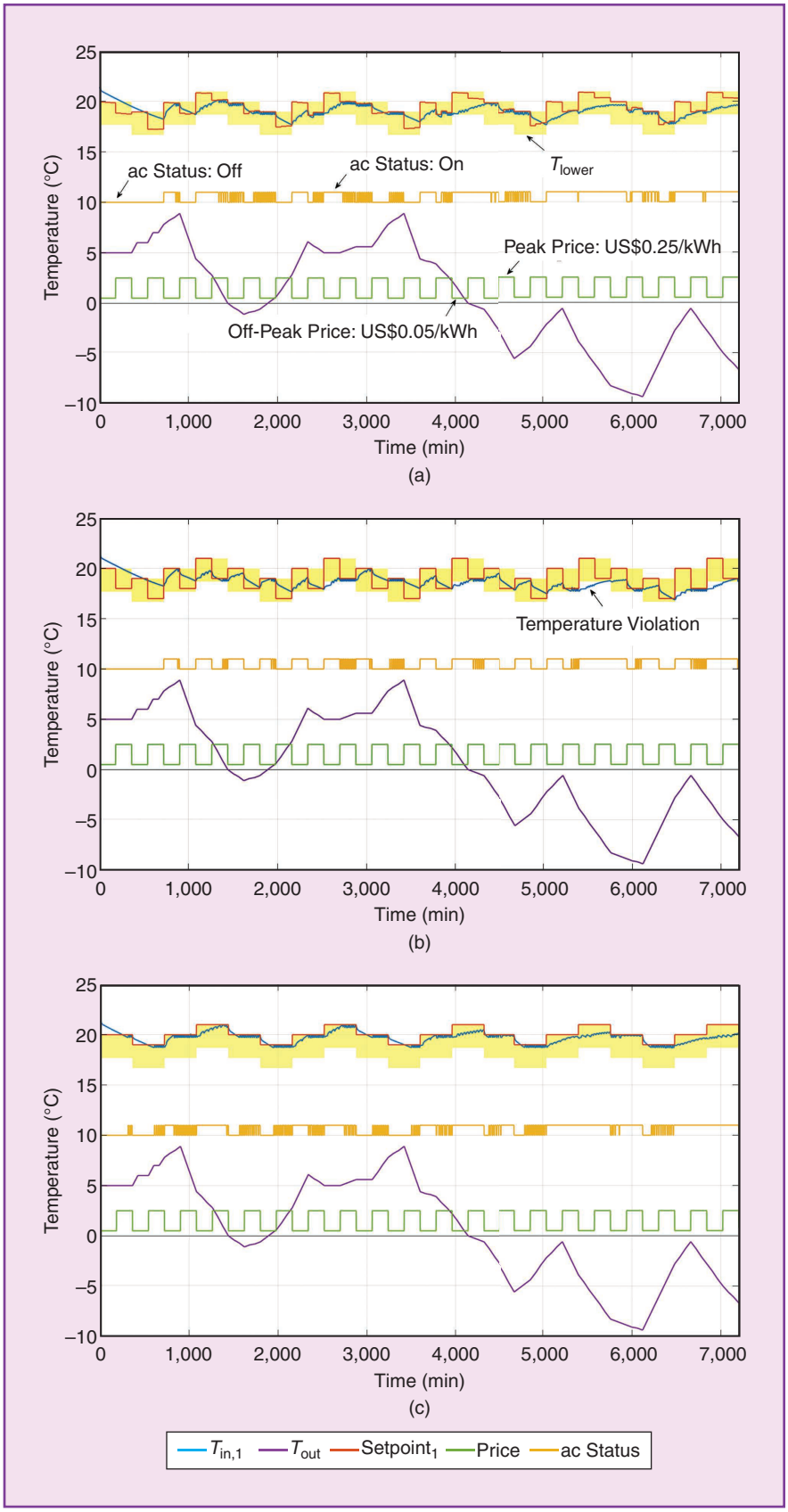

figure 6. A comparison of the setpoint control strategies for the first five test days (zone 1): (a) DDPG, (b) rule based, and (c) fixed setpoint.

violations. Thus, we may conclude that the DDPG control can effectively solve unseen physical environments and provide an efficient and flexible HVAC control strategy after its offline training.

From these simulation studies, both the DQN and DDPG controls demonstrate a better HVAC control performance compared with conventional approaches, like a fixed-setpoint or simple rule-based control strategy, which implies their considerable potential for online deployment. As shown in Table 3, the DDPG-based RL algorithm shows energy cost savings in the range of 19–29% (average = 25.9%) compared to the fixed-setpoint control in the simulation.

Since the DQN-based RL approach achieves slightly higher energy savings (e.g., >32% energy cost savings) compared to the fixed-setpoint control, we selected the DQN-based RL algorithm as the HVAC control strategy to deploy in the real house. The detailed deployment process and experiment results are introduced in the next section.

Deployment of a Deep RL-Based HVAC Control Strategy

To further validate the control performance of the deep RL approach in real-world scenarios, the pretrained DQN control for controlling the multizone HVAC system was deployed at the Yarnell Station research house in Knoxville, Tennessee, as shown in Figure 7. The house is equipped with a two-stage heat pump, two-zone control system, and two smart thermostats. The two-story house is zoned by floor with a smart thermostat located centrally on each floor. Supply air dampers are used to control the delivery of conditioned air from the heat pump to the appropriate zone(s) based on the call for conditioning from the thermostats.

The zone controller manages the staging of the heat pump and indoor airflow rate, which is adjusted based on the number of zones calling for conditioning and the "size" settings of those zones (set using jumpers on the control board during the commissioning of the system). The staging of the heat pump is controlled based on a supply air temperature sensor located in the duct downstream of the air handler. The staging is controlled to maintain a heating mode supply air temperature of at least 32.2 °C (90 °F). The power consumption of the heat pump, therefore, is dependent on the outdoor air temperature, indoor temperature, and combination of zones calling for conditioning. During some condition combinations, it may only be possible to elicit a single-stage response from the two-stage heat pump. With these many different conditions affecting the achievable power response, the system is difficult to accurately model, making it a challenging real-world application for the deep RL approach.

table 3. A comparison of results for different control strategies

	DDPG		Rule-Based		Fixed Setpoint	
Building Index	Cost (US$)	Comfort Violation (min)	Cost (US$)	Comfort Violation (min)	Cost (US$)	Comfort Violation (min)
1	42.22	31	27.78	1,296	57.98	0
2	44.13	41	29.22	1,586	60.13	0
3	52.14	45	36.51	2,347	68.52	0
4	59.66	101	43.94	3,364	75.61	0
5	45.84	41	31.3	1,879	62.91	0
6	42.49	39	27.68	1,398	59.06	0
7	37.47	24	23.51	1,012	53.42	0
8	61.21	81	45.42	3,520	76.44	0
9	35.34	25	21.98	818	49.9	0
10	43.19	59	28.41	1,323	58.46	0

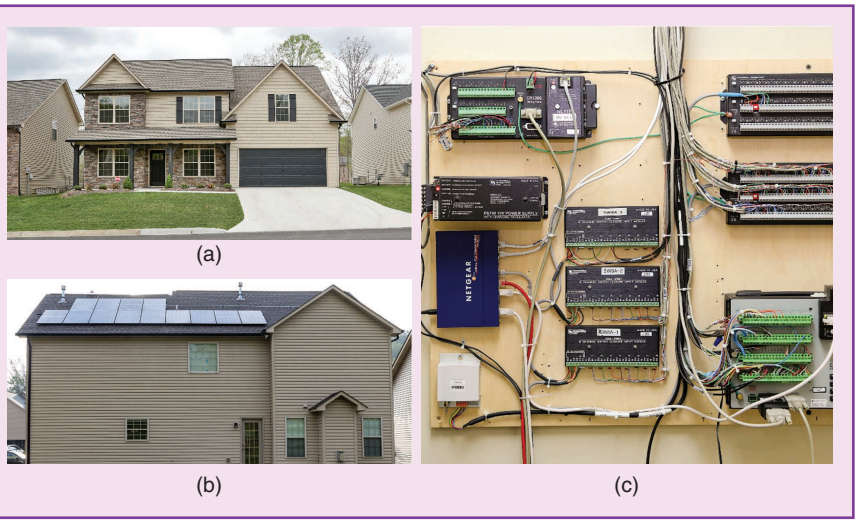

figure 7. The Yarnell Station research house in Knoxville, Tennessee, United States: the (a) front view, (b) back view, and (c) data acquisition system. (Source: Oak Ridge National Laboratory, https://www.ornl.gov/content/smart-buildings; used with permission.)

The DQN control strategy was deployed at the research house from 25 March to 5 April 2021. Since this period often has mild temperatures, the user comfort level was shifted higher than usual, i.e., 22.2–24.4 °C (72–76 °F), to ensure the home would experience adequate heating load. The fixed-setpoint baseline case is used to compare the DQN control strategy's performance during the deployment phase. In the baseline case, the setpoint is fixed to 22.2 °C (72 °F) to minimize the energy cost while maintaining the user comfort level. However, it is not feasible to deploy the fixed-setpoint control under the same weather conditions in the same research house as the DQN deployment, so it is not straightforward to have a truly fair comparison between the DQN and baseline controls.

Instead, we can create a "simulated DQN" case using the recorded thermostat setpoints and measured weather data during the DQN deployment. This allows us to evaluate the accuracy of the DQN simulation by comparing the measured to simulated results. More importantly, this makes a more direct comparison between the "simulated DQN" case and the baseline (fixed-setpoint) simulation case. The model parameters were adjusted based on the comparison between the measured data from the DQN deployment and "simulated DQN" control to increase the accuracy of the simulation. In the following discussion, we present the details of this procedure and performance comparison.

The procedure of the comparison study, which is based on the DQN deployment, is as follows:

- ✔ First, we fine-tune the building simulation model used for DQN training with the new data collected during the DQN deployment period.

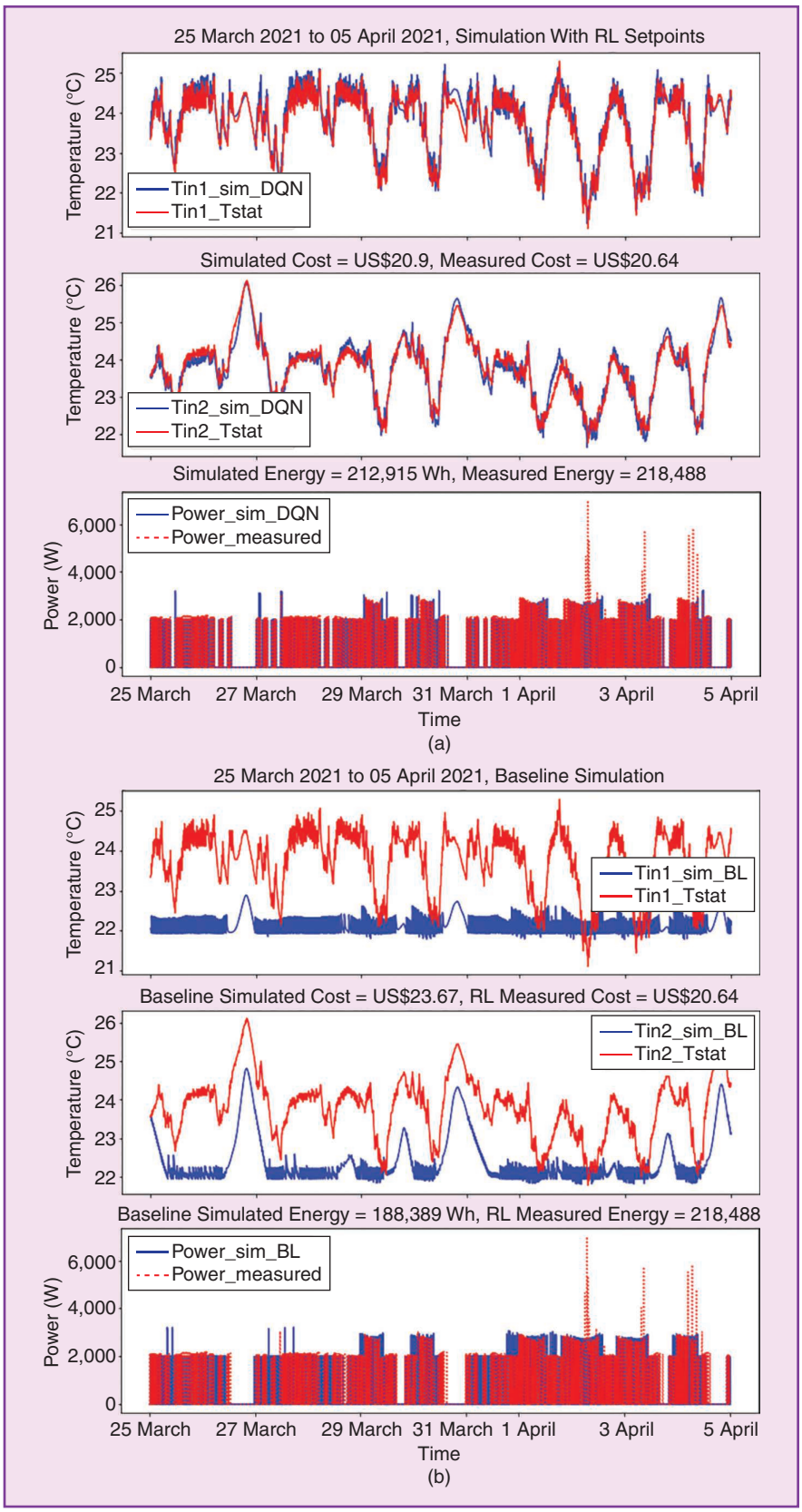

figure 8. A comparison of the simulation and deployment results: (a) a simulated DQN case versus measured data for DQN control and (b) a simulated baseline case versus measured data for DQN control. BL: baseline.

With these many different conditions affecting the achievable power response, the system is difficult to accurately model, making it a challenging real-world application for the deep RL approach.

- ✔ Next, we re-create the DQN's deployment (in simulation) by running the DQN's setpoints through this building simulation for the duration of 1 March–5 April 2021. This is referred to as the "simulated DQN" case, stemming from the actual DQN deployment.
- ✔ Then, we use this fine-tuned building simulation for the baseline case by using a fixed setpoint of 22.2 °C for the same duration. This later run mimics the deployment of the fixed-setpoint baseline control in the research house during the deployment period.

It should be noted that, since we do not have any measured data for the baseline case to use for setting the initial temperatures in the simulation, we instead simulate an additional 3.5 weeks (i.e., 1–24 March 2021) of baseline operation before the period of interest. Simulating this additional time allows the modeled temperatures to stabilize to realistic values and minimizes the effect of any error associated with the selected initial temperatures.

The results from the "simulated DQN" case using the retrained model and baseline simulation case employing the fixed-setpoint control are shown in Figure 8(a) and (b), respectively. In the top plot of Figure 8(a), the *Tin1_sim_DQN* curve (in blue) shows the temperatures of the "simulated DQN" case of the first floor in the research house, while *Tin1_sim_BL* (in blue) in Figure 8(b) shows the temperatures of the simulated baseline case (i.e., fixed-setpoint control) on the first floor. Note that, in both top plots of Figure 8(a) and (b), the red *Tin1_Tstat* curve, which represents the temperatures measured by thermostats during the DQN deployment, is plotted as a reference to signify the difference between the "simulated DQN" and baseline cases. The "simulated DQN" case closely resembles the DQN deployment because the two curves in the top plot of Figure 8(a) are very close. Thus, the comparison between the "simulated DQN" and baseline is highly credible. It is also evident that the fixed-setpoint control gives quite different results from the DQN deployment.

The mid plots in Figure 8(a) and (b) represent the second floor of the research house and show similar patterns to the first floor. The bottom plots show the simulated and measured power use of the HVAC system. We may observe in the bottom plot of Figure 8(a) that the measured power usage during the DQN deployment (in red) matches well with the power usage in the simulated DQN case (in blue). This also

table 4. The daily electricity cost and energy use comparison between DQN/RL and the fixed-setpoint baseline.

	Baseline		DQN/RL				Cost Reduction	
Date	Simulated Cost (US$)	Simulated Energy (Wh)	Simulated Cost (US$)	Measured Cost (US$)	Simulated Energy (Wh)	Measured Energy (Wh)	DQN/RL Simulated Versus Baseline Simulated (%)	DQN/RL Measured Versus Baseline Simulated (%)
25 March	1.19	10,475	1.18	1.47	16,338	19,602	0.8	–23.5
26 March	0.84	5,005	0.64	0.73	5,670	6,824	23.8	13.1
27 March	1.63	13,886	1.29	1.35	18,094	18,892	20.9	17.2
28 March	1.06	9,766	0.97	1.17	12,874	15,650	8.5	–10.4
29 March	2.69	19,316	2.23	2.09	21,025	21,506	17.1	22.3
30 March	2.18	14,412	1.59	1.57	14,311	14,871	27.1	28
31 March	1.11	13,453	1.08	1.11	15,884	16,089	2.7	0
1 April	3.38	29,741	3.02	2.82	31,657	29,552	10.7	16.6
2 April	3.79	31,228	3.7	3.47	33,272	31,867	2.4	8.4
3 April	3.31	24,237	3.31	3.04	26,208	25,751	0	8.2
4 April	2.48	16,872	1.87	1.82	17,582	17,884	24.6	26.6
All	**23.67**	**188,389**	**20.9**	**20.64**	**212,915**	**218,488**	**11.7**	**12.8**

The results signify that the deep RL approach is of considerable potential for online applications in solving complex control and optimization problems like residential demand management.

verifies the accuracy of the fine-tuned building simulation model. In addition, the bottom plot of Figure 8(b) shows the power consumption from the baseline case (in blue) and measured power (in red); the difference between the baseline power consumption and measured power in Figure 8(b) is greater than that in Figure 8(a), which demonstrates the considerable difference between the fixed-setpoint baseline case and "simulated DQN" case.

Table 4 shows a breakdown of the daily energy cost and consumption comparisons for the DQN-based RL control approach and fixed-setpoint baseline case. The daily cost savings from the DQN control (either simulated or deployed) range from 0 to 28% (except for a couple of outlier days with negative cost savings), depending on the day. While there are some fluctuations in the day-to-day comparison of energy use as well as the cost of the simulated and measured DQN cases, the overall energy use and cost over the 11-day deployment have a difference of less than 3%. This indicates that the building simulation is well calibrated. The significant observation from Table 4 is that the DQN cases (either simulated or deployed) consumed more energy than the baseline while still managing to reduce the total cost by 11.7% and 12.8%, respectively, in comparison with the baseline case. This is because the DQN control preheats the home to a higher temperature during low-price periods such that the total cost is decreased, while more energy is consumed, and the comfort level is better.

Note that the minor difference between the simulated DQN case and DQN deployment is likely due to a combination of factors, including small inaccuracies associated with the building model compared to the real-world building response and difference in HVAC response over the decision interval of 5 min. Future performance improvements during deployment could be achieved by decreasing the decision interval of the DQN control to a shorter interval and the inclusion of online learning to fine-tune the decisions of the DQN control over a longer operational period.

Summary

This article explores the application of deep RL approaches to implement energy management in a multizone residential HVAC system to minimize energy costs and maintain user comfort. Both simulation and real-world deployment results demonstrate that the deep RL approach can learn an HVAC control strategy that is more economical, generalized, and adaptive than either the rule-based or simple fixed-setpoint control strategy. The results signify that the deep RL approach is of considerable potential for online applications in solving complex control and optimization problems like residential demand management.

For future directions, an interesting topic would be to further investigate the generalization ability of deep RL approaches, for example, to make the control workable for various scenarios, including both cooling and heating as well as idle scenarios, without additional retraining efforts. Another promising direction would be physics-informed deep RL, which would introduce physical laws to guide the exploration of the approach and further improve learning efficiency. Further, similar approaches based on machine learning may be extended from HVAC control to other building energy controls.

For Further Reading

V. Mnih *et al*., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015, doi: 10.1038/nature14236.

T. P. Lillicrap *et al*., "Continuous control with deep reinforcement learning," 2015, arXiv:1509.02971.

Y. Du *et al*., "Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning," *Appl. Energy*, vol. 281, p. 116,117, Jan. 2021, doi: 10.1016/j.apenergy.2020.116117.

Y. Du *et al*., "Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control," *Electric Power Syst. Res.*, vol. 192, p. 106,959, Mar. 2021, doi: 10.1016/j.epsr.2020.106959.

K. Kurte *et al*., "Evaluating the adaptability of reinforcement learning based HVAC control for residential houses," *Sustainability*, vol. 12, no. 18, p. 7727, Sep. 2020, doi: 10.3390/su12187727.

Biographies

Yan Du is with the University of Tennessee, Knoxville, Tennessee, 37996, USA.

Fangxing Li is with the University of Tennessee, Knoxville, Tennessee, 37996, USA.

Kuldeep Kurte is with Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37830, USA.

Jeffrey Munk is with Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37830, USA.

Helia Zandi is with Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37830, USA.