2021

Reinforcement Learning for Building Management Systems

Parastoo Delgoshaei
National Institute of Standards and Technology, p.delgoshaei@gmail.com

Amanda Pertzborn
National Institute of Standards and Technology

Mohammad Heidarinejad
Illinois Institute of Technology

Reinforcement Learning for Building Management Systems

Parastoo Delgoshaei[1]*, Amanda Pertzborn[2], Mohammad Heidarinejad[3]

[1]Mechanical Systems and Controls Group, Engineering Laboratory, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA (Parastoo.Delgoshaei@nist.gov)

[2]Mechanical Systems and Controls Group, Engineering Laboratory, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA (Amanda.Pertzborn@nist.gov)

[3] Department of Civil, Architectural, and Environmental Engineering, Illinois Institute of Technology, Chicago, IL 60605, USA (muh182@iit.edu)

* parastoo.delgoshaei@nist.gov

ABSTRACT

It is increasingly common to design buildings with advanced sensing and control systems to improve energy efficiency, indoor air quality which impacts health and productivity. However, there has been limited progress in making building automation systems "intelligent," as the performance of such buildings is often limited by reactive control systems, primarily using setpoint limits and fixed operation schedules. The complex nature of building control problems motivates the application of state-of-the-art software engineering methods and techniques. Agent-based models (ABM) are well-suited for controlling complex engineering systems such as those employed in building heating, ventilation, and air-conditioning (HVAC) systems. In this paradigm, a collection of interacting autonomous components (i.e., agents) adapt and make decisions in changing environments. There is a growing body of literature on adaptive agents in ABMs in many industries, but few have looked at the compatibility of ABMs with artificial intelligence (AI) optimization approaches. In most cases, conventional optimization techniques, such as mixed integer linear programming and gradient descent, have been used to find an optimal solution. This paper explores the use of an actor-critic, model-free algorithm based on a deterministic policy gradient that provides continuous control to generate the desired supply air temperature. The case study develops a thermal energy storage (TES) agent that determines the optimal valve position to manage the temperature of the cooling water flow. The case study was developed using the Intelligent Building Agents Laboratory at the National Institute of Standards and Technology. Future work will use multiple agents (i.e., air handling unit, TES, chiller) acting in cooperation or competition.

1. INTRODUCTION

There is a need for more intelligent and efficient building control algorithms because people spend 85 % of their time inside buildings, and the building sector uses about 40 % of the total primary energy consumption in the U.S. (DOE, 2012). However, better control algorithms can result in significant reduction in energy consumptions in buildings. In general, building control can be divided into three broad categories: rule-based, model-based, and data-driven.

Rule-based methods can be further categorized into prescriptive and heuristic approaches. Prescriptive methods can be obtained from best practices, equipment schedules, and rules of thumb, with a goal of improving indoor air quality. A heuristic approach adjusts setpoints based on previous experiences; for example, setting the temperature back during unoccupied hours or demand response events. ASHRAE Guideline 36 summarizes heuristic rules (ASHRAE, 2018). Although rule-based approaches are effective, they are not optimal and are not necessarily designed for the specific system in use. Moreover, these algorithms are not dynamic; they do not change over time based on the operational status of the system and do not use predictions about the future.

Model-based approaches are generally based on physics-based mathematical models. One example of such an approach is model predictive control (MPC), which uses the thermal or energy dynamics of the system and disturbance predictions to make optimal decisions. The optimization algorithm computes the optimal strategy based on predictions

over a specific time horizon. In this approach, the focus is prediction of the disturbance (i.e., weather, cost, occupancy). MPC has been shown to save energy or lower operating costs in both simulation (Paris et al., 2010) and field tests (Prívara et al., 2011). Although MPC algorithms have the potential to outperform conventional control algorithms, they are not widely adopted in buildings since each building and its energy systems are unique. It is challenging to find a general-purpose building energy model that would be suitable for a variety of buildings.

Data-driven approaches are tailored for each building since they are based on historical building data. Due to advances in measurement techniques, the availability of large volumes of data in the built environment, and abundant and inexpensive computational power, recent studies have examined machine learning algorithms for building control. Reinforcement learning (RL) is a branch of machine learning that is used for optimal decision making. RL has been used extensively in gaming (Szita, 2012) and robotics (Kober & Peters, 2009) and is becoming more popular in other applications that require intelligent decision making.

RL has been successfully applied to many areas of building energy management and control. The work proposed by (Zhang et al., 2019) implemented a model-based RL approach that learns the system dynamics using a neural network. Then MPC uses the learned system dynamics to perform control via a random-sampling shooting method. Their approach reduced the total energy consumption, of a simulated two zone data center, by 17.1% to 21.8%. Q-Learning is one of the most popular RL algorithm, introduced in 1989 (Cornish & Watkins, 1989). Advances in computational power have enabled the combination of RL algorithms with deep neural networks, significantly increasing the effectiveness of RL methods and enabling them to be used in complex and safety critical applications such as autonomous cars. Another study (Barrett & Linder, 2015) used Q-learning combined with Bayesian learning for occupancy prediction. They implemented a heating, ventilation, and air conditioning (HVAC) controller to optimize occupant comfort and energy costs. The study shows that a learning thermostat can achieve cost savings of 10 % over a programmable thermostat, while maintaining high levels of occupant comfort. In a comparison to a rule-based system, Yang et al. (2015) developed a batch Q-learning approach using neural networks to control a photovoltaic powered heating system (Yang et al., 2015). The results indicate a 10 % improvement over the rule-based approach. Using a deep neural network, the studies in (Wei et al., 2017, 2018) developed RL-based HVAC control and the results suggest energy saving improvements of 20 % to 70 % over conventional Q-learning methods. The above studies indicate that RL algorithms can be promising for building mechanical systems control. This paper attempts to explore? how the performance of a chilled water system can be optimized using an actor-critic model. This is a proof of concept effort to investigate the efficacy of the model and the impact of targeted data collection to train the RL model.

2. BACKGROUND

RL is a sub-category of machine learning that includes at least one agent interacting with an environment. Through its interactions with the environment, the agent learns a series of optimal actions by determining through trial-and-error which actions yield the best reward based on the selected criteria. Table 1 describes commonly used concepts in the RL field.

Table 1. Concepts and their descriptions that are commonly used in RL

Concept	**Description**
Environment	Physical world in which the agent operates
Reward	Feedback from the environment
Policy	A mapping from perceived states of the environment to actions to be taken when in those states
State	Current state of the environment
Value	The expected (long term) cost of a certain policy

2.1 Subcategories of Reinforcement Learning

RL algorithms can be categorized as model-based or model-free. Algorithms that purely sample from experience are "model-free" RL algorithms. Model-free approaches do not use a model of the transition of the environment from one state to another state. For example, Monte Carlo Control, SARSA (Rummery & Niranjan, 1994), Q-learning (Cornish & Watkins, 1989), and actor-critic (Konda, 2002) algorithms are model-free. These algorithms learn the optimal policy based on a trial-and-error method. Model-based methods, however, learn a model of the environment. RL algorithms belong to two different categories:

Value-Based: These algorithms focus on finding an approximation of the value function or functions of states (state-action pairs).
Policy-Based: These algorithms try to find an optimal policy directly without using the Q-value.

Each of these algorithms have their advantages and disadvantages. For example, value-based algorithms are more stable while policy-based approaches converge faster and are better for continuous environments. The actor-critic method merges the advantages of both approaches. Figure 1 illustrates the actor-critic approach used in this study.

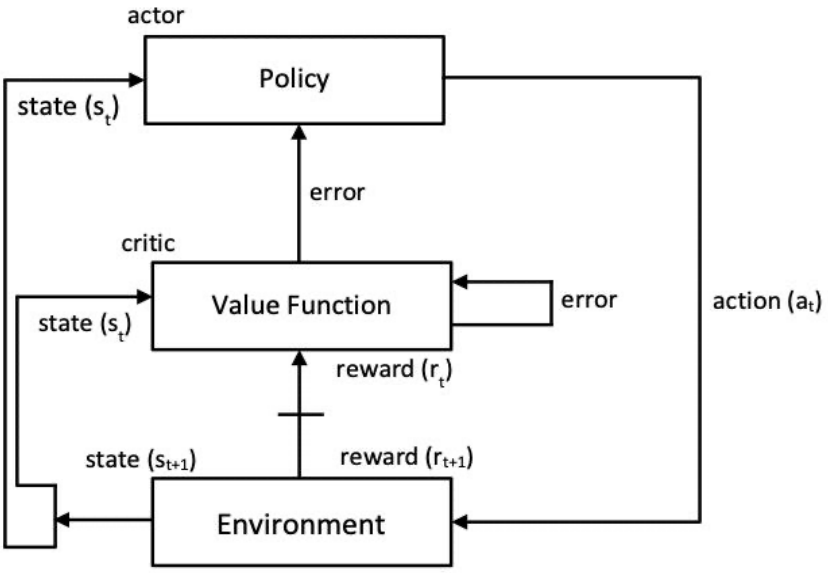

Figure 1: Architecture of the actor-critic method

In this figure, the agent is composed of an actor and a critic. The actor takes the state and the error as inputs to determine the best action. The actor uses a policy-based approach to learn the optimal policy. The critic uses a value-based approach to evaluate the action taken by the actor.

3. CASE STUDY

The Intelligent Building Agents Laboratory (IBAL) at the National Institute of Standards and Technology (NIST) was designed to emulate the HVAC and cooling loads in a small commercial building (Pertzborn & Veronica, 2018). The IBAL contains experimentally controllable weather, air system zones, and a hydronic system, which includes the cooling plant. Figure 1 is a schematic of the portion of the IBAL that is used in this study. The plant includes two water cooled chillers, Chiller1 and Chiller2, and thermal energy storage (TES). Chiller1 has a nominal capacity of 26.4 kW and Chiller2 has a nominal capacity of 52.8 kW, but both chillers use variable speed compressors, so they have variable capacity. The chillers are used to meet the building load or charge the TES, which is an ice-on-coil design with a capacity of 233 kWh. The chillers are plumbed in parallel with each other and in series with the TES. The hydronic system uses a 30 % propylene glycol (PG) mixture. The left portion of the figure, which includes the

plant, is the primary loop (PL) and the right portion of the figure, which includes the building load (HX1), is the secondary loop (SL).

One use of TES is for load shifting. Instead of using a chiller on a hot afternoon to meet a building load, which is a time of peak energy usage, the TES can be used. The chiller can instead operate during times of off-peak energy usage, such as at night, to charge the TES. This type of operation takes pressure off the grid during times of peak use and can reduce energy cost to the consumer if the utility uses a tariff structure that incentives load shifting. The TES can also be used in conjunction with a chiller. In that scenario, there are two benefits. First, a smaller chiller can be selected in the design phase because the chiller does not have to meet all the load, which can reduce both first and operating costs. Second, the TES can meet enough of the load to maintain chiller power consumption below the level that would trigger demand charges.

For this research, the building load is emulated by use of a heat exchanger, HX1. HX1 is a plate heat exchanger with hot water (approximately 71 °C) flowing through one side and PG flowing through the other side. The building load is set by modulating valve V11, which sets the flow rate of the hot water through the heat exchanger. For this scenario, the temperature T_{SA} is a proxy for the supply air temperature downstream of a cooling coil. Figure 2 shows the schematic of the equipment and measurements.

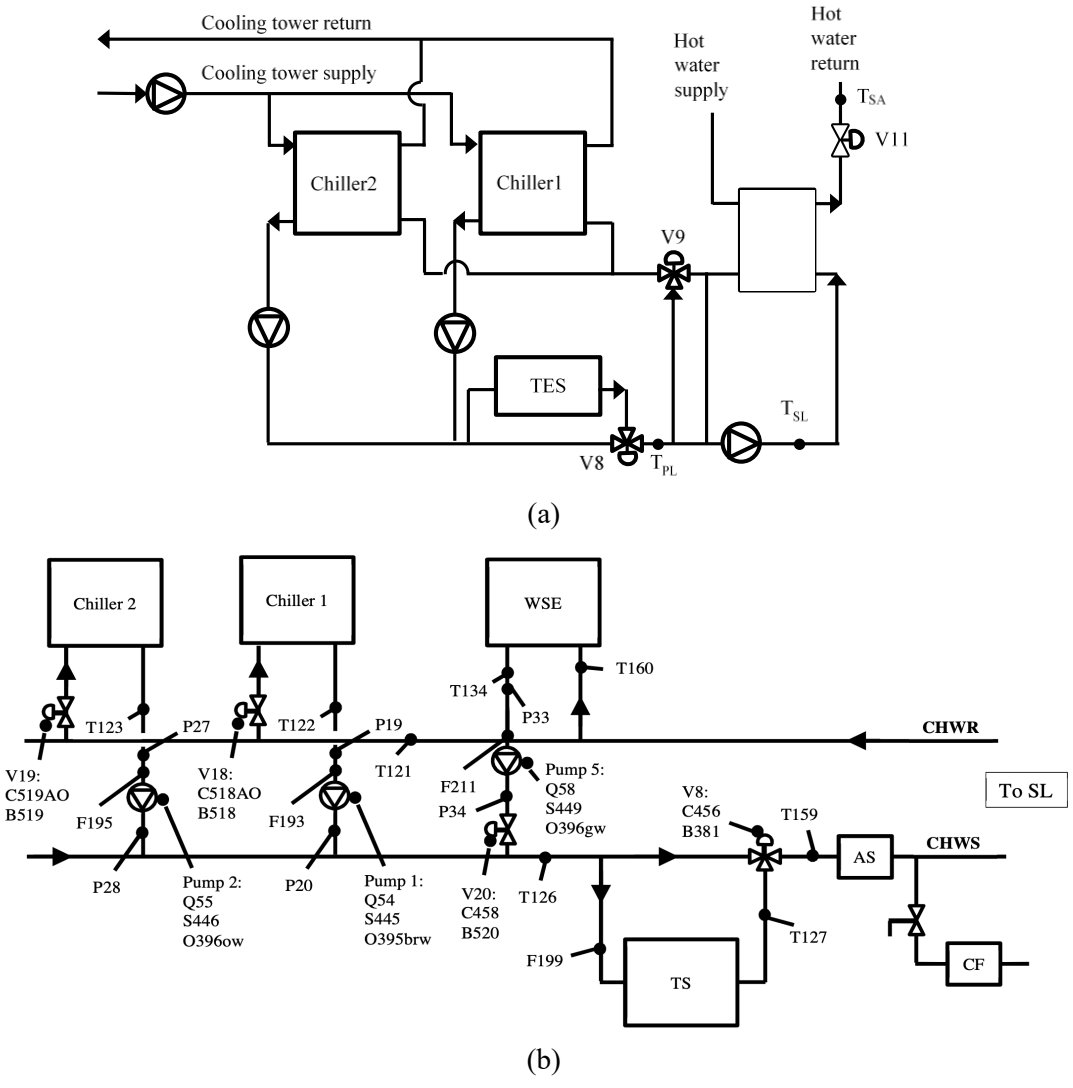

Figure 2: Schematic of the chiller, thermal energy storage, and building load (HX1) components in the IBAL: (a) high-level and (b) detailed sensor installation

The lab can support the following modes of operation:

- Chiller1 meets the building load
- Chiller1 charges the TES (charge)
- Chiller2 meets the building load
- Chiller2 charges the TES (charge)
- TES meets the building load (discharge)

When a chiller is used to meet the load, valve V8 is positioned so that the TES is bypassed. When a chiller is used to charge the TES, V8 is positioned so that all the PG from the chiller flows through the TES. When the TES is used to meet the building load, if V8 is positioned such that all the PG flows through the TES, the temperature leaving the PL, T_{PL}, would be 0 °C, which is typically colder than needed in the SL. Instead, V8 modulates so that T_{PL} is generated by the mixing of PG from the TES and returning from the SL. The temperature in the SL, T_{SL}, can be further controlled by modulating V9 to mix the fluid from the PL and the outlet of the SL. This provides fine control of the temperature.

For this study, we focus on the mode where the TES is used to meet the building load and examine an approach for optimizing the control signal for V8.

4. METHODS

The goal of the case study was to optimally control the position for the three-way mixing valve V8, shown in Figure 2, using the RL actor-critic method. V8 regulates the flow distribution between the chiller and thermal storage branches. When the control signal for V8 is 0 V, the resistance in the thermal storage branch will force all the flow to come from the chiller branch. Similarly, when the control signal is 10 V, the flow will be provided from the thermal storage branch. This mechanism will mix the two flow branches to provide the right temperature to the secondary loop to ensure the supply air temperature can meet the building needs.

4.1 Model Formulation

In the actor-critic formulation used here, the policy estimator (actor) and the value estimator (critic) are both implemented as fully connected deep neural networks with one output and no activation functions. These neural networks are trained using a stochastic gradient descent optimization algorithm with learning rates of 0.001 and 0.1 for the actor and critic, respectively. At each iteration the actor network predicts an action that will cause the environment to transition to a new state. Subsequently, the critic evaluates the value of the new state of the environment based on the penalty function, which is the negative of the reward, and provides an error to the actor. If the error is negative, the actor will weaken the tendency for that action to be taken. The reward function can consider factors such as energy, cost, or performance. In this case study, we focused on meeting the setpoint while reducing the chiller power consumption. In this problem formulation, the goal is to minimize the reward and thus the reward is defined as a penalty function, described in Equation (1).

$$R = -A \times (Setpoint - Supply)^2 - B \times (Chiller\ Power)^2 \quad (1)$$

In our case study, the agent is an actor-critic model that predicts the optimal valve position. To mimic the behavior of the zone that is the environment in our model, we used data from the heat exchanger (HX1) that serves as a building load. The state in this formulation is the temperature that leaves the heat exchanger, T_{SA} in Figure 2a, and serves as the proxy for the supply temperature. In this work we used regression to model the transition of the state (supply temperature) of the environment based on the received actions.

4.2 Results

The following graphs were obtained using the model with test data collected from the IBAL with a temperature setpoint of 12.2 °C (54 °F) for T_{SA}. Figure 3 shows that in the first 250 episodes, the trial and error phase, the reward varies between 20 and 65. After episode 250, the penalty asymptotically reaches the minimum value of about 20 (the

algorithm is trying to minimize the negative of the reward) and stays around that value until the last episode. Thus, our reward is at its maximum value.

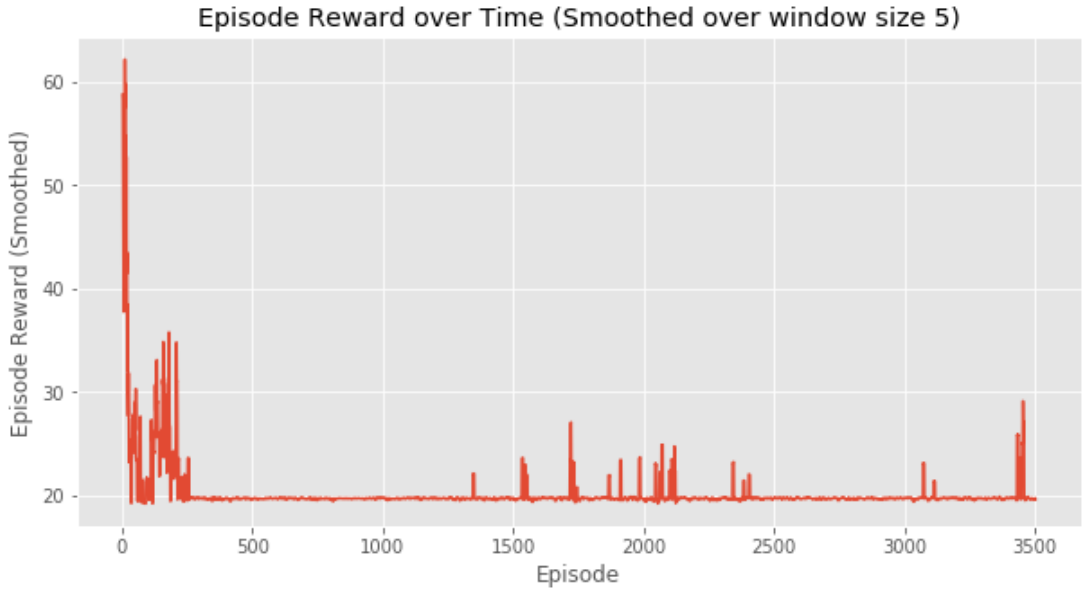

Figure 3: Episode reward over time

Similarly, as shown in Figure 4, the state of the environment converges to a value around 12.6 °C (54.6 °F) after episode 250. This is within an acceptable range for the target value of 12.2 °C (54 °F).

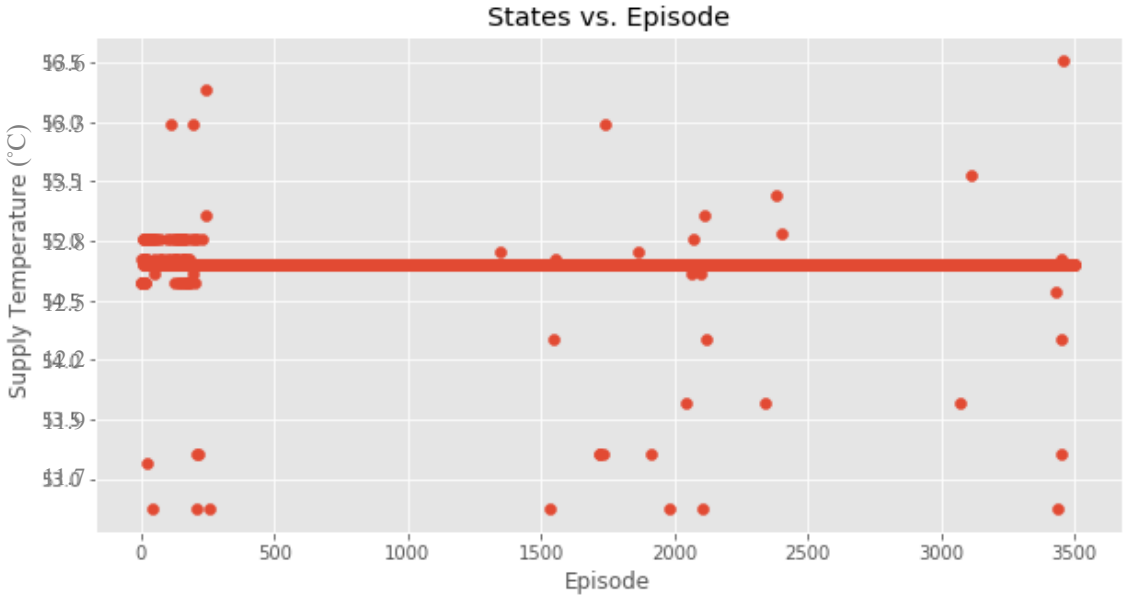

Figure 4: States versus episodes

Figure 5 shows a similar pattern; more potential actions (valve positions) were explored before episode 250. After the exploration phase, the valve position converges to a value between 1.2 and 1.6. This result indicates that the majority of the flow will bypass the ice tank because the building load is relatively small.

Figure 5: Actions versus episodes

5. DISCUSSION

This study applied an actor-critic RL algorithm to find an optimal solution for a mixing valve on thermal energy storage in a building HVAC system using real data. This approach is a first step towards making more complex control decisions and moving away from prescriptive sequences of operation for HVAC control. One observation we made during this study is that simply using generic historical experimental data that are not specific to the problem being studied can lead to convergence problems. In addition, the noise inherent in experimental data can make them unusable in RL algorithms in a raw form, requiring extensive preprocessing, which complicates the use of RL. Development of more complex control systems will require more targeted and less noisy datasets. Overall, compared to supervised or unsupervised algorithms that have immediate feedback or no feedback, RL algorithms that provide delayed feedback are suitable for HVAC building systems with careful consideration of the reward function. The next step of this work to implement this approach in the lab and validate the optimization results.

6. CONCLUSIONS AND FUTURE WORK

This paper deployed reinforcement learning (RL) to develop optimal control for a part of the hydronic system in the Intelligent Building Agents Laboratory (IBAL) at the National Institute of Standards and Technology (NIST). The actor-critic algorithm was used to find an optimal position of a three-way valve on a thermal energy storage system. The case study in this work serves as a proof of concept for applying RL for building HVAC control. RL algorithms are promising approaches for building control. One caveat is that these algorithms require the right dataset for proper training and exploration. Constructing the reward function is not a trivial task; for the purpose of this case study, reward was determined by minimizing energy consumption and achieving the setpoint.

An extension of this work will focus on a multi-agent model with agents such as the TES, chiller and the air handling unit that compete and cooperate to achieve the common goal of system optimization. Factors to consider include selecting a chiller to meet the building load based on chiller capacity and power consumption and selecting the TES to meet the building load based on its capacity and the power consumption associated with charging the TES.

NOMENCLATURE

ABM	Agent-based model
AI	Artificial Intelligence
HVAC	Heating Ventilation and Air Conditioning

IBAL	Intelligent Building Agent Lab
MPC	Model Predictive Control
NIST	National Institute of Standards and Technology
RL	Reinforcement Learning
SL	Secondary Loop
TES	Thermal Energy Storage

REFERENCES

ASHRAE. (2018). *ASHRAE Guideline 36-2018: High-Performance Sequences Of Operation For HVAC Systems*. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

Barrett, E., & Linder, S. (2015). Autonomous HVAC Control, A Reinforcement Learning Approach. In A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, & M. Spiliopoulou (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 3–19). Springer International Publishing.

Cornish, C. J., & Watkins, H. (1989). *Learning from Delayed Rewards*. King's College.

DOE. (2012). *Buildings Energy Data Book*. Table 1.1.3: Buildings Share of U.S. Primary Energy Consumption. http://buildingsdatabook.eren.doe.gov/TableView.aspx?table=1.1.3

Kober, J., & Peters, J. (2009). Learning motor primitives for robotics. *2009 IEEE International Conference on Robotics and Automation*, 2112–2118. https://doi.org/10.1109/ROBOT.2009.5152577

Paris, B., Eynard, J., Grieu, S., Talbert, T., & Polit, M. (2010). Heating control schemes for energy management in buildings. *Energy and Buildings*, *42*(10), 1908–1917. https://doi.org/10.1016/j.enbuild.2010.05.027

Pertzborn, A., & Veronica, D. (2018). *Intelligent Building Agents Laboratory: Air System Design* (Technical Note (NIST TN)-2025). National Institute of Standards and Technology.

Prívara, S., Široký, J., Ferkl, L., & Cigler, J. (2011). Model predictive control of a building heating system: The first experience. *Energy and Buildings*, *43*(2), 564–572. https://doi.org/10.1016/j.enbuild.2010.10.022

Rummery, G. A., & Niranjan, M. (1994). *On-Line Q-Learning Using Connectionist Systems*. Cambridge University Engineering Department.

Szita, I. (2012). Reinforcement Learning in Games. In M. Wiering & M. van Otterlo (Eds.), *Reinforcement Learning: State-of-the-Art* (pp. 539–577). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27645-3_17

Wei, T., Chen, X., Li, X., & Zhu, Q. (2018, November). *Model-based and data-driven approaches for building automation and control*. ICCAD '18: Proceedings of the International Conference on Computer-Aided Design. https://doi.org/10.1145/3240765.3243485

Wei, T., Wang, Y., & Zhu, Q. (2017, June 18). *Deep Reinforcement Learning for Building HVAC Control*. DAC '17, Austin, TX, USA.

Konda, V. R. (2002). *Actor-Critic Algorithms*. Massachusetts Institute of Technology.

Yang, L., Nagy, Z., Goffin, P., & Schlueter, A. (2015). Reinforcement learning for optimal control of low exergy buildings. *Applied Energy*, *156*, 577–586. https://doi.org/10.1016/j.apenergy.2015.07.050

Zhang, C., Kuppannagari, S. R., Kannan, R., & Prasanna, V. K. (2019, November 13). *Building HVAC Scheduling Using Reinforcement Learning via Neural Network Based Model Approximation*. ACM BuildSys '19, New York, NY, USA.