Investigations on the Influence of Model Accuracy in Deep Reinforcement Learning Control for HVAC Applications

Mingyue Guo
Student Member ASHRAE

Yangyang Fu, PhD
Full Member ASHRAE

Mingzhe Liu, PhD
Full Member ASHRAE

Zheng O'Neill, PhD, PE
Fellow ASHRAE

ABSTRACT

Recent research has highlighted the effectiveness of advanced building controls in reducing the energy consumption of heating, ventilation, and air-conditioning (HVAC) systems. Among advanced building control strategies, deep reinforcement learning control (DRL) shows the potential to achieve energy savings for HVAC systems and has emerged as a promising strategy. However, training DRL requires an interactive environment for the agent, which is challenging to achieve with real buildings due to time and response speed constraints. To address this challenge, a simulation environment serving as a training environment is needed, even though the DRL algorithm does not necessarily need a model. The error between the model and the real building is inevitable in this process, which may influence the efficiency of the DRL controller. To investigate the impact of model error, a virtual testbed was established. A high- fidelity Modelica-based model is developed serving as the virtual building. Three reduced-order models (ROMs) (i.e., 3R2C, Light Gradient Boosting Machine (LightGBM) and artificial neural network (ANN) models) were trained with the historical data generated from the virtual building and were embedded in the training environments of DRL. The sensitivity of ROMs and the Modelica model to random and periodical actions were tested and compared. Deploying the policy trained based on a ROM-based environment, which stands for a surrogate model in reality, into the Modelica-based virtual building testing environment, which stands for real-building, is a practical approach to implementing the DRL control. The performance of the practical DRL controller is compared with rule-based control (RBC) and an ideal DRL controller which was trained and deployed both in the virtual building environment. In the final episode with best rewards of the case study, the 3R2C, LightGBM, and ANN-based DRL outperform the RBC by 7.4%, 14.4%, and 11.4%, respectively in terms of the reward, comprising the weighted sum of energy cost, temperature violations, and the slew rate of the control signal, but falls short of the ideal Modelica-based DRL controller which outperforms RBC by 29.5%. The DRL controllers based on data-driven models are highly unstable with higher maximum rewards but much lower average rewards which might be caused by the significant prediction defect in certain action regions of the data-driven model.

INTRODUCTION

Background and Literature Review

In 2022, the commercial and residual buildings consumed 35.1% and 38.9% of total U.S. retail sale electricity. 32.1% of power in residential buildings and 12.9% of power in commercial buildings is used for space conditioning(EIA, 2023), which indicates Heating, Ventilation and Air-conditioning (HVAC) system is a large end-use for electricity

Mingyue Guo is a PhD student, **Yangyang Fu, PhD**, is a Research Engineer, **Mingzhe Liu, PhD**, is a Postdoctoral Researcher, **Zheng O'Neill, PhD, PE**, is an Associate Professor in the J. Mike Walker 66' Department of Mechanical Engineering, Texas A&M University, College Station, TX.

consumption. Numerous studies have shown that advanced HVAC control can significantly reduce energy use and mitigate greenhouse gas emissions (Fu et al., 2019).

However, a large portion of buildings today are still operated under simple rule-based control (RBC) strategies which have limited energy-saving potential, especially under a flexible energy price, e.g., time-of-use price. Advanced model-based control strategies e.g., Model Predictive Control (MPC) are often inefficient in practice, due to the compromise between the complexity of the building thermal dynamic modeling and the difficulty to get the feasible optimal solution (Zong et al., 2017).

Compared with model-based control, deep reinforcement learning (DRL) control has advantages due to its model- free nature, i.e., DRL agents train their policies via direct interaction with an environment to learn the optimal control strategies. Although many efforts have been made to incorporate RL in HVAC control (Kazmi et al., 2018; Qiu et al., 2020; Wei et al., 2017) , the application of DRL control in practice of HVAC industry is still limited. The most notorious reason is that it may take a long time for an RL agent to converge to a stable control policy, e.g., 40 days reported by (Vázquez-Canteli et al., 2019) and over 50 days presented by (Fazenda et al., 2014). Besides, the DRL agent is supposed to learn policies based on the environment states under various actions to make a good decision. However, it is not practical to traverse the whole action space because the thermal comfort of the building needs to be maintained to avoid complaints from occupants. Therefore, simulation-based pre-training of DRL agents was adopted in various research (Chen et al., 2019; Luo et al., 2022). Fu et al., 2023 investigated the performance of both MPC and DRL control in a single zone building energy system assuming that no modeling error exists, i.e., using the same high-fidelity model for model predictions in MPC or agent training and control signal implementations in DRL. However, a model error is inevitable in practice, and developing a high-fidelity model is time-consuming and requires a lot of data and effort to calibrate (Zhang et al., 2019). Adopting a reduced-order model (ROM), such as Resistance-Capacitance (RC) model or data-driven model, is a more feasible approach than using a high-fidelity model such as Modelica and EnergyPlus.

This paper aims to investigate the following two questions:

1. How to evaluate the control-oriented model? The error between building energy modeling and real building is inevitable. The accuracy of models should not be the only criterion for a control-oriented model. The model's insensitivity to its actions strongly affected the agent's performance even though the absolute error of the predictive model is small (Luo et al., 2022). The sensitivity of different ROMs and the Modelica model to random and periodical control signals are compared.
2. How will the DRL controller perform when trained in ROM environments and deployed in a virtual building environment? Training DRL in a model-based environment offline and deploying the trained policy in a real building will raise problems such as unstable control policy (Zhang et al., 2019). However, the degree of the influence is still unclear. This paper compared the control performances of a DRL controller trained in ROM environments, the RBC and the ideal DRL controllers trained in a high-fidelity virtual building environment.

METHODOLOGY

A virtual testbed that can perform an offline training interacting with ROM environments and an online deployment in an environment with a high fidelity model was developed. The framework is shown in Figure 1. The training and testing environments and the DRL agents are packaged and deployed in a Docker container (Anderson, 2015) that could be a plug-and-play application in different computation environments. The major difference between training and testing environments is the building and HVAC system model. The testing environment (i.e., Modelica- based virtual building) is described in detail in (Fu et al., 2023). In the testing environment, the Modelica model is compiled by the co-simulation using Functional Mockup Unit (FMU) and interacts with the DRL agent through the ModelicaGym (Lukianykhin & Bogodorova, 2019). In the training environment, the building and HVAC system are simulated by ROMs. ROMs are trained based on the historical operation data generated by virtual buildings. Currently, the 3R2C, Light Gradient Boosting Machine (LightGBM), and artificial neural network (ANN) models are developed to predict the zonal indoor air temperature, and a polynomial model is established to simulate the HVAC system's

performance. The weather data and time-of-use electricity price are provided for both the training and testing environment. The interface of training environments and DRL agents is built on OpenAI Gym (Brockman et al., 2016). The virtual testbed is implemented in Python. The DRL agent gives actions according to the state of the environment to maximize the value of the reward function with the guidance of policy. The policy will be updated during various trials over time. In this research, the practical DRL agent was pre-trained interacting with the training environment offline and then deployed the well-trained policy in the testing environment for online control while the ideal DRL agent was trained and deployed both in the testing environment to get the optimal control performance. In this way, the deviation between the high-fidelity model and ROMs mimics the model error between predictive models and real building in practice.

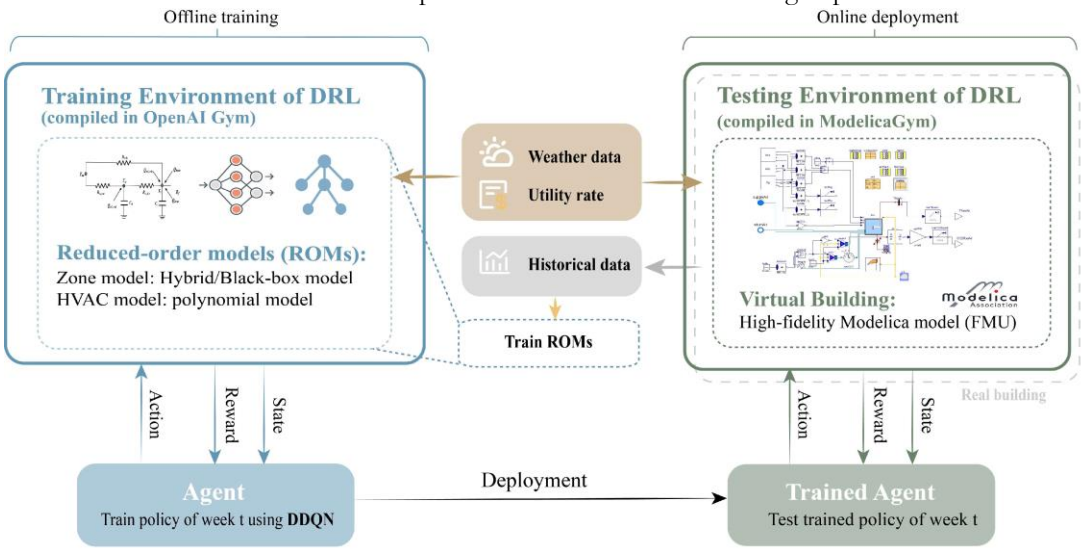

Figure 1 Framework of DRL virtual testbed.

This research takes a single zone building with a fan coil unit system as a case study to compare to the controllers' performance reported in (Fu et al., 2023) that assumed there were no model errors. All the parameters of the building, HVAC system and control problem formulation in this research are set the same as (Fu et al., 2023) to ensure a fair comparison. The single zone building is *case900FF* with a heavyweight construction (Judkoff & Neymark, 1995). The building consists of four exterior walls and a flat roof with a floor dimension of 6 m-by-8 m. The floor-to-ceiling height is 2.7 m. The east and west walls have the shorter dimension and the south wall contains two 3 m-by-3m windows. The building is assumed to be located in Chicago, Illinois, USA, and the Typical Meteorological Year 3 weather data is used for the simulation. The HVAC system is a direct expansion coil system. This research only studied the cooling scenario. The supply air temperature is assumed to be a constant value at 14 ◦C for the simplification. The zone temperature is controlled by modulating the fan speed, which is the control variable or action in our DRL controller. The capacity of the fan is 0.45 m^3/s (953.50 CFM). The control variable, i.e., the speed of the fan consists of 50 discrete control actions ranging from 0 to 1 with an increment of 0.02 to meet the requirement of the DRL algorithm, i.e., Double Deep Q- Network (DDQN) used in the case study. The control step is every 15 minutes.

DRL CONTROLLER DESIGN

Control Target and Reward Function

The rewards function is formulated as Eq. 1. The target of the DRL controller is to reduce the energy cost while maintaining the indoor thermal comfort and stability of the HVAC system. So, the reward function contains three terms, i.e., energy cost, thermal comfort, and action slew rate.

$$R_t = -(\omega_1 p_t P_t \Delta t + \omega_2 \epsilon_t^2 + \omega_3 \Delta u_t^2) \quad \text{Eq. 1}$$

In Eq. 1, ω_1, ω_2. ω_3 are the weights for the three terms. The product of energy price at time t, p_t, the electricity power consumption at time t, P_t , and the time interval, Δt , stands for the energy cost. The difference between zone temperature setpoint and indoor air temperature at time t, ϵ_t , denotes the temperature violation. The slew rate of control signal Δu indicates the system oscillation. The reward value at time t, R_t , is the negative sum of the three terms because the DRL agent is designed to maximize the reward. The weights in the rewards function are set to [100, 1, 10] to facilitate the comparison of the control results with the ideal DRL controller in (Fu et al., 2023).

Action-space Design

The action in DRL is what the agent gives to the environment, corresponding to the control signal in the control problem. The action in this case is the speed ratio of the fan. The action space is designed as 50 discrete fan speed ratios ranging from 0 to 1 to meet the requirement of the DDQN algorithm.

State-space Design

States are the feedback that the agent receives from the environment. The states at each timestep include current measurements, historical measurements and future disturbances which can be obtained by prediction. Table 1 summarizes all the states and their boundaries designed for the DRL environment, which includes the time index, indoor air temperature, outdoor air temperature, solar radiation, power consumption and utility rate. Historical and current states usually can be collected by building automation systems or weather stations, the future weather data can be accessed via weather forecasting and the future utility rate is feasible to get from energy providers. Therefore, the design of environment states is practical to implement. In this paper, the previous step m and the future step k are set to 4 which means information from 8 timesteps need to be provided at each timestep.

Table 1 Summary of States in DRL environment.

Symbol	Bound	Description
t	[0, 86400]	Current time index in seconds, [s]
$T_{z,t}$	[5,40]	Zone temperature at time t, [°C]
$T_{oa,t}$	[-10,40]	Outdoor air temperature at time t, [°C]
$\dot{q}_{s,t}$	[0, 1000]	Solar radiation at time t, [W/m2]
P_t	[0, 1500]	Power at time t, [W]
p_t	[0,1]	Energy price at time t, [$/kWh]
$T_{oa,\{t+1\ldots t+k\}}$	[-10,40]	Outdoor air temperature at the next k steps from t, [°C]
$\dot{q}_{s,\{t+1\ldots t+k\}}$	[0, 1000]	Solar radiation at the next k steps from t, [W]
$p_{\{t+1\ldots t+k\}}$	[0,1]	Energy price at next k steps from t, [$/kWh]
$T_{z,\{t-m\ldots t-1\}}$	[5,40]	Zone air temperature at previous m steps from t, [°C]
$P_{\{t-m\ldots t-1\}}$	[0, 1500]	Power at previous m steps from t, [W]

REDUCED-ORDER MODEL (ROM)

Developing and calibrating a high-fidelity model is laborious work. ROMs are introduced to surrogate high-fidelity model for pre-training the DRL agent. This research developed a hybrid model, i.e., 3R2C model, and two data-driven models using LightGBM and ANN algorithms for predictions of zone air temperatures and a polynomial model to predict HVAC power consumption.

Hybrid Zone Model: 3R2C

The reduced-order RC model simplified the energy balance in the zone with multiple heat resistances and capacitances. The structure of the lumped 3R2C model is shown in Figure 2 and the mathematic formulation is given

by Eq. 2, where C_z, C_w represent the thermal capacity of zone and walls; $R_{w,int}$, $R_{w,ext}$, R_{win} denote the lumped thermal resistance of interior wall surface convection and conductance, exterior wall surface convection and conductance, and window; T_{oa} , T_w , T_z represent the outdoor air temperature, wall temperature and indoor air temperature; $Q_{sol.wall} = \alpha_{wal} A_{wal} q_{sol}$, $Q_{sol.win} = \tau_{win} A_{win} q_{sol}$ are solar heat gains absorbed by external wall surface, solar heat flux transmitted through the window, α_{wal}, τ_{win} are the absorptivity of wall and transmissivity of window; Q_{inter}, Q_{hvac} are internal heat gains and cooling energy conveyed by the HVAC system, respectively. The cooling energy conveyed by the HVAC system can be calculated via $Q_{hvac} = \dot{m}_{ret} * h_{ret} - \dot{m}_{sup} h_{sup}$, where supply air rate ,$\dot{m}_{sup}$, and and return airflow, $\dot{m}_{ret}$, can be calculated by the control signals, enthalpy of supply air and return air h_{sup}, h_{ret} can be approximated using the fixed supply air temperature and indoor air temperature. The thermal resistances, thermal capacities, the absorptivity of wall and transmissivity of the window α_{wal}, τ_{win} were identified using one week data before the control period. The HVAC system was excited with a free floating case during the night and random fan speeds in the daytime to fully capture the dynamics of the zone. The reasons for utilizing only one week of data are twofold: first, it expedites the solution process, and second, the benefit from increasing the size of training data is marginal. The measurements of the dataset contain outdoor air temperature, zone temperature, internal heat gain, solar radiation and cooling energy from the HVAC system which are feasible to get or estimate in real building operations. The energy flow from infiltration is not considered for this preliminary study as the infiltration in this study was neglectable.

$$C_z \frac{\delta_{T_z}}{\delta_t} = \frac{T_w - T_z}{R_{w,int}} + \frac{T_{oa} - T_z}{R_{win}} + Q_{hvac} + Q_{sol.win} + Q_{inter} \quad \text{Eq. 2a}$$

$$C_w \frac{\delta_{T_w}}{\delta_t} = \frac{T_{oa} - T_w}{R_{w,ext}} + \frac{T_w - T_z}{R_{w,int}} + Q_{sol.wal} \quad \text{Eq. 3b}$$

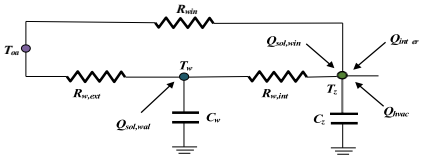

Figure 2 Structure of 3R2C model for single zone building.

Data-driven Zone Model: LightGBM and ANN

LightGBM is a gradient-boosting framework that uses tree-based learning algorithms (Ke et al., 2017). It is famous for its faster training speed, higher efficiency and low memory usage. LightGBM framework showed a high accuracy in building energy prediction applications (Miller et al., 2020). The artificial neural network (ANN) is a widely used algorithm in the machine learning field. Considering the data availability in real-life operation, only four features, including outdoor air temperature, global solar irradiance, hour of the day and airflow rate of the fan which is the control variable, are adopted as the input of the LightGBM model and the ANN model in this study. Auto machine learning (autoML) techniques are utilized to find the optimal hyperparameters for the LightGBM model and ANN model. The extrapolation of the data-driven model is typically poor. So, the training dataset should fully cover the action space and state space. Most variables in the state space are external disturbances that we can't control in reality. A comprehensive dataset, from fully excited action space was established. Initially, the training data comprises data from one month, with one week for free float setting, one week for periodical signal increasing fan speed from 0% to 100%, and two weeks for random fan speed signal. However, the LightGBM model had a bad performance when responding to the control signal, even though the root mean square errors (RMSEs) for the zone air temperature predications in the testing period for one day with random fan speeds at daytime and fan turned off at night was small of only 0.94 ◦C (1.69 °F). The

LightGBM model trained by the aforementioned dataset tends to underestimate the zone temperature when the fan speed is small, so that the DRL controller pre-trained the LightGBM environment tends to turn down the fan speed because the zone temperature received by the DRL agent is still lower than the upper boundary even the zone temperature in the virtual building is very high. This indicates that the low model errors in testing dataset cannot guarantee the performance of controller. Their response to control signals should be examined. To get more reliable data-driven models, the period of training data was increased to 3 months with one month with free float setting, one month with periodicals signal, and one month with random fan speed signals.

Performances of Zone Models

What has a significant impact on the performance of the DRL controller may not be the model accuracy using the testing dataset but the model's sensitivity to the actions of the agent (Luo et al., 2022). So, in this research, the models' performances are evaluated by their response to random control signals and periodical control signals increasing the fan speed from 0% to 100% in the testing week. The random or periodical 50 discrete fan speeds, aligning with the action design in DRL, are exerted to ROMs and high-fidelity Modelica model, respectively. The zone temperature predictions of ROMs are compared with the response of the high-fidelity Modelica model. Figure 3 gives an example of the periodical signal and the responses of ROMs and the Modelica model. Generally, all the ROMs can capture the up and down trends of zone temperature. LightGBM and ANN models have bad performances when the fan speed is low. The mean and maximum absolute error (AE) which are widely used to measure the difference between predicted values and the actual values are used to quantify the response error. Table 2 summarizes the evaluation matrices of ROMs. LightGBM model has the lowest mean AE while the 3R2C model has the lowest maximum AE. Data-driven models have large maximum AEs. The LightGBM and ANN model have some significant errors in response to control signals within specific regions, notably at lower fan speeds. The response of the 3R2C model is the most stable.

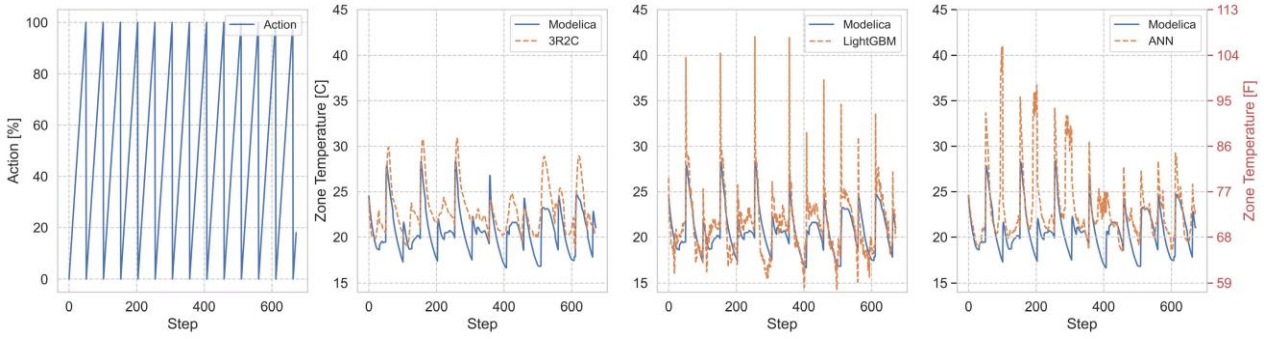

Figure 3 Responses to Periodical Control Singal of Zone Models.

Table 2 Matrices to evaluate ROMs' performances.

Model	Random Signal				Periodical Signal			
	Mean AE [∘C]	Mean AE [°F]	Max AE [∘C]	Max AE [°F]	Mean AE [∘C]	Mean AE [°F]	Max AE [∘C]	Max AE [°F]
3R2C	2.18	3.93	5.66	10.18	2.36	4.25	5.72	10.30
LightGBM	1.56	2.81	14.69	26.44	1.59	2.86	16.29	29.31
ANN	2.61	4.69	25.35	45.63	2.56	4.62	23.61	42.50

HVAC Model: Polynomial

A polynomial model described by Eq. 4 was developed to surrogate the Modelica HVAC model in the cooling season, where α_1, α_2, α_3, α_4 are the coefficients identified by the historical data; m denotes the airflow rate of the fan; T_z and T_s are the indoor air temperature and supply air temperature, respectively; COP is the Coefficient of Performance of the chiller, which is a constant value of 3 in the ideal Modelica chiller model. The first four terms represent the power

consumption of the fan and the last term is the power consumption of the chiller. One-month historical data is used to identify α_1, α_2, α_3, α_4. The polynomial model fits the historical data exceptionally well, achieving a Coefficient of the Variation of the Root Mean Square Error (CV-RMSE) (Hwang et al., 2022) of 1.6% in the training dataset and 1.5% in the testing dataset since the Medelica model is an ideal HVAC system model.

$$P = \alpha_1 + \alpha_2 m + \alpha_3 m^2 + \alpha_4 m^3 + \frac{1008 * m * (T_z - T_s)}{COP} \quad \text{Eq. 4}$$

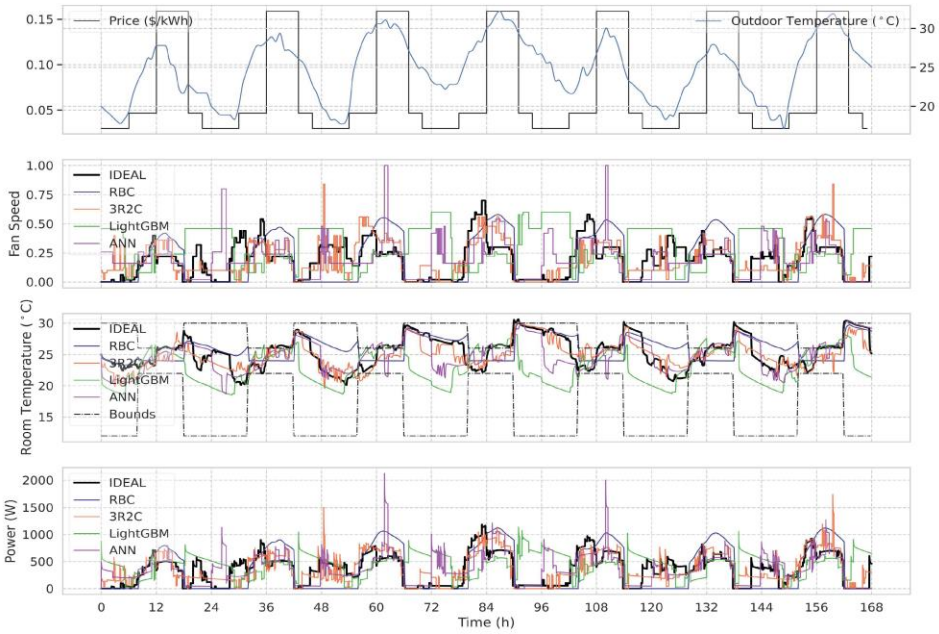

Figure 4 Comparison of performance by RBC, ideal DRL controller, and DRL based on different ROMs.

RESULTS AND DISCUSSIONS

The practical DRL controller trained in the ROM environment and directly deployed in the virtual environment of the high-fidelity Modelica model was compared with the ideal DRL controller trained and deployed in the high- fidelity environment and the baseline, i.e., rule-based control, which sets the zone air cooling setpoint to 24 ◦C for occupied hours and 30 ◦C for unoccupied hours. Due to the stochastic properties of DRL algorithms, six experiments were run under different random seeds. The time-of-use electricity price, outdoor air temperature, and the controller's performance under the seed with the best rewards in the test week (seven days) are shown in Figure 4. The ideal solution reveals the upper bound of the DRL controller under the DDQN algorithm in this case. The control signals of all ROM controllers fluctuate more than those from the ideal DRL controller. The DRL controller based on the 3R2C training environment has the most similar control output as the ideal DRL controller. The DRL controller based on the LightGBM training environment maximizes the utilization of nighttime low electricity prices for precooling the zone compared with controllers with other training environments. The key performance indicators, the maximum rewards of the final epoch, and the maximum, average rewards of the high-fidelity environment in the training process are summarized in Table 3. The ROM-based DRL controllers lower the energy cost at the expense of some thermal comfort. As shown in Table 3, all the ROM-based DRL controllers under the "best" random seed have higher reward values than the RBC while having lower reward values than the ideal DRL controller. In the final episode of the case study, the 3R2C, LightGBM, and ANN-based DRL outperformed the RBC by 7.4%, 14.4%, and 11.4%, respectively while the ideal DRL outperformed the RBC by 29.5% in terms of reward. Even though the data-drive-based DRL controllers have a lower best reward, the average reward is much worse than the 3R2C-based DRL controller., which

indicates that the data-driven-based DRL controller has a poor stability and could be problematic to implement in practice directly. The deviation of the 3R2C-based DRL controller under different random seeds is much lower than the other ROM-based DRL controllers. This may come from the fact that the 3R2C model has a relatively accurate response to any action in the action space, as shown in Figure 3 of the model performance section.

Table 3 Control performances of different DRL controllers

Scenario	Energy Cost [$]	Total Thermal Discomfort [Kh]	Maximum Temperature Violation [K]	Action Changes [-]	Reward of Final Episode[a] [-]	Best Reward[b] [-]	Average Reward[c] [-]
RBC	7.16	2.00	2.32	1.84	-746.90	-746.90	-746.90
Ideal DRL	5.27	12.01	0.96	3.45	-564.02	-526.29	-557.36
DRL-3R2C	5.98	8.91	2.55	6.95	-691.70	-638.20	-754.05
DRL-LGB	5.63	12.32	1.71	3.34	-639.67	-582.00	-1747.11
DRL-ANN	5.65	12.08	2.19	5.13	-661.74	-637.77	-1377.94

a. Episodic reward in the final episode with the "best" random seed. The final episode for 3R2C, LGB and ANN-based DRL is 300, 200, and 300. The reward in the virtual building trained by the LGB-based DRL will fluctuate after 200 episodes even though the reward in the training environment is stable and convergent.
b. Best reward is given by the testing virtual building environment during the training process among 6 random seeds.
c. Average reward is given by the testing virtual building environment during the training process of 6 random seeds.

CONCLUSIONS AND FUTURE WORK

This study investigated the influence of model error in the deep reinforcement learning (DRL) based controller. The performances of the practical reduced-order model (ROM)-based DRL controllers which are trained in different ROM environments are compared with a rule-based controller (RBC) and an ideal DRL controller which is trained and deployed both in the virtual building environment. The conclusions are drawn as follows:

- The preliminary results indicate that all the ROM-based DRL controllers outperform the RBC but are inferior to the ideal DRL controller in terms of rewards in the final episode. The ROM-based DRL controllers are more unstable than the ideal DRL controller with a higher maximum temperature violation and higher deviation between maximum reward and average reward among experiments with 6 random seeds.
- DRL controllers trained in data-driven environments could have higher maximum rewards but have much lower average rewards and larger deviations with different random seeds than controller trained in the 3R2C environment. The 3R2C model is the most robust among all ROM cases in this preliminary study.
- Data-driven models tend to be accurate when the training data is sufficient but perform poorly in specific ranges of actions, such as low fan speed regions. The performance of the 3R2C model is stable in the full action space despite having a slightly larger mean AE than the lightGBM model. This might be the reason that the 3R2C-based DRL controller is the most robust among ROM-based DRL controllers.

It's worthwhile to propose metrics to evaluate the control-oriented models comprehensively. The measurements to improve the performance and stability of the practical ROM-based DRL controller, e.g., adaptive deployment, integration of model uncertainty and policy gradient algorithms will be investigated in the future.

NOMENCLATURE

ANN	=	Artificial Neural Network
COP	=	Coefficient of Performance
DDQN	=	Double Deep Q Networks
DRL	=	Deep Reinforcement Learning
HVAC	=	Heating, Ventilation, and Air Conditioning
AE	=	Absolute Error
RBC	=	Rule-Based Control
ROM	=	Reduced-order Model

REFERENCES

Anderson, C. (2015). Docker [Software engineering]. *IEEE Software*, *32*(3), 102-c3. https://doi.org/10.1109/MS.2015.62

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). *OpenAI Gym* (arXiv:1606.01540). arXiv. https://doi.org/10.48550/arXiv.1606.01540

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*, 316–325. https://doi.org/10.1145/3360322.3360849

EIA. (2023, April 20). *Use of electricity—U.S. Energy Information Administration (EIA)*. https://www.eia.gov/energyexplained/electricity/use-of-electricity.php

Fazenda, P., Veeramachaneni, K., Lima, P., & O'Reilly, U.-M. (2014). Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems. *Journal of Ambient Intelligence and Smart Environments*, *6*(6), 675–690. https://doi.org/10.3233/AIS-140288

Fu, Y., Xu, S., Zhu, Q., O'Neill, Z., & Adetola, V. (2023). How good are learning-based control v.s. model-based control for load shifting? Investigations on a single zone building energy system. *Energy*, *273*, 127073. https://doi.org/10.1016/j.energy.2023.127073

Fu, Y., Zuo, W., Wetter, M., VanGilder, J. W., Han, X., & Plamondon, D. (2019). Equation-based object-oriented modeling and simulation for data center cooling: A case study. *Energy and Buildings*, *186*, 108–125. https://doi.org/10.1016/j.enbuild.2019.01.018

Hwang, R.-L., Liao, W.-J., & Chen, W.-A. (2022). Optimization of energy use and academic performance for educational environments in hot-humid climates. *Building and Environment*, *222*, 109434. https://doi.org/10.1016/j.buildenv.2022.109434

Judkoff, R., & Neymark, J. (1995). *International Energy Agency building energy simulation test (BESTEST) and diagnostic method* (NREL/TP-472-6231). National Renewable Energy Lab. (NREL), Golden, CO (United States). https://doi.org/10.2172/90674

Kazmi, H., Mehmood, F., Lodeweyckx, S., & Driesen, J. (2018). Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems. *Energy*, *144*, 159–168. https://doi.org/10.1016/j.energy.2017.12.019

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html Lukianykhin,

O., & Bogodorova, T. (2019). ModelicaGym: Applying reinforcement learning to Modelica models. *Proceedings of the 9th International Workshop on Equation-Based Object-Oriented Modeling Languages and Tools*, 27–36. https://doi.org/10.1145/3365984.3365985

Luo, J., Paduraru, C., Voicu, O., Chervonyi, Y., Munns, S., Li, J., Qian, C., Dutta, P., Davis, J. Q., Wu, N., Yang, X., Chang, C.-M., Li, T., Rose, R., Fan, M., Nakhost, H., Liu, T., Kirkman, B., Altamura, F., … Mankowitz, D. J. (2022). *Controlling Commercial Cooling Systems Using Reinforcement Learning* (arXiv:2211.07357). arXiv. https://doi.org/10.48550/arXiv.2211.07357

Miller, C., Arjunan, P., Kathirgamanathan, A., Fu, C., Roth, J., Park, J. Y., Balbach, C., Gowri, K., Nagy, Z., Fontanini, A. D., & Haberl, J. (2020). The ASHRAE Great Energy Predictor III competition: Overview and results. *Science and Technology for the Built Environment*, *26*(10), 1427–1447. https://doi.org/10.1080/23744731.2020.1795514

Qiu, S., Li, Z., Li, Z., Li, J., Long, S., & Li, X. (2020). Model-free control method based on reinforcement learning for building cooling water systems: Validation by measured data-based simulation. *Energy and Buildings*, *218*, 110055. https://doi.org/10.1016/j.enbuild.2020.110055

Vázquez-Canteli, J. R., Ulyanin, S., Kämpf, J., & Nagy, Z. (2019). Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities. *Sustainable Cities and Society*, *45*, 243–257. https://doi.org/10.1016/j.scs.2018.11.021

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep Reinforcement Learning for Building HVAC Control. *Proceedings of the 54th Annual Design Automation Conference 2017*, 1–6. https://doi.org/10.1145/3061639.3062224

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. https://doi.org/10.1016/j.enbuild.2019.07.029

Zong, Y., Böning, G. M., Santos, R. M., You, S., Hu, J., & Han, X. (2017). Challenges of implementing economic model predictive control strategy for buildings interacting with smart energy systems. *Applied Thermal Engineering*, *114*, 1476– 1486. https://doi.org/10.1016/j.applthermaleng.2016.11.141