2022

Comparison Study Of High-performance Rule-based HVAC Control With Deep Reinforcement Learning-based Control In A Multi-zone VAV System

Xing Lu

Yangyang Fu

Shichao Xu

Qi Zhu

Zheng O'Neill

See next page for additional authors

Comparison study of high-performance rule-based HVAC control with deep reinforcement learning-based control in a multi-zone VAV system

Xing Lu[1], Yangyang Fu[1], Shichao Xu[2], Qi Zhu[2], Zheng O'Neill[1]*, Zhiyao Yang[1]

[1]Texas A&M University,
College Station, Texas, USA

[2]Northwestern University,
Evanston, USA
* Corresponding Author

ABSTRACT

The design, commissioning, and retrofit of heating, ventilation, and air-conditioning (HVAC) control systems are crucially important for energy efficiency but often neglected. Generally, designers and control contractors adopt ad-hoc control sequences based on diffused and fragmented information and therefore the majority of the existing control sequences are diverse and sub-optimal. ASHRAE Guideline 36 (GDL36), High-performance Sequences of Operation for HVAC Systems, is thus developed to provide standardized and high-performance rule-based HVAC control sequences with the main focus on maximizing energy efficiency. However, these high-performance rules-based control sequences are still under-development, and only a few studies verify their overall effectiveness. In addition, the performance evaluations in most existing studies only focus on the energy-saving potentials compared with the conventional rule-based control strategies. In this study, the high-performance rule-based control sequences from GDL36 was compared to the state-of-the-art deep reinforcement (DRL) control in terms of the energy efficiency in a multi-zone VAV system. The system-level supervisory controls (i.e., supply air temperature and supply differential pressure setpoint) in ASHRAE GDL36 were replaced by the counterpart in the DRL control, of which action space is a bi-dimensional continuous space. A five-zone medium office building model in Modelica was utilized as a virtual testbed. Particularly, the plant side power consumption uses a regression model to reflect the real condition of the plant loop operation, Proximal policy optimization (PPO) was selected as the DRL algorithm due to its stable performance for the continuous space and easiness of the hyper-parameter tuning. The DRL algorithm was implemented using the Tianshou library in Python. A containerized OpenAI gym environment was leveraged to enable the connection between the Modelica building model and the DRL algorithm. Typical load conditions in Chicago, 5A (high and mild load weeklong simulation) were considered. The simulation results show that control sequences from GDL36 perform comparable performance in terms of energy efficiency and thermal comfort as the DRL controls.

1. INTRODUCTION

Innovations in building controls at the supervisory level have great potential to achieve the whole-building level energy savings on the order of 30% and higher (Pritoni et al., 2020). Despite the significant role of the HVAC control systems in energy efficiency, its design, commissioning, and retrofit have long been an intricate and complicated issue, considering that only diffuse and fragmented information on system operation is available for decision making in most of the scenarios. Due to this limitation, designers and control contractors can only rely on ad-hoc control sequences for system operation in practice, which is one of the major reasons why buildings are operated sub-optimally. To provide standardized and high-performance HVAC control sequences, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) has developed the Guideline 36 (GDL36) High Performance Sequences of Operation (SOO) for HVAC Systems to maximize energy efficiency.

Control sequences collected in GDL36 belong to prescriptive and feedback-based reactive control. In this type of control, different setpoints or schedules are determined at the supervisory level based on the heuristic rules and then the Proportional-Integral-Derivative (PID) local controls are used to track the setpoints. The energy efficiency performance of GDL36 has been demonstrated by limited lab/field tests and simulation-based studies (Paliaga et al., 2020; Pritoni et al., 2020; Wetter et al., 2018), as tabulated in Table 1. Rodriguez (Rodriguez, 2019) conducted an

experiment in a commercial building air handling unit (AHU) and showed that more outside air was used and building cooling energy use was reduced after GDL36 was implemented. The control retrofit using the GDL36 airside SOO in 555 County Center (Talyor Engineering, 2020), a five-story office building in California, indicated that 15% of whole building electricity use and 56% of natural gas use were saved in the first year of the testing. A medium building in Vallejo, California, was retrofitted by implementing GDL36 SOO (Kiriu & Stein, 2021). The control sequence retrofits reduced annual energy bills by over $200,000 and heating energy use by more than 55%. These two field demonstrations showed a simple payback year of 6.7 and 8.9 years, respectively.

Wetter et al. (Wetter et al., 2018) implemented the airside GDL36 SOO in a single-floor medium office building model and reported a 30% annual site energy usage saving for the HVAC system with acceptable thermal comfort compared to the old SOO published in 2006 by ASHRAE (ASHRAE, 2005). Zhang et al. (Zhang et al., 2020) implemented and verified both airside and waterside control sequences in a Modelica-based simulation environment for a single-zone VAV system. Their simulation results showed that the GDL36 yielded 17.3 % of annual HVAC energy saving compared to the conventional baseline control strategy. In a follow-up study (Zhang et al., 2022), they estimated the energy saving of the control retrofit for multi-zone variable air volume systems using Spawn of EnergyPlus. The results showed the GDL36 SOO could provide a wide range of HVAC energy savings with an average savings of 31% in different climates, internal loads, and HVAC system operation periods. Overall, the energy-saving potential from retrofitting existing controls to the GDL36 SOO has yet to be shown for a wide range of cases (Zhang et al., 2022).

Table 1: Summary of existing evaluation studies of GDL36 SOO on energy efficiency

Study	Approach	Test Conditions	Baseline	Results
(Wetter et al., 2018)	Simulation	Modelica-based single-floor five-zone medium office VAV system (airside)	Typical rule-based	- 30% annual site energy use of the HVAC system with acceptable thermal comfort.
(Zhang et al., 2020)	Simulation	Modelica-based single-zone system (plant & airside)	Typical rule-based	- 17.3% of annual HVAC energy saving with acceptable thermal comfort.
(Zhang et al., 2022)	Simulation	21-zone VAV system (airside)	Typical rule-based	- an average of 31% HVAC energy saving
(Rodriguez, 2019)	Field test	AHU in a commercial building	Not mentioned	- More outside air was used and building cooling energy use was reduced.
(Talyor Engineering, 2020)	Field test	555 County Center, five-story office building, CA (Airside)	Not mentioned	- In the first year, 15% of whole building electricity use and 56% of natural gas use. - Estimated payback period was 6.7 years.
(Kiriu & Stein, 2021)	Field test	A medium hospital building in Vallejo, CA (Airside & Waterside)	Not mentioned	- Reduced annual energy bills by over $200,000 and heating energy use by more than 55%. - Estimated payback period was 8.9 years.

The reviewed literature indicates the design and commissioning of the HVAC system control sequences could have a substantial impact on building energy efficiency. The airside control retrofits following high-performance SOO could achieve energy savings up to 30% and more compared to the conventional rule-based control. However, achievable savings from the high-performance SOO in other HVAC system parts are still unknown (e.g., airside controls together with waterside controls). In addition, the energy saving potential of high-performance SOO compared to the state-of-the-art intelligent controls as the benchmark is also unclear. Therefore, there is a practical need to benchmark the SOO in GDL36 with other intelligent controls such as deep reinforcement learning (DRL)-based control.

Reinforcement learning (RL) is a category of machine learning algorithms that aims to learn an optimal control policy from the direct interaction between the agent and the environment. The agent will perform empirical learning and decide on the action to drive the environment towards a favorable trajectory according to a predefined reward function. Different researchers have investigated the application of RL controllers in building systems, including single air-conditioning units (Costanzo et al., 2016; Leurs et al., 2016), VAV systems (Azuatalam et al., 2020; Jia et al., 2019; Wang et al., 2017; Yu et al., 2020; Yuan et al., 2021), radiant heating systems (Chen et al., 2019; Zhang et al., 2019), building envelope(Chen et al., 2018), and whole HVAC systems (Ahn & Park, 2020; Liu & Henze, 2006, 2007; Yang

et al., 2015). Despite the reported benefits after the successful controller tuning, the model-free RL controllers are subject to the issues such as the long training period and stability issues.

There exist several studies on developing RL strategies for commercial building VAV systems. (Yuan et al., 2021) applied Q-learning-based RL in both single-zone and multi-zone VAV systems to optimize the supply airflow rate. Despite achieved energy saving over the baseline control, details on the baseline control are unclear. In addition, it is not clear whether the control design of AHU-level control loops is considered or not. Wei et al. (Wei et al., 2017) used deep Q network algorithms for optimal airflow control of multi-zone VAV systems. The simulation experiments demonstrated the DRL-based algorithm was more effective in an energy cost reduction compared with the rule-based controllers. Hanumaiah et al. (Hanumaiah & Genc, 2021) proposed a distributed multi-agent DRL framework for the optimal control of multi-zone systems. The impact of the reward ratio and the weather on different model-free RL performances was discussed. Ding et al. (Ding et al., 2020) developed a model-based RL with a model predictive path integral control method to the multi-zone VAV. The results showed that the proposed controller could achieve 10.65% more energy savings compared to the rule-based benchmark while maintaining similar thermal comfort. Furthermore, the training time was reduced significantly compared to the model-free RL benchmark.

Although aforementioned literature review demonstrates the energy saving potential of DRL, there exist few studies that compared the performance of DRL-based controllers with high-performance rule-based control sequences of operation, i.e., ASHRAE Guideline 36. To be specific, the benchmark control strategies in most existing evaluation studies are the PID and on-off controllers. For example, the benchmark rule-based controller for the zone air temperature (ZAT) controller is the on-off control (Wei et al., 2017). Essentially speaking, the GLD36 SOO is much simpler compared to the DRL controllers. The heuristic rules in GLD36 could improve the energy efficiency; however, the improvement may be constrained by the incapability of predictive and adaptive learning. OBC and RL controllers have their own obvious challenges and limitations, which prevent their wide applications in the field at the current stage. In this context, this study presents an energy performance comparison of GDL36 SOO and DRL-based control within a medium office building virtual testbed in Modelica. Section 3 describes the case study description and the simulation testbed. Section 4 details the formulation and implementation of the DRL-based controller (DRLC). Section 5 discusses the energy efficiency comparison results of the DRLC and GDL36.

2. DESCRIPTION OF CASE STUDY

This case study aims to compare energy efficiency and thermal comfort performance of DRLCs with ASHRAE Guideline 36 for a multi-zone VAV cooling system, which is a typical HVAC system configuration in commercial buildings. The energy efficiency is reflected by the cooling energy use for the whole HVAC system and total ZAT violation during the system operation hours (i.e., 7 am – 7 pm) is calculated as a thermal comfort metrics.

$$E_{Coo,tot} = \sum_{t_0}^{t_N} E_{HVAC}(t_i) = \sum_{t_0}^{t_N} E_{Fan}(t_i) + \sum_{t_0}^{t_N} E_{Plant}(t_i) \quad (1)$$

$$dt_{tot}(t_0, t_N) = \sum_{z \in Z} \sum_{t_0}^{t_N} |s_z(t_i)| \quad (2)$$

where N is the sampling number for each operation time step point t. E_{Fan} and E_{Plant} are the energy use for the AHU fan and the plant system. z is the zone index for the set of zones, and s_z is the deviation from the lower and upper setpoint temperatures. The zone air cooling temperature setpoint is 24 ℃, and the allowable deviation in this study is ± 0.5 ℃ from the setpoint.

Since the published version of GDL36-2018 only contains the control sequences on the airside systems, this study focuses on the comparison of rule-based airside high-performance sequences with the intelligent controllers. Another assumption is that the comparison is conducted at the supervisory level, which is the overall control of the local subsystems (Wang & Ma, 2008). For the airside control SOO of multi-zone VAV systems in GDL36, there exist several critical supervisory level controls, e.g., AHU supply air temperature (SAT) reset and static differential pressure (DP) setpoint reset, and economizer damper controls. In this comparison study, the first two controls are replaced by the counterpart in two intelligent controllers. For the AHU SAT reset, the SAT is reset based on the outdoor air temperature (OAT) and the setpoint request from the zone terminal units to find a balance between the fan energy and cooling energy. To be specific, the setpoint shall be reset from minimum cooling SAT when the outdoor air temperature is maximum OAT and above, proportionally up to maximum SAT when the outdoor air temperature is

minimum OAT and below. The maximum OAT is reset using T&R logic based on the zone-level reset requests between minimum and maximum cooling SAT. The static DP reset is enabled by the Trim & Respond control. Under that control logic, the system will tend towards minimum static pressure but respond to the increasing demand from the zone terminal units.

Compared to the high-performance SOO in GDL36, the SAT and the static DP are determined by control policy in DRLC. To ensure an apples-to-apples comparison, the local controls (e.g., zone-level PID controls) remain the same for the three controllers. Table 2 lists the differences of SOO between two controller types.

Table 2: Difference between two controller types

Supervisory control loop name	GDL36	DRLC
SAT setpoint reset	Rule-based, i.e., determined by the OAT and zone requests	Determined by the optimal control policy after the training
Static DP reset	Rule-based, i.e., determined from zone requests	

The simulation experiment was for Chicago, IL, USA (ASHRAE climate zone 5A) in two typical weeks of different cooling loads, i.e., a cooling week (07/24-07/31) and a shoulder week (06/09-06/16). The cooling week has a high average outdoor air temperature, and the shoulder week has a mild cooling load which enables a long operation period of the airside economizer. The simulated medium office building was the single-floor five-zone VAV system as described in the reference (Lu, Fu, et al., 2021) under both airside and waterside control sequences of GDL36. The original model was developed from the GDL36 model in Modelica Buildings Library 7.0.0 (Wetter et al., 2014). To ease the computational cost, the detailed waterside model was replaced by the data-driven regression model for both the cooling season and the shoulder season, respectively, as shown in Eq. (3) and (4).

$$P_{pla} = 11188 + 0.18 \cdot \mathrm{Q_{coo}} + 24.24 \cdot \mathrm{T_{db}} - 44.44 \cdot \mathrm{T_{wb}} - 0.05, \quad (3)$$

$$P_{pla} = 726656 + 4.12 \cdot Q_{coo} - 2816.1 \cdot T_{db} - 2638.7 \cdot T_{wb} + 26.2 \cdot H_{gh} - 0.0037 Q_{coo} \cdot T_{db} - 0.0097 Q_{coo} \cdot T_{wb} + 2.01 \cdot 10^{-5} \cdot Q_{coo} \cdot H_{gh} + 10.2 \cdot T_{db}\, T_{wb} + 0.16 \cdot T_{db} \cdot H_{gh} - 0.25 \cdot T_{wb} \cdot H_{gh}, \quad (4)$$

where Q_{coo} is the cooling load at the cooling coil, T_{db} is the dry bulb outdoor air temperature, $\mathrm{T_{wb}}$ is the wet bulb outdoor air temperature and H_{gh} is the global horizontal solar radiation. Table 3 shows the statistical metrics of the regression model. The coefficient of determination (R2), root mean square error (CV-RMSE), and normalized mean bias error (NMBE) both indicate the high accuracy of the regression models. As mentioned earlier, the system is sized under the ASHRAE climate zone 5A Chicago, IL (Lu, Adetola, et al., 2021; The U.S. Department of Energy (DOE), 2020).

Table 3: Regression model accuracy statistical results

Regression model	Model type	R2	CV-RMSE	NMBE
Cooling season model	Linear	0.99	2.1%	1.3%
Shoulder season model	Interactions linear	0.99	3.2%	2.3%

3. CONTROLLER FORMULATION AND IMPLEMENTATION

The DRLC was formulated to minimize the HVAC total energy consumption while mitigating the ZAT violation by adjusting the AHU SAT setpoint and the AHU static pressure setpoint, were bounded within [12 ℃,18 ℃] and [25 Pa,410 Pa]. Figure 1 depicts the DRLC formulation. The reward R for the DRL is shown in Eq. (5).

$$R_i = E_{HVAC,i} + \alpha \cdot dT_{vio,i}, \quad (5)$$

where $E_{HVAC,i}$ are the HVAC energy consumption and $dT_{vio,\mathrm{i}}$ are and ZAT violation at the i$^{\mathrm{th}}$ control interval (i.e., 15 minute each). α is the penalty coefficient that balances the energy consumption and thermal comfort. Generally, a small α corresponds to a smaller HVAC energy consumption but a larger temperature violation and vice versa. An appropriate α needs to be tuned to keep a similar level of ZAT violation with the GDL36 controller.

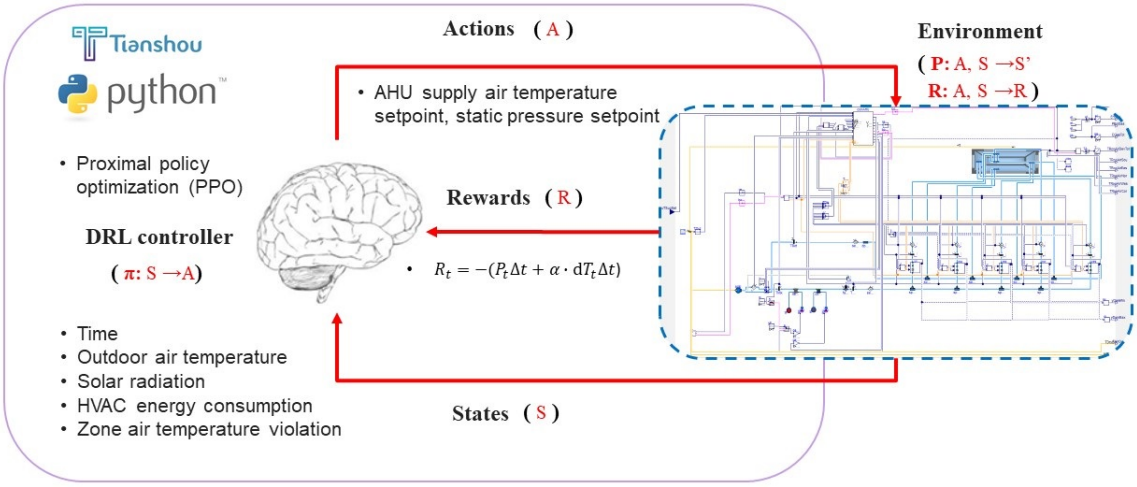

Figure 1: Schematics of DRLC formulation and implementation

The control actions are bidimensional and the action space is a continuous space to avoid the curse of the dimension in the discretized action space. The states are determined based on HVAC engineering knowledge. We consider three different combinations of states, as shown in Table 4. The commonly used states are time, outdoor air temperature, solar radiation, HVAC energy consumption, and ZAT violation. The simulation environment is the virtual medium office in Modelica which provides the reward value (R) and next observations (S') during the interaction with the DRLC.

Table 4: Summary of different state design

State Space	Number of States	State Variables
S1	7	time, outdoor air temperature, solar radiation, and HVAC energy consumption, ZAT violation, fan speed, maximum/minimum zone terminal damper position
S2	6	time, outdoor air temperature, solar radiation, HVAC energy consumption, ZAT violation, fan speed
S3	5	time, outdoor air temperature, solar radiation, HVAC energy consumption, and ZAT violation

For the implementation, a flexible containerized framework (Fu et al., 2021) was leveraged where the building model was interfaced with a state-of-the-art DRL library Tianshou (Weng et al., 2021) through the functional mockup unit (FMU). Tianshou is a highly modularized DRL library in Python based on pure PyTorch (Paszke et al., 2019) and has supported more than 20 classic algorithms. Tianshou's performances are reported to be comparable or better than the best reported results for most algorithms in the open literature (Weng et al., 2021). In this study, the Proximal Policy Optimization (PPO) (Schulman et al., 2017) is selected as the DRL algorithm because it suits the continuous bi-dimensional action space in our case. Still, several critical hyperparameters in PPO need to be tuned (Raval, 2021). For example, Step Per Collect (also called Time Horizon), i.e., how many steps to collect before adding it to the experience buffer; Batch size (also called Minibatch), i.e., how many experiences are used for each gradient descent update; Entropy coefficient i.e., a regularizer that helps the exploration; and Updated time, i.e., how many times the data collected are reused, etc. are hyperparameters that could have significant impacts on the DRLC performance.

To fine-tune the proposed DRLC, various other factors were considered, including the penalty coefficient, the state design, the neural network layer number, as suggested in this reference (Andrychowicz et al., 2020). The penalty coefficient α was first swept to determine the appropriate value that keeps a similar level of ZAT violation with the GDL36 controller. Then different common values are grid-searched in other aspects to find the best hyperparameters. Table 5 lists the sweeping parameters for tuning the DRLC. The DRL policy is trained and tested for 800 epochs in each scenario (i.e., different combinations of parameters). One epoch length is one week. The computation time for one epoch training takes around 10 minutes on a Windows 10 machine with Intel® Core™ i5-9500 @3.00 GHz CPU and 16 GB RAM. That being said, the training time for 800 epochs in single scenario would take about 5.5 days. Due

to the large computation cost for the hyperparameter tuning of the DRLC, different scenarios are assigned to different cores in the high-performance computing clusters in Texas A&M University.

Table 5: Sweeping parameters for DRLC tuning considering various factors

Aspects	State design	Entropy coefficient	Step per collect	Batch size	Repeat per collect	State normalization	Advantage normalization, Value Clip	Number of Neural Network Layers
Value	S1, S2, S3	0, 0.01	384, 512	64, 128	5, 10	True, False	True, False	3, 4, 5
Number	3	2	2	2	2	2	2	3

4. RESULTS AND DISCUSSION

In this section, the experimental results for DRLC after the hyperparameter tuning are reported analyzed. Recall that for each scenario (i.e., the combination of the hyperparameters), the DRL policy has been trained and tested for 800 epochs (each epoch denotes one week).

4.1. Cooling Season Results

Figure 2 illustrates the rewards under different scenarios using the parallel coordinate plot. Each line denotes one epoch under different combinations of hyperparameters. The line color represents the reward value. The redder represents a higher reward value, while the bluer represents the lower reward value. The reward value of the GDL36 SOO is annotated in the color bar on the right side. It can be seen that the DRLC under the cooling season needs to be trained for at least 300 epochs. The larger number of Step Per Collect and Repeat Per Collect is generally beneficial to the final rewards but increases the DRLC training time.

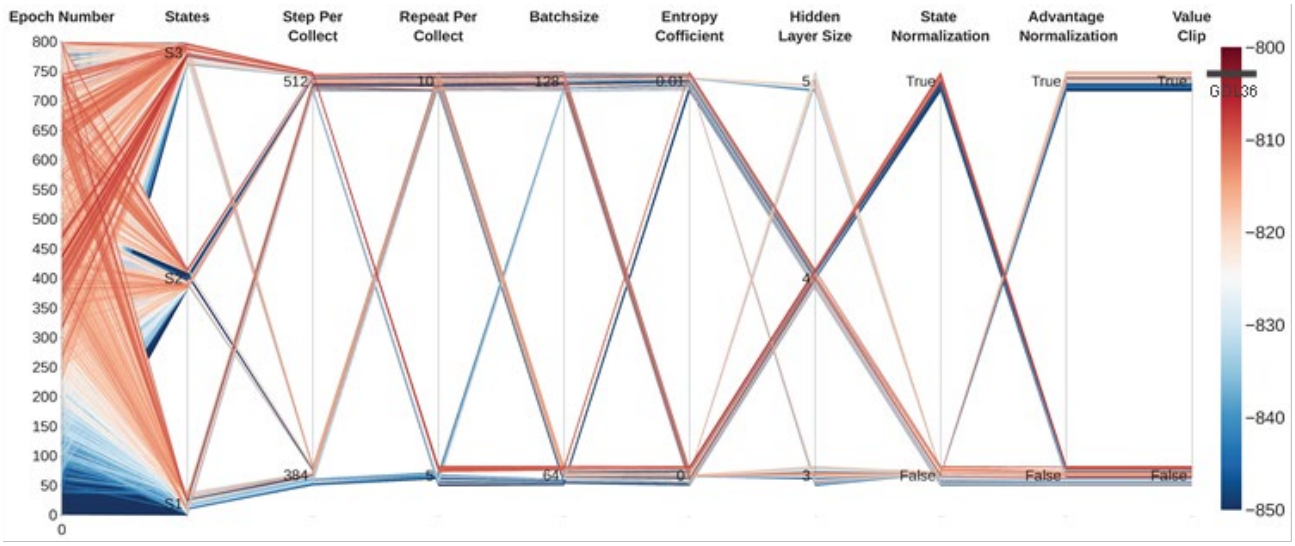

Figure 2 Parallel coordinate plot of rewards under different scenarios in cooling season week

Figure 3 shows the reward evolution throughout the epochs for the best scenario in the cooling season week. The hyperparameter setting for the best scenario is also provided. The blue line represents the entire reward for the GDL36 SOO. It can be seen that after 800 epochs of training, the DRLC performance in the cooling season week could nearly chase up with the GDL36 SOO in terms of the reward. The HVAC energy consumption for the best scenario of DRLC increases 2.2% compared to the GDL36 SOO while decreasing 0.83 K·h temperature violation in the cooling season week. This indicates the energy efficiency performance of GDL36 SOO is comparable to DRLC in the cooling week for this specific study. In addition, the DRLC is still subject to the curse of the high training time to achieve comparable performance.

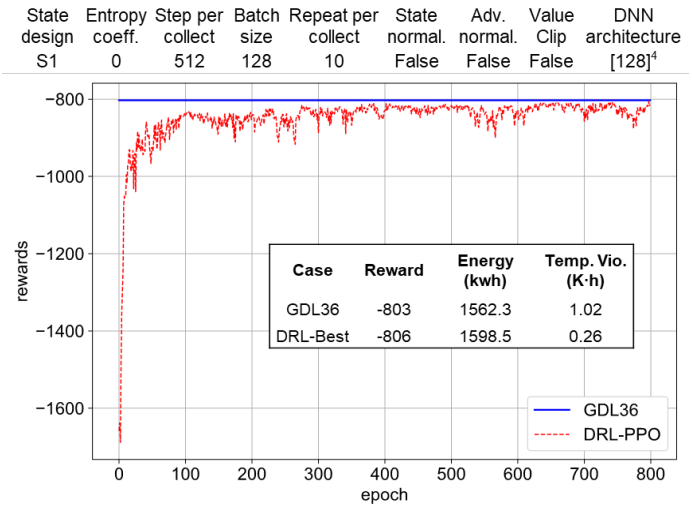

Figure 3: Reward per epoch for the best scenario in the cooling season week

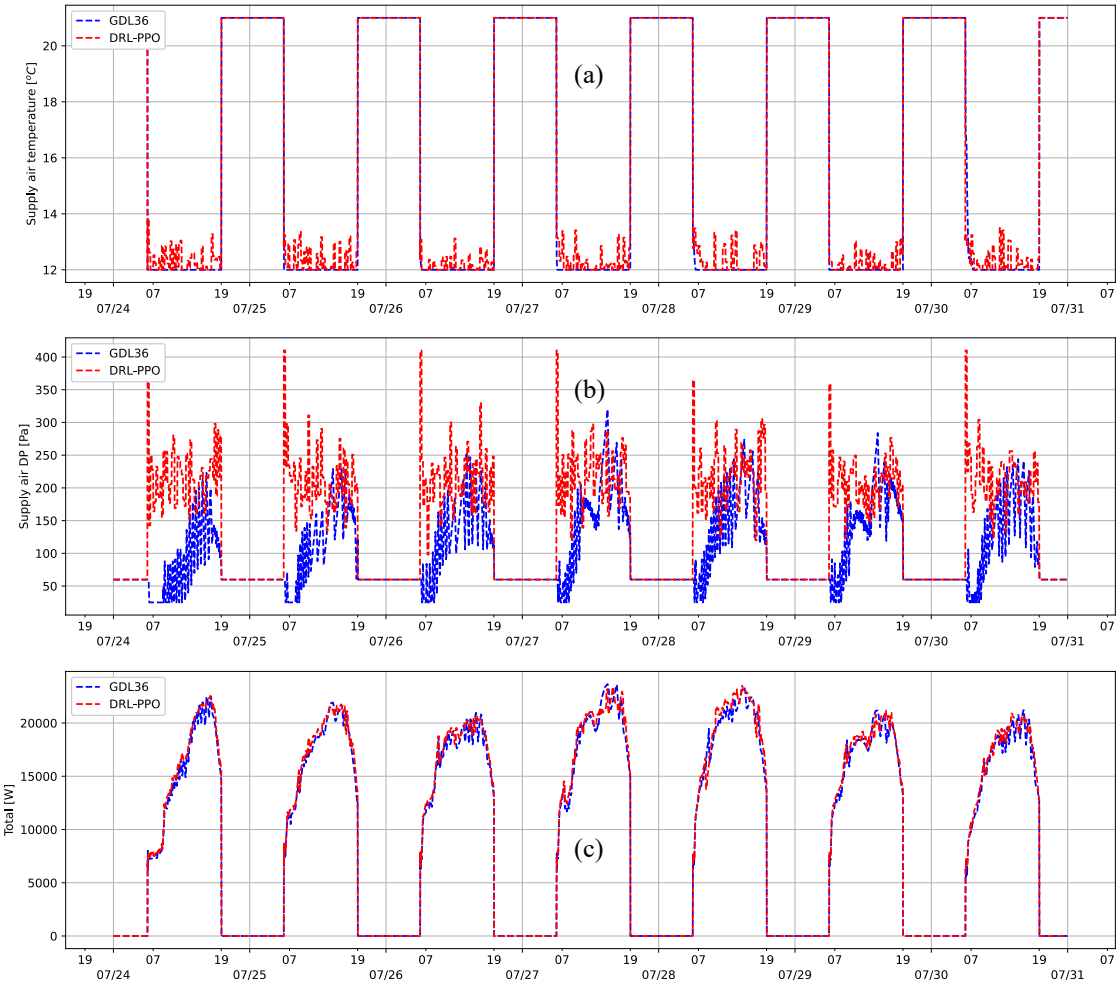

Figure 4: Time series of (a) SAT setpoint (b) static DP setpoint (c) Total HVAC power consumption during cooling season week for DRLC

Figure 4 depicts the detailed energy results for the DRLC in the cooling season week. Compared to the SAT setpoints in GDL36 staying at the lowest value (i.e., 12 ℃), the SAT setpoints have a frequent variation between 12-14 ℃ in

the case of DRLC. For the static DP setpoints, DRLC generally has a higher value throughout the operation hours. Figure 4(c) shows that the line of HVAC power consumption for DRLC overlap for most of the days with GDL36 SOO, while DRLC expending slightly less power consumption at some periods.

4.2. Shoulder Season Results

Similarly, the rewards under different scenarios are illustrated in Figure 5. The reward value of the GDL36 SOO is annotated in the color bar on the right side. Roughly after 150 epochs' training, DRLC under the shoulder season week could achieve an equivalent level performance with GDL36 SOO. Similarly, Figure 6 depicts the details regarding the best scenario in the shoulder season week. The best reward for the shoulder season week is -596, which is 4.64% higher than the baseline GDL36.

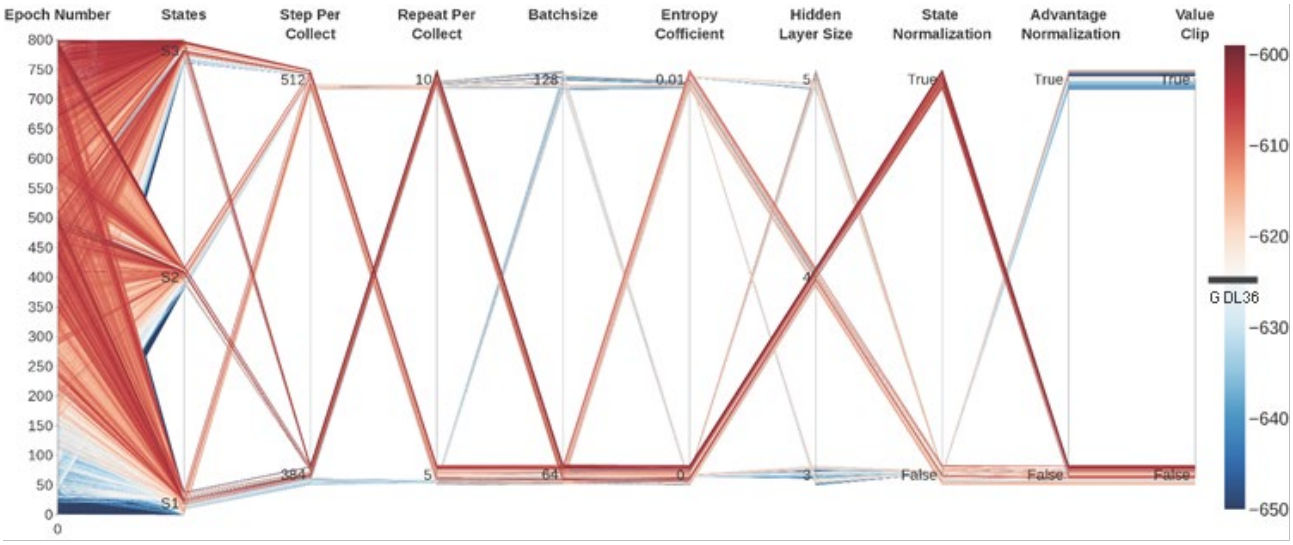

Figure 5: Parallel coordinate plot of rewards under different scenarios in shoulder season week

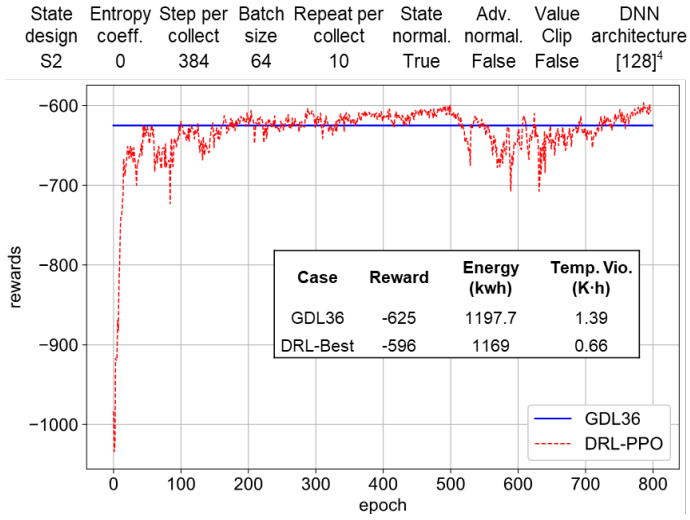

Figure 6: Reward per epoch for the best scenario in the shoulder season week

4.3. Summary

Based on the above results, following findings could be summarized. For the HVAC energy consumption in the cooling season week, the DRL consumes 2.3% more compared to GDL36 SOO. In the shoulder season week, the DRL save 2.4%. It is noted that this amount of saving is less than the penalties of uncertainties from typical sensor measurements in HVAC systems. For the thermal comfort metric, two controllers can all maintain the ZAT within the

predefined comfort bounds with minor temperature violation. The DRLC have slightly less temperature violation than the GDL36 SOO.

5. CONCLUSIONS

ASHRAE GDL36 has demonstrated energy efficiency benefits over the conventional rule-based controls. In this study, the energy and thermal comfort performance of GDL36 are compared with DRLC. This study is conducted with a five-zone VAV cooling system virtual testbed in Chicago, IL. The baseline control system is implemented with the high-performance airside and waterside GDL36 SOO. The DRLC replace the airside supervisory level control loops. In other words, the optimal SAT and static DP are determined by trained control policy in DRLC. The DRLC was formulated to minimize the HVAC energy consumption and zone air temperature violations. The DRLCs with different hyperparameters in the PPO algorithm were studied and fine-tuned. The results showed that the GDL36 SOO has a comparable energy performance (within a 3% deviation) with DRLC in both high and mild cooling loads. For the thermal comfort metric, the GDL36 has slightly more ZAT violation in both typical weeks compared to DRLC.

For this case study, the GDL36 has demonstrated its comparable performance in terms of energy efficiency and thermal comfort with the two intelligent controllers. The GDL36 is good enough considering the complexity, long training time, and tuning efforts of the DRLC. However, there are several limitations of this study to be noted. First, the DRLC are formulated ideally only for theoretical comparison studies. They are not deployable for real applications. The control policies are trained and tested for the same week, which is also not realistic in practice. Second, the performance might be further improved by considering more complex aspects. For DRLCs, only the PPO algorithm is explored and other DRL algorithms are not studied. Third, the simulation-based study is only experimented in a five-zone medium office building, one single climate zone, and only cooling season. The effect of the climate, building type, internal loads, and operation time on the final results are not investigated. Therefore, the future work includes the expansion of the evaluation studies to other building types with different HVAC systems and climate zones; and the comparison studies for more complicated intelligent controllers.

REFERENCES

Ahn, K. U., & Park, C. S. (2020). Application of deep Q-networks for model-free optimal control balancing between different HVAC systems. *Science Technology for the Built Environment*, *26*(1), 61-74.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., . . . Michalski, M. (2020). What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*.

ASHRAE. (2005). *Sequences of Operation for Common HVAC Systems*. American Society of Heating Refrigerating and Air-Conditioning Engineers.

Azuatalam, D., Lee, W.-L., de Nijs, F., & Liebman, A. (2020). Reinforcement learning for whole-building HVAC control and demand response. *Energy and AI*, *2*, 100020.

Chen, B., Cai, Z., & Bergés, M. (2019). Gnu-rl: A precocial reinforcement learning solution for building hvac control using a differentiable mpc policy. Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation,

Chen, Y., Norford, L. K., Samuelson, H. W., & Malkawi, A. (2018). Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings*, *169*, 195-205.

Costanzo, G. T., Iacovella, S., Ruelens, F., Leurs, T., & Claessens, B. (2016). Experimental analysis of data-driven control for a building heating system. *Sustainable Energy, Grids Networks*, *6*, 81-90.

Ding, X., Du, W., & Cerpa, A. E. (2020). Mb2c: Model-based deep reinforcement learning for multi-zone building control. Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation,

Fu, Y., Xu, S., Zhu, Q., & O'Neill, Z. (2021). Containerized framework for building control performance comparisons: model predictive control vs deep reinforcement learning control. Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation,

Hanumaiah, V., & Genc, S. (2021). Distributed Multi-Agent Deep Reinforcement Learning Framework for Whole-building HVAC Control. *arXiv preprint arXiv:.13450*.

Jia, R., Jin, M., Sun, K., Hong, T., & Spanos, C. (2019). Advanced building control via deep reinforcement learning. *Energy Procedia*, *158*, 6158-6163.

Kiriu, R., & Stein, J. (2021). Medical Office Building Thrives With Advanced Control Sequences. *ASHRAE Journal*, *63*, 62-67.

Leurs, T., Claessens, B. J., Ruelens, F., Weckx, S., & Deconinck, G. (2016). Beyond theory: Experimental results of a self-learning air conditioning unit. 2016 IEEE International Energy Conference (ENERGYCON),

Liu, S., & Henze, G. P. (2006). Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 2: Results and analysis. *Energy and Buildings*, *38*(2), 148-161.

Liu, S., & Henze, G. P. (2007). Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory.

Lu, X., Adetola, V., & O'Neill, Z. (2021). What are the Impacts on the HVAC System when it Provides Frequency Regulation?–A Comprehensive Case Study with a Medium Office Building. *Energy and Buildings*, 110995.

Lu, X., Fu, Y., O'Neill, Z., & Wen, J. (2021). A holistic fault impact analysis of the high-performance sequences of operation for HVAC systems: Modelica-based case study in a medium-office building. *Energy and Buildings*, *252*, 111448.

Paliaga, G., Singla, R., Snaith, C., Lipp, S., Mangalekar, D., Cheng, H., & Pritoni, M. (2020). Re-Envisioning RCx: Achieving Max Potential HVAC Controls Retrofits through Modernized BAS Hardware and Software.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Pritoni, M., Prakash, A., Blum, D., Zhang, K., Tang, R., Granderson, J., . . . Paliaga, G. (2020). Advanced control sequences and FDD technology. Just shiny objects, or ready for scale?

Raval, S. (2021). *Best Practices when training with PPO*. https://github.com/llSourcell/Unity_ML_Agents/blob/master/docs/best-practices-ppo.md

Rodriguez, A. (2019). *Utilizing ASHRAE G36 for Free Outside Air Cooling (Economizer)* The George Washington University].

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:.06347*.

Talyor Engineering. (2020). *Case Study: Advanced HVAC Controls*. https://taylorengineers.com/wp-content /uploads/2020/04/2018-09-18-Advanced-HVAC-Controls-Case-Study-555-County-Center.pdf

The U.S. DepartmentofEnergy(DOE). (2020). *DOE Commercial Prototype Building Models* https://www. energycodes.gov/development/commercial/prototype_models

Wang, S., & Ma, Z. (2008). Supervisory and optimal control of building HVAC systems: A review. *HVAC & R Research*, *14*(1), 3-32.

Wang, Y., Velswamy, K., & Huang, B. (2017). A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. *Processes*, *5*(3), 46.

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep reinforcement learning for building HVAC control. Proceedings of the 54th annual design automation conference 2017,

Weng, J., Chen, H., Yan, D., You, K., Duburcq, A., Zhang, M., . . . Zhu, J. (2021). Tianshou: a Highly Modularized Deep Reinforcement Learning Library. *arXiv preprint arXiv:.14171*.

Wetter, M., Hu, J., Grahovac, M., Eubanks, B., & Haves, P. (2018). OpenBuildingControl: Modeling feedback control as a step towards formal design, specification, deployment and verification of building control sequences. Building Performance Modeling Conference and SimBuild,

Wetter, M., Zuo, W., Nouidui, T. S., & Pang, X. (2014). Modelica buildings library. *Journal of Building Performance Simulation*, *7*(4), 253-270.

Yang, L., Nagy, Z., Goffin, P., & Schlueter, A. (2015). Reinforcement learning for optimal control of low exergy buildings. *Applied Energy*, *156*, 577-586.

Yu, L., Sun, Y., Xu, Z., Shen, C., Yue, D., Jiang, T., & Guan, X. (2020). Multi-agent deep reinforcement learning for HVAC control in commercial buildings. *IEEE Transactions on Smart Grid*, *12*(1), 407-419.

Yuan, X., Pan, Y., Yang, J., Wang, W., & Huang, Z. (2021). Study on the application of reinforcement learning in the operation optimization of HVAC system. Building Simulation,

Zhang, K., Blum, D., Cheng, H., Paliaga, G., Wetter, M., & Granderson, J. (2022). Estimating ASHRAE Guideline 36 energy savings for multi-zone variable air volume systems using Spawn of EnergyPlus. *Journal of Building Performance Simulation*, *15*(2), 215-236.

Zhang, K., Blum, D. H., Grahovac, M., Hu, J., Granderson, J., & Wetter, M. (2020). *Development and Verification of Control Sequences for Single-Zone Variable Air Volume System Based on ASHRAE Guideline 36*.

Zhang, Z., Chong, A., Pan, Y., Zhang, C., & Lam, K. P. (2019). Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy and Buildings*, *199*, 472-490.